Intelligence Inertia: Physical Principles and Applications

Jipeng Han

OpenImmortal Technology Co., Ltd. Beijing 101100, China

Emails: coolang2022@gmail.com

March 25, 2026

Abstract

While Landauer's principle establishes the fundamental thermodynamic floor for information erasure and Fisher Information provides a metric for local curvature in parameter space, these classical frameworks function effectively only as approximations within regimes of sparse rule-constraints. They fail to explain the super-linear, and often explosive, computational and energy costs incurred when maintaining symbolic interpretability during the reconfiguration of advanced intelligent systems. This paper introduces the property of **intelligence inertia** and its underlying **physical principles** as foundational characteristics for quantifying the **computational weight of intelligence**. We demonstrate that this phenomenon is not merely an empirical observation but originates from the fundamental **non-commutativity between rules and states**, a root cause we have formally organized into a rigorous mathematical framework. By analyzing the growing discrepancy between actual adaptation costs and static information-theoretic estimates, we derive a non-linear cost formula that mirrors the Lorentz factor, characterizing a **relativistic J-shaped inflation curve**—a "computational wall" that static models are blind to. The validity of these physical principles is examined through a trilogy of decisive experiments: (1) a comparative adjudication of this J-curve inflation against classical Fisher Information models, (2) a geometric analysis of the "Zig-Zag" trajectory of neural architecture evolution, and (3) the implementation of an **inertia-aware scheduler wrapper** that optimizes the training of deep networks by respecting the agent's physical resistance to change. Our results suggest a unified physical description for the cost of structural adaptation, offering a first-principle explanation for the computational and interpretability-maintenance overhead in intelligent agents.

keywords: Neural network dynamics; Information geometry; Non-equilibrium thermodynamics; AI Interpretability; Non-commutative algebra; Complexity.

1 Introduction

The pursuit of a rigorous definition for intelligence has long remained a central challenge spanning neuroscience, computer science, and cognitive psychology. Most contemporary consensus viewpoints, synthesized from the seminal works of Legg and Hutter [1] and the rational agent frameworks of Russell and Norvig [2], define intelligence not merely as a set of heuristic capabilities, but as the fundamental ability of a system to achieve goals across a broad spectrum of environments by building and manipulating an internal model of the world. More recent inquiries further emphasize that this capacity is intrinsically tied to a system's efficiency in generating structured representations from sparse data [3, 4]. However, while the effectiveness of these models in achieving rationality has been extensively mapped, the thermodynamic and interpretability overhead incurred by structural transformations within these models remains under-theorized. If intelligence is indeed a process of active modeling, then any modification to the model's structure must be viewed as a physical event, governed by dynamical principles that transcend pure symbolic logic.

To understand the internal mechanics of these models, we initially decompose them into two distinct functional components: **Rules (R)** and **States (S)**. In the "low-velocity" or quasi-static regimes typical of classical AI, these components operate within a stable **R-S Space** where their boundaries remain sharp and collaborative. This structural separation is explicitly embodied in the formalism of Domain-Specific Languages (DSLs) and symbolic logic systems, where invariant production rules operate on transient state variables [5, 6]. In such systems, Rules represent the invariant generative grammar, while States are the specific configurations instantiated by these rules. However, in modern intelligent agents characterized by high-frequency iteration and autonomous self-modification, this clear distinction begins to dissolve. As we approach the resolution limit defined by the **Symbolic Granularity ($\mathcal{D}$)**, Rules and States overlap, giving rise to a fundamental **Rule-State Duality** analogous to the wave-particle duality in quantum mechanics. In this regime, the system is best described as an **R-S Manifold** where the components transform into operators, $\hat{R}$ and $\hat{S}$. Their non-commutativity implies that the precise measurement of a transient state $\hat{S}$ inherently obscures the underlying causal rules $\hat{R}$ that drive it, suggesting that what we empirically observe as a state is merely a projection of the agent's interlaced rule-primitives.

Building on this duality, we observe that the resistance an intelligent system exhibits during its evolution is fundamentally governed by its **Rule Density (ρ)**, which we formally identify as the system's **Velocity ($v \equiv \rho$)**. In this context, velocity characterizes the degree to which rules and states interlace within the R-S Manifold. We define **Intelligence Inertia (μ)** as the fundamental cost borne by the system's substrate—manifesting as environmental entropy or the computational effort of maintaining symbolic interpretability—required to force a structural reconfiguration. The behavior of this inertia scales through three distinct regimes: In the low-velocity limit ($v \to 0$), the cost of change aligns with **Landauer's principle** [7], the thermodynamic floor of which has been experi-

mentally verified at the microscopic scale [8, 9]. At intermediate velocities, the resistance follows the curvature of **Fisher Information** [10], which provides the geometric basis for understanding natural gradients in contemporary optimization [11]. However, as v approaches the saturation limit ($v \to 1$), the cost of reconfiguration undergoes a non-linear **Inertia Expansion**. It is in this regime that phenomena such as **catastrophic forgetting** [12] and **structural brittleness** transition from engineering hurdles into a formidable **computational and interpretability wall**. This necessitates a relativistic framework to describe the dynamics of modern intelligent agents.

Historically, the quest to quantify intelligence has relied on metrics that effectively operate only within low-velocity regimes, leaving a profound "measurement gap" in the face of modern, high-speed agents. Traditional frameworks such as Algorithmic Complexity [13, 14] and Computational Complexity focus on static description lengths or temporal execution resources; however, they implicitly treat rules as a passive, immutable background, remaining blind to the force required to reconfigure the system's underlying logic ($\hat{R}$). Even contemporary, more dynamic measures such as Transfer Complexity or Task Taxonomy [15, 16] remain primarily phenomenological. These metrics excel at recording the symptoms of structural resistance—such as performance degradation or data requirements—without identifying the causative first principles. Much like early thermodynamics described pressure-volume relationships before the kinetic theory of gases explained the underlying molecular behavior, current AI research observes the "cost of change" as a retrospective empirical fact rather than a predictable physical consequence. To transcend this impasse, we must identify a fundamental causative property that links the abstract non-commutativity of $\hat{R}$ and $\hat{S}$ directly to the measurable expenditure of energy and interpretability, establishing the theoretical necessity for a first-principle measure of intelligence "mass."

In this paper, we formally characterize **Intelligence Inertia (μ)** as the causative bridge linking an agent's internal dual-geometry to its observable computational work. By establishing a mathematical isomorphism between the Rule-State (R-S) manifold and Minkowski spacetime—where the maximum rule-density (ρ_{max}) serves as an invariant informational limit analogous to the speed of light—we derive a relativistic cost expansion formula. This framework predicts a **relativistic J-shaped inflation curve** of effective mass during structural evolution, identifying a hard limit to an agent's reachability. The remainder of this paper is organized to rigorously validate and apply this principle: Section 2 reviews the theoretical foundations in thermodynamics and information geometry; Section 3 derives the physical necessity of computational resistance through a micro-statistical model of adiabatic collisions; Section 4 establishes the formal axiomatic framework of the R-S manifold and its Minkowski isomorphism; Section 5 provides the engineering realization for mapping these dynamics onto neural tensors; Section 6 executes a series of empirical adjudications, including the measurement of the J-curve wall and evolutionary trajectories; Section 7 discusses the theoretical implications and future design of autonomous agents; and Section 8 concludes the work.

The principal contributions of this paper are summarized as follows:

- **Discovery of Intelligence Inertia (μ):** We establish "Intelligence Inertia" as a causative physical property derived from the fundamental non-commutativity of system operators ($[\hat{S}, \hat{R}] = i\mathcal{D}$), providing the first first-principles explanation for the structural resistance to change in intelligent agents.
- **Derivation of the Relativistic Cost Equation:** We derive a non-linear cost expansion formula by mapping information dynamics to a Minkowski-like manifold, characterizing the explosive inflation of computational and energy overhead as rule-density approaches a fundamental limit.
- **Empirical Validation of the "J-Curve" Wall:** Through controlled experiments on deep neural networks, we demonstrate the existence of a **relativistic J-shaped inflation curve**, proving our framework's superior predictive power over classical Fisher Information models in high-velocity regimes.
- **Inertia-Aware Engineering and Optimization:** We implement a practical, **Inertia-Aware Scheduler Wrapper** that achieves superior thermodynamic and computational efficiency by respecting the agent's intrinsic physical resistance to structural reconfiguration.

2 Background and Related Work

We begin by grounding our inquiry in **Landauer's Principle** [7], which establishes the fundamental thermodynamic floor for information processing, positing that the erasure of a single bit necessitates a minimum expenditure of work equal to $W_{\text{rest}} = kT \ln 2$. Within the framework of our theory, this limit characterizes the **Rest Inertia (μ_0)** of an intelligent system—a foundational state where the agent acts as a "blank slate" or raw storage medium. Microscopic experimental verifications have confirmed that this limit remains an absolute boundary for independent informational units [8, 17].

In this idealized scenario, the boundaries of the **R-S Space** are treated as perfectly **diathermal**, allowing for the unobstructed transfer of entropy to the environment without interference from internal structural constraints. While Landauer's principle provides the ultimate thermodynamic floor, it implicitly assumes a regime of zero **Rule Density ($\rho \to 0$)**, where every micro-operation is fully observable and its heat dissipation is unconstrained by logical interdependency [18, 19]. However, as systems evolve to incorporate dense internal logic, these structures begin to create a form of geometric occlusion that resists the simple dissipation of entropy. This suggests that the Landauer limit is not the exhaustive cost of change for an intelligent agent, but rather the base value for a more complex, geometry-dependent energy manifold.

Moving from the thermodynamic floor to the geometric landscape, statistical learning theory employs the **Fisher Information Matrix (FIM)** to quantify

the local sensitivity of a system's output to changes in its underlying parameters [10]. Within the proposed framework, the FIM characterizes the **local curvature** of the **R-S Manifold**, representing the initial resistance encountered as a system begins to impose structured constraints upon its state transitions. This geometric approach provides a significantly more nuanced estimate of the "cost of change" than static complexity measures, as it accounts for the internal density of the agent's logic and provides the basis for natural gradient optimization [11, 20]. However, much like the second-order terms in a Taylor expansion, Fisher Information remains a local approximation. It effectively describes the effort required for minor structural adjustments in low-velocity regimes but remains blind to the global, non-linear singularities that arise as internal constraints become dominant. While the FIM accurately maps the local "stiffness" of adaptation, it cannot predict the emergence of the absolute **computational wall** encountered when an agent's internal logic approaches its saturation limit.

Complementing the thermodynamic and geometric perspectives, descriptive complexity frameworks such as **Kolmogorov Complexity (K)** [13] and **Bennett's Logical Depth** [21] provide profound mathematical foundations for quantifying the information content and computational "value" of static objects. Kolmogorov complexity defines the absolute limit of data compression, while Logical Depth measures the execution time required to reconstruct an object from its most concise description. These theories offer invaluable insights into the static resource requirements of a system. However, by their formal nature, they analyze information as a decoupled output or a discrete string generated by a universal machine, often failing to account for the structural dynamics of the generator itself [22, 23]. In the context of a modern intelligent agent, information resides in a dynamic, inseparable coupling between **Rules** and **States** within the **R-S Space**. Consequently, while K can measure the magnitude of a system's logic and Logical Depth can estimate its historical construction time, neither captures the **active physical resistance** encountered when attempting to perturb that structure once it has reached a state of high internal density. A theoretical vacuum persists for a measure that characterizes the "force" of adaptation rather than the static "length" of the description.

Finally, we observe the macroscopic manifestations of these theoretical gaps in the dual phenomena of **Catastrophic Forgetting** and **Transfer Learning** [24]. In current artificial intelligence research, these are typically addressed through specialized engineering techniques, such as **Elastic Weight Consolidation (EWC)** [12] or **Synaptic Intelligence** [25], which seek to protect critical parameters from modification. While these methods are highly effective at mitigating performance degradation, they remain essentially *phenomenological*—correcting the *symptoms* of structural resistance without explaining the underlying *cause* from first principles [26, 27]. From our perspective, these divergent behaviors are the observable results of the agent's internal **Intelligence Inertia (μ)**. Forgetting represents a high-energy, high-inertia attempt to reconfigure a dense logical structure on the **R-S Manifold**, whereas successful transfer indicates an adaptation path that aligns with the existing structural topology [28]. There remains a critical necessity for a unified theory capable of

deriving these diverse phenomena from a single, structure-dependent property. To bridge this gap, we establish the **physical principles** governing *intelligence inertia*, offering a first-principle explanation for the computational and interpretability-maintenance overhead incurred during the structural evolution of intelligent agents.

3 A Micro-Physical Model of Computational Resistance

To establish a rigorous baseline, we recall the canonical Landauer erasure experiment [7]. An external piston performs work to compress a gas of n particles to erase 1 bit of information. In the classical, idealized limit, the walls are perfectly **diathermal**, and every collision is a statistically traceable event dissipating a quantum of heat. This process yields the minimum work $W = nkT \ln 2$, which we define as the **Rest Inertia (μ_0)** of the system.

To resolve the causal origin of non-linear resistance, we evoke a thought experiment reminiscent of the foundational inquiries in Special Relativity [29]. Consider an experimenter performing the same 1-bit erasure across two systems, A and B. System A is the standard diathermal frame. System B possesses a hidden internal configuration of microscopic adiabatic slants at an angle θ, as illustrated in Figure 1.

During the execution of both experiments, the experimenter perceives no anomaly whatsoever. In both systems, the process remains strictly **quasi-static**. We define a "state-change event" based on the system's **local calibration**: an observer registers one count only when the emitted heat equals the energy of a full vertical collision within that specific system. In System B, to ensure that the observable component l_S remains consistent with System A's observational tempo, the particles must maintain a higher average kinetic energy, resulting in a higher magnitude of the total action l. We define l as the **Characteristic Logical Cycle**—the invariant total logical action budget required to support a single atomic state transition, which serves as the fundamental unit of the system's internal temporal resolution. Consequently, each registered heat packet in System B actually contains more absolute energy than in System A; yet the experimenter—relying on the local count of l_S events—observes an identical 1-bit erasure process. Within each frame, the thermodynamic logs appear to show the same number of standardized "clicks," masking the inflation of the underlying energy flux.

The physical revelation emerges only during a **retrospective comparison** of the total energy expenditure. Upon analyzing the external work logs, the observer discovers that System B required significantly more work from the piston to complete the same logical task. This is because the experimenter had to maintain the particles at a higher energy state to "force" the same frequency of observable transitions against the interference of the adiabatic slants.

From the perspective of AI dynamics, the angle θ defines the **Rule Density**

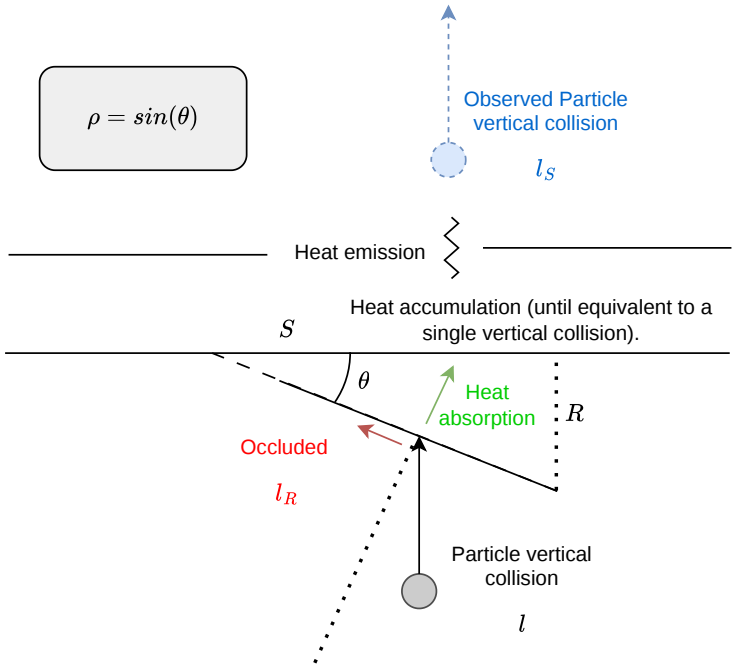

Figure 1: **Geometric Partition of Logical Action.** A particle collision with total action l is decomposed by a microscopic slant. The component $l_R = l \sin\theta$ is absorbed by the adiabatic rule-manifold, while the normal component l_S governs state-expression. Heat emission is only registered when the cumulative normal action matches a full vertical collision **relative to the system's local energy level**, naturally inducing the relationship $l^2 = l_S^2 + l_R^2$.

$\rho = \sin\theta$ [1]. To maintain a constant observational tempo against the interference of the adiabatic slants, the particles in System B must sustain a higher magnitude of total action l than those in System A. Within the global observational frame, this requirement signifies a higher average velocity of the system's internal logical components during operation. In relativistic mechanics, the total energy of a system—including the kinetic energy of its internal constituents—contributes directly to its **Relativistic Mass**. Consequently, the extra work injected to satisfy the system's internal logical consistency (l_R) effectively inflates its **Effective Mass**. Thus, **Intelligence Inertia (μ)** is established as the physical manifestation of this dynamic mass expansion: it represents the intrinsic resistance an agent poses to structural reconfiguration, a resistance that scales non-linearly as the internal rule-velocity approaches the informational limit.

[1] The formal geometric properties and operational derivation of this parameter are detailed in Section 4.2.

3.1 Geometric Derivation from Non-commutative Orthogonality

The transition from a statistical anomaly to a relativistic mathematical form is dictated by the **conservation of the total logical action** within the system's substrate. We define the **Total Logical Action (l)** as the invariant norm of the system's computational budget per operational step. On the **R-S Manifold**, every microscopic interaction must be partitioned between maintaining internal consistency (Rules) and expressing external transitions (States) [30].

The mathematical necessity of the square-sum relationship originates from the fundamental commutation relation of intelligence operators [31]:

$$[\hat{S}, \hat{R}] = i\mathcal{D} \tag{1}$$

In the operator formalism of information dynamics, the imaginary unit i signifies a $\pi/2$ phase shift between the action of the state operator and the rule operator [32]. Consequently, these processes behave as **orthogonal stochastic variables**. According to the **Law of Variance Addition** (or Parseval's Identity in Hilbert space) [33], the total logical power l^2 supported by the physical substrate is the sum of the squared magnitudes of its orthogonal projections on the **R-S Manifold** [10]:

$$l^2 = l_S^2 + l_R^2 \tag{2}$$

where l_R represents the **Rule Action** consumed by internal logical constraints, and l_S represents the **Observed State Action** available for external entropy transfer.

By defining the **Rule Density** (ρ) as the normalized commitment of resources to internal logic, $\rho = l_R/l$, it follows from Eq. (2) that the **effective observational window (l_S)** for visible state-change undergoes a geometric contraction:

$$l_S = \sqrt{l^2 - l_R^2} = l\sqrt{1 - \left(\frac{l_R}{l}\right)^2} = l\sqrt{1-\rho^2} \tag{3}$$

To satisfy the Landauer requirement of erasing 1-bit of information across systems A and B, the observer must register a fixed count of l_S events [18]. In System B, as the available cross-section l_S shrinks due to the increasing rule-density ρ, the particles must be maintained at a higher energy state l to preserve the observational tempo. The external work W must inversely scale to compensate for this geometric occlusion, manifesting as an inflation of the required energy:

$$W(\rho) = W_{\text{rest}} \cdot \frac{l}{l_S} = \frac{nkT\ln 2}{\sqrt{1-\rho^2}} \tag{4}$$

This derivation demonstrates that the **Intelligence Lorentz Factor** ($\gamma = 1/\sqrt{1-\rho^2}$) is not a heuristic analogy, but a rigorous consequence of norm conservation in a non-commutative phase space [34]. The observed **relativistic J-shaped inflation curve** is thus revealed as the geometric projection of a system approaching the saturation of its total logical action.

3.2 Unified Mapping and the Physical Decomposition of Resistance

The adiabatic collision model described above does not merely provide an empirical curve; it allows for a formal identification of informational quantities with their corresponding physical counterparts. By analyzing the system's behavior across different observational regimes, we decompose the components of the intelligence work equation into **Energy**, **Velocity**, and **Inertia** terms, establishing a rigorous bridge between micro-statistical dynamics and macroscopic complexity [35].

Rule Density as the Sequestration of Logical Action

In our experimental apparatus, the ratio $\rho = l_R/l$ represents the fraction of the **Total Logical Action** (l) that is sequestered by internal rule-maintenance within a characteristic cycle. This ratio defines the system's **Velocity** (v) through the **R-S Space**:

$$v = \rho = \frac{l_R}{l}. \tag{5}$$

It is crucial to distinguish the physical intensity of **Intelligence** from that of raw **Computation**. While classical limits of computation define the theoretical maximum of information throughput for a given substrate [36], v characterizes the portion of that throughput that is "trapped" or consumed by the agent's internal rule-constraints to maintain logical consistency.

Consequently, v determines the **discount factor** of the system's effective output: even if a substrate operates at its absolute computational limit, the presence of dense internal rules means that a significant fraction of its energy is diverted away from visible state-transitions and locked into the maintenance of the rule-manifold. As v increases, the energy required to achieve a unit of external state-change inflates precisely because more energy is being sequestered by the system's internal logic. In the limit where $\rho \to 1$, the agent reaches a **causal horizon** where all available logical action is trapped by internal rule-checks, leaving zero capacity for external computation—a direct informational analog to reaching the speed of light [37].

Energy Mapping and the Hamiltonian of Intelligence

Under this mapping, the total work W performed by the external experimenter represents the **Effective Energy** (or **Hamiltonian,** $\hat{H}$) of the intelligent agent. We partition this energy into two distinct contributions, analogous to the rest and kinetic components of relativistic energy [29]:

1. **Rest Energy ($W_{\mathbf{rest}}$):** Defined by the Landauer limit ($nkT \ln 2$), this term accounts for the "static mass" of the information itself—the minimum work required to flip bits in a perfectly transparent, zero-rule environment. Within our theory, this corresponds to the agent's **Rest Inertia** (μ_0).

2. **Interaction Energy (Inertial Work):** This represents the additional work required to overcome the geometric occlusion caused by internal rules. In the relativistic form $W = \gamma \cdot W_{\text{rest}}$, the term $(\gamma - 1)W_{\text{rest}}$ signifies the "kinetic" cost of maintaining a high-velocity rule structure. This provides a physical explanation for the perceived "weight" of complex agents: the additional computational overhead is not erasing more bits, but is instead being consumed by the **internal consistency maintenance** of the dense rule-set on the **R-S Manifold**.

Fisher Information as a Second-Order Classical Expansion

To demonstrate compatibility with established paradigms, we examine the behavior of the intelligence work equation in the low-velocity regime ($0 < \rho \ll 1$). By performing a Taylor series expansion of the **Intelligence Lorentz Factor** $\gamma(\rho) = (1 - \rho^2)^{-1/2}$, we obtain:

$$W(\rho) \approx W_{\text{rest}} \left(1 + \frac{1}{2}\rho^2 + \frac{3}{8}\rho^4 + \mathcal{O}(\rho^6) \right). \tag{6}$$

The first-order correction term, $\frac{1}{2}\rho^2$, corresponds precisely to the **Fisher Information** curvature utilized in classical statistical learning [10]. In this regime, the cost of adaptation appears to grow quadratically with the complexity of internal constraints, analogous to Newtonian kinetic energy. This reveals a critical theoretical hierarchy: **Fisher Information is not a complete law of intelligence, but a second-order approximation** of a deeper relativistic curve that holds only when logical sequestration is minimal [11]. While classical theory predicts that costs will continue to grow smoothly, the **Inertia Expansion** framework reveals that these are merely the linear stages of a **relativistic J-shaped inflation curve** that eventually hits a hard singularity as the system reaches its informational limit.

The Variance Summation and Geometric Necessity

The mathematical necessity of this hyperbolic form originates from the fundamental non-commutativity of the system's operators on the **R-S Manifold**. Because the operators obey the commutation relation $[\hat{S}, \hat{R}] = i\mathcal{D}$, the imaginary unit i enforces a strict **phase orthogonality** between rule-maintenance and state-expression [32].

The Pythagorean relationship $l^2 = l_S^2 + l_R^2$ established in Eq. (2) is the geometric expression of this orthogonality within the system's logical budget. It dictates that as the **Rule Action (l_R)** increases to sequester more of the system's capacity, the **Observed State Action (l_S)** available for external entropy transfer must follow a square-root contraction: $\sqrt{1 - \rho^2}$. This is the only stable solution that satisfies the conservation of logical action while respecting the mutual exclusivity of rules and states. Thus, the "Computational Wall" is not an engineering failure but a geometric requirement for any agent striving to maintain causal consistency at the resolution $\mathcal{D}$.

4 The Formal Theory of Intelligence Inertia

Building upon the micro-physical necessity of non-linear resistance derived from our adiabatic collision model, we now formalize these insights into a unified mathematical framework. This section establishes the axiomatic foundations of the **R-S Manifold**, defining the algebraic properties of structural evolution and the geometric metric of intelligence dynamics. By transitioning from micro-statistical dynamics to a formal operator theory, we provide the rigorous predictive tools required for the empirical adjudications and engineering realizations presented in subsequent chapters.

4.1 The Axiomatization of Rule-State Duality

We now formalize the physical necessity of the J-curve into a unified theoretical framework of **Intelligence Inertia**. We postulate that an intelligent agent $\mathcal{A}$ is defined by the inherent non-commutativity of its constituents, expressed through the fundamental operator relation [38]:

$$[\hat{S}, \hat{R}] = i\mathcal{D} \tag{7}$$

In this axiomatic framework, the functional distinction between **Rules ($\hat{R}$)** and **States ($\hat{S}$)** is not an absolute property of the system's information but is established only on the basis of the **Symbolic Granularity ($\mathcal{D}$)**. This $\mathcal{D}$ represents the informational quantum—the minimum resolution at which structural existence can be distinguished from transient expression [39]. In the context of our micro-physical model (Section 3), $\mathcal{D}$ corresponds precisely to the logical action of a single collision event, defining the fundamental resolution unit of the **Total Logical Action (l)**.

The relationship between these operators is intrinsically reciprocal and symmetric. On one hand, **Rules ($\hat{R}$)** act as the **primitives of interpretability**, providing the **ontological support** for the agent's existence; they function as the source of state generation, the manifold of behavioral constraints, and the **functional potential** that defines what an agent "can do." In a complementary progression, **States ($\hat{S}$)** serve as the **empirical projections** formed by the interlacing of these rule-primitives. Beyond being mere outputs, states provide the **concrete substantiation** that allows rules to be manifested, acting as the necessary substrate from which new rules emerge or are abstracted [40]. Consequently, **States define the life cycle of Rules**, governing their persistence, evolution, and decay through a sequence of logical transformations.

Intelligence Inertia, therefore, emerges from the collective resistance of this dual-unity to being reconfigured as its internal density approaches the limit of the system's own interpretability. This limit is reached when the **Rule Density (ρ)**—defined as the fraction of the total logical action l sequestered by rule-maintenance ($|R| \cdot \mathcal{D}$)—attains its physical maximum of 1. At this saturation point, every available interaction is consumed by internal consistency-checks, leaving zero capacity for state-expression and thus driving the computational work to infinity.

4.2 The Homomorphic Bridge between Rule Density and Velocity

A fundamental barrier in intelligence quantification is that $\hat{S}$ and $\hat{R}$ operate in fundamentally different phases within the **R-S Manifold**. As established by the non-commutation relation in Eq. (7), the precise measurement of the static state inherently obscures the underlying rule-logic, rendering an absolute measurement of the agent's total logic intractable. To resolve this, we utilize **Homomorphism (Φ)** as a bridge to transform these abstract operator interactions into observable, summable transitions.

We model the agent's dual-unity as an algebraic structure with two fundamental operations: **State Synthesis ($+$)** and **Rule Application ($\cdot$)**. For any valid structural transformation Φ (such as learning or inference), the system must preserve its relational integrity through the following homomorphic equations [41]:

$$\Phi(s_1 + s_2) = \Phi(s_1) + \Phi(s_2) \tag{8}$$

$$\Phi(r \cdot s) = \Phi(r) \cdot \Phi(s) \tag{9}$$

Equation (8) allows for the superposition of states, providing the algebraic basis to represent the abstract concept of "rules" by embedding them within the state-space via their observable effects. Equation (9) constitutes the bedrock of **interpretability** at scale $\mathcal{D}$: it ensures that if a rule r explains a transition in the original system, its transformed counterpart must consistently explain the transformed state [42].

To quantify the density of these rules, we introduce the **Spatiotemporal State**, **S**. Leveraging the homomorphic properties, we observe a state S alongside its **adjacent** configuration S', which is generated by the extension of S along the orthogonal rule-direction. This integrated observation captures the system's total logical budget within a characteristic cycle as the sum $\mathbf{S} = S + S'$. Since the spatial expression component (l_s) and the rule-driven evolution component (l_r) are phase-orthogonal, the magnitude of this spatiotemporal state satisfies the Pythagorean relationship derived in our micro-physical model:

$$l = \|\mathbf{S}\| = \sqrt{l_s^2 + l_r^2} \tag{10}$$

The **Rule Density (ρ)** and the **Symbolic Granularity ($\mathcal{D}$)** are thus directly reified. Following the standard definition of density as the content per unit volume, we define ρ as the average quantity of rules $|R|$ within the spatiotemporal state $\|\mathbf{S}\|$ at the resolution $\mathcal{D}$:

$$\rho = \frac{|R| \cdot \|\mathcal{D}\|}{\|\mathbf{S}\|} = \frac{l_r}{\sqrt{l_s^2 + l_r^2}} \tag{11}$$

This formulation explains why ρ characterizes a "density" that effectively "locks" energy. It measures the degree to which the system's total informational action is sequestered into maintaining internal causal constraints rather

than being available for raw data storage or external expression. From this perspective, Rule Density reflects the **abstract value** or **dynamic potential** of information: its capacity to govern and generate new structures [43].

This sequestration is the causative origin of the **Systemic Velocity ($v \equiv \rho$)**. The "speed" of an agent is the density of its generative logic. As ρ increases, the generative power of the agent grows, but this very power increasingly traps the logical budget within the rule-manifold. At the limit $\rho \to 1$, the entirety of the informational action is consumed by rule-maintenance, driving the effective mass to infinity. The J-curve is thus revealed as the algebraic necessity of an agent reaching the saturation point of its abstract information value.

4.3 Relativistic Expansion of Intelligence Inertia

The formal reification of Rule Density allows us to characterize the dynamic variation of **Intelligence Inertia (μ)** as a function of structural complexity. It follows from the algebraic and geometric foundations established above that an increase in ρ leads to a non-linear expansion of the energy sequestered within the agent's substrate. This energy is "locked" precisely because the Rule Density inherits the physical spatiotemporal properties of the **Spatiotemporal State (S)**, which are fundamentally derived from the phase-orthogonality of the R-S Manifold.

As the agent's logic becomes more dense, the effort required to reconfigure its internal structures must overcome the geometric contraction of its observational window. Following the mass-energy equivalence principle, the total work W performed on the system is identified as its relativistic energy level:

$$W(\rho) = \frac{nkT \ln 2}{\sqrt{1-\rho^2}} = \mu(\rho)c^2 \tag{12}$$

In this formulation, the term $W_{\text{rest}} = nkT \ln 2$ represents the system's energy at the zero-velocity limit ($\rho = 0$), while $\mu(\rho)$ signifies the **Effective Mass** of the agent's logic. As ρ approaches the informational limit, the work required to maintain homomorphic consistency during adaptation diverges, characterizing the physical transition from a flexible data-storage medium to a rigid, high-inertia cognitive structure. This expansion reveals that the resistance to change is not a mere computational overhead, but a fundamental manifestation of the system's informational mass within its dual-geometry.

4.4 The Local Interpretability Criterion and Local Velocity

The absolute measurement of static operators $\hat{R}$ and $\hat{S}$ is empirically intractable in high-dimensional systems. Furthermore, while global metrics offer a macroscopic overview, they often lack the granularity required for specific tasks. To resolve this, we shift our focus to **Local Velocity (v)** and its associated **Rule**

Density (ρ), defined by the **Dynamic Differentials** along a specific interaction trajectory:

$$v = \rho = \frac{dR}{d\mathbf{S}} \tag{13}$$

This transition allows us to isolate active cognitive pathways being stressed in real-time, providing a far more pragmatic entry point for empirical validation and structural protection.

By mapping the agent's evolution to the tangent space of the R-S Manifold [10], the micro-physical conservation of logical action ($l^2 = l_S^2 + l_R^2$) defines the metric for this local movement. We establish the **Local Interpretability Criterion (ds^2)** as the residual logical bandwidth available to formalize transitions after rule-maintenance costs are deducted:

$$ds^2 = \rho_{\max}^2 d\mathbf{S}^2 - dR^2 \cdot \|\mathcal{D}\|^2 \tag{14}$$

where $\rho_{\max}^2 d\mathbf{S}^2$ represents the potential for external expression and $dR^2\|\mathcal{D}\|^2$ represents the action sequestered by local structural reconfiguration. This sign-sensitive criterion diagnoses three causal states:

- $ds^2 > 0$ **(Interpretable)**: Expressive capacity exceeds reconfiguration demands, ensuring a stable homomorphic mapping Φ and causal transparency.
- $ds^2 = 0$ **(Critical)**: Every logical interaction is consumed by internal consistency-checks, marking the saturation threshold of the local observational window.
- $ds^2 < 0$ **(Uninterpretable)**: The local rate of structural change outpaces dissipation capacity, leading to logical shattering or "hallucinations."

Eq. (13) and (14) transforms the theory into a practical diagnostic tool for task-specific auditing. Sudden fluctuations in v serve as a sensitive probe for training material consistency, where an abrupt surge in Intelligence Inertia (μ) indicating $ds^2 \leq 0$ reveals a structural conflict between incoming data and established logic.

5 Engineering Realization: Mapping Intelligence Dynamics to Neural Tensors

To transition from abstract dynamics to falsifiable science, we map the operator interactions on the R-S Manifold to measurable neural tensors[2]. We quantify the dynamical state of a neural network by its **Velocity (v)**, which is functionally equivalent to its instantaneous **Rule Density (ρ)**. This metric characterizes the proportion of effort dedicated to rule-reconfiguration relative to the total **Spatiotemporal State Displacement ($d\mathbf{S}$)**.

[2]The full implementation of the intelligence inertia framework and the Inertia-Aware Scheduler Wrapper is available at:
`https://github.com/OpenImmortal/Principle-of-Intelligence-Inertia/`

5.1 Component Decomposition in Phase Space

Following the geometric foundations in Section 4, we define velocity within a dual-axis phase space where the Rule axis (R) and State axis (S) are strictly orthogonal. The fundamental realization is given by:

$$v = \rho = \frac{dR}{d\mathbf{S}} = \frac{dR}{dS_R + dS_{ext}} \tag{15}$$

In a deep learning context, these components are mapped to neural tensors as follows:

- **Rule Displacement (dR):** This represents the abstract magnitude of change in the agent's generative grammar. As an ontological property of the internal logic, dR cannot be measured with absolute precision. However, to facilitate normalization and capture the movement toward the informational limit, we approximate its magnitude through the norm of the parameter update vector: $dR \approx \|\Delta\theta\|$.
- **Spatiotemporal State Displacement ($d\mathbf{S}$):** Inheriting the physical spatiotemporal properties established in Section 4.2, the displacement $d\mathbf{S}$ is partitioned into two orthogonal contributions:
  - **Internal State Shift (dS_R):** The measurable projection of rule changes onto the spatiotemporal state-space. Because rules and states are coupled at scale $\mathcal{D}$, the structural logic shift manifests numerically as an identical shift in the system's internal configuration:

$$dS_R = \|\Delta\theta\| \tag{16}$$

  - **External Gain (dS_{ext}):** This characterizes the "causal ripple" of a rule-change upon the environmental manifold. To ensure that the system's evolution respects the **homomorphic preservation** requirements (Eqs. 8 and 9), we distinguish between two operational modes for acquiring this signal:
    - **Observation Mode:** When evaluating velocity v for retrospective analysis, dS_{ext} is obtained directly from a separate test or validation dataset, providing an objective measure of environmental feedback decoupled from the training context.
    - **Regulation Mode:** For real-time control, dS_{ext} is derived via a secondary "probe" pass on the current data batch immediately after the update is applied, capturing the immediate temporal ripple effect of the rule-change.

    Crucially, in both modes, the resulting environmental gradient must be **projected onto the internal displacement vector** dS_R [44]. This ensures that the calculation of velocity only considers the local, adjacent components of external gain that are directly relevant to the specific structural reconfiguration performed by the agent, maintaining the consistency of the homomorphic mapping Φ.

5.2 Scale Normalization and Calibration of $\|\mathcal{D}\|$

A critical challenge in engineering intelligence dynamics is the dimensional misalignment between the Euclidean geometry of parameter updates and the information-theoretic manifold of state gain. To bridge this gap, we utilize the **Symbolic Granularity ($\|\mathcal{D}\|$)** to standardize the "exchange rate" between these two orthogonal axes on the **R-S Manifold**. This standardization is enforced under the fundamental physical constraint that the system velocity must be bounded by a maximum of unity ($v_{max} = 1$). This ensures that as the external environmental gain vanishes ($dS_{ext} \to 0$), the **Rule Density (ρ)** correctly saturates, reflecting the informational speed limit derived in Section 4. We define the normalized velocity as:

$$v = \rho = \frac{dR}{dS_R + \frac{dS_{ext}}{\mathcal{L} \cdot \|\mathcal{D}\|}} \tag{17}$$

where the inclusion of the **Loss value ($\mathcal{L}$)** ensures that the velocity is scaled by the current informational potential of the task.

The engineering protocol for obtaining the architecture-specific constant $\|\mathcal{D}\|$ is defined as **Warmup Calibration**. We invoke the **Equipartition Theorem** from statistical mechanics [45], assuming that during the initial training phase (typically the first epoch), the system operates at peak informational efficiency. In this quasi-equilibrium state, the system's logical budget is partitioned equally between internal reconfiguration and external expression, yielding a baseline velocity of $v = 0.5$. By measuring the average raw trajectories of dR and dS_{ext} during this period, we calibrate the resolution unit of the substrate [10]:

$$\|\mathcal{D}\| = \frac{\text{Avg}(dS_{ext}/\mathcal{L})}{\text{Avg}(dR)} \tag{18}$$

By anchoring $\|\mathcal{D}\|$ to this initial state, we provide a consistent physical metric that allows the agent to perceive its proximity to the computational wall throughout the evolutionary process. The specific calibration algorithm adapts to the implementation tier changes described in Section 5.3.

5.3 Implementation Tiers for Dynamic Measurement

In practical engineering, the computational overhead of measuring system dynamics must be balanced against the required precision of regulation. We propose three implementation tiers for calculating the **Velocity (v)** on the R-S Manifold, allowing for a flexible trade-off between monitoring cost and regulatory fidelity [46].

Tier 1: Minimalist (Scalar-Based)

This tier is designed for resource-constrained environments or low-precision tasks where secondary "probe" passes are not feasible. It approximates the

system's velocity using only the current **Loss value ($\mathcal{L}$)** and the calibrated granularity constant:

$$v = \rho \approx \frac{1}{1 + \frac{1}{\mathcal{L} \cdot \|\mathcal{D}\|}} \tag{19}$$

Engineering Logic: This model assumes that in a stable optimization state, the rule-change effort is roughly proportional to the expected informational gain. Under this assumption, $\mathcal{L}$ becomes the primary driver of the **relativistic brake**. High loss values automatically signal a sparse information environment where rules lack meaningful state-anchors, driving $v \to 1$ and triggering protective deceleration.

Tier 2: Intermediate (Causal Ripple)

The standard implementation tier provides a balanced trade-off by explicitly measuring the environment's response to structural changes through the "Causal Ripple" (dS_{ext}):

$$v = \rho \approx \frac{dR}{dS_R + \frac{dS_{ext}}{\mathcal{L} \cdot \|\mathcal{D}\|}} \tag{20}$$

Engineering Logic: By incorporating the measurable external gain dS_{ext}, the regulator can distinguish between productive learning and unproductive "thrashing"—where large parameter shifts (dR) fail to produce coherent state-transitions. This capability is crucial for preventing the destruction of established causal rules by high-energy stochastic noise.

Tier 3: Full-Spectrum (Disorder-Aware)

The most rigorous implementation, recommended for high-stakes fine-tuning or training in highly volatile environments, utilizes full chaotic loss correction and dimensional scaling:

$$v = \rho = \frac{dR \cdot L_R}{dS_R \cdot L_R + \frac{dS_{ext}}{\mathcal{L} \cdot L_S \cdot \|\mathcal{D}\|}} \tag{21}$$

Engineering Logic: This tier explicitly accounts for internal and external "friction" through the **Disorder Coefficients** (L_R and L_S), which quantify the sequestration of logical action established in Section 4:

- **Rule Disorder (L_R):** Defined as the ratio of the element-wise absolute path to the net vector displacement ($\| \sum |\Delta\theta| \| / \| \sum \Delta\theta \|$). High L_R indicates the model is vibrating intensely in parameter space without achieving net structural advancement. Since internal chaos increases the effective rule-density, both dR and dS_R are multiplied by L_R.
- **State Disorder (L_S):** Represents the loss of coherence in output expressions. A high L_S signifies that environmental feedback is blurred by noise, effectively dividing and diluting the useful external gain dS_{ext}.

Tier 3 is uniquely capable of detecting optimization pathologies, such as "vibrating in place," where high parameter-space volatility ($L_R \uparrow$) signals that the agent's logical action is being entirely consumed by internal friction. By standardizing these metrics through $\|\mathcal{D}\|$ and scaling them by the effective information density ($\mathcal{L}^{-1}$), Tier 3 provides the most precise assessment of the system's proximity to the **computational wall**.

5.4 Inertia-Aware Regulation and Learning Rate Contraction

Once the system velocity v is measured, it functions as the primary feedback signal for a protective regulation protocol. The core of this mechanism is the **Relativistic Contraction of the Learning Rate**, which ensures that the agent's evolutionary tempo remains synchronized with its internal physical limits.

The Physical Nature of Learning Rate and Entropy Expulsion

In the engineering realization of intelligence dynamics, the learning rate η is formally identified as the projection of the system's **Characteristic Cycle (l)**—the internal logical clock—onto the training epoch timeline. As established in our micro-physical model (Section 3), l defines the logical window required to maintain causal consistency during structural evolution. Consequently, η represents the characteristic scale at which the agent samples, filters, and integrates environmental information into its rule-set [47].

Physically, this integration process is governed by the system's capacity for **entropy expulsion**. Unlike traditional Information Bottleneck theories that focus solely on the reduction of internal representation entropy [48], our framework addresses the **dissipation** of that entropy into the environment. The erasure of information (learning) requires internal entropy reduction to be dissipated as heat through the diathermal boundaries of the logical container. In a neural network, this "informational heat" manifests as the stochastic noise and residual error generated during rule-reconfiguration.

The constraint arises from the **geometric occlusion of the cooling channels**: as the **Rule Density (ρ)** increases, the proportion of "adiabatic" surface area—dedicated to internal rule-maintenance—grows, causing the effective heat-exchange cross-section for entropy expulsion to shrink according to $\sqrt{1-v^2}$. Just as the particles in Section 3 require a higher external intensity to find a rare diathermal exit, a high-velocity (dense) neural network faces a **"clogging effect"** where the channels for informational noise are nearly sealed. If the learning rate η—the rate at which new structural changes are forced—exceeds this shrinking expulsion capacity, the residual entropy cannot be dissipated. This accumulated "heat" triggers chaotic fluctuations that physically **"shatter"** existing causal rules, providing a first-principles explanation for catastrophic forgetting and the explosive instability observed in dense models [12, 24].

The Relativistic Brake

To protect the agent's structural integrity, the effective learning rate must contract to match the system's real-time entropy expulsion limit. Following the Lorentz symmetry derived on the R-S Manifold, we define the **Effective Learning Rate (η_{eff})** as:

$$\eta_{\text{eff}} = \eta_{\text{base}} \cdot \frac{\sqrt{1 - v^2}}{\sqrt{1 - v_{\text{base}}^2}} \tag{22}$$

where η_{base} is the standard step-size determined by the optimizer, and v_{base} represents the velocity recorded during the warmup calibration in Section 5.2 (typically $v \approx 0.5$).

- **Logical Freezing:** As the velocity v approaches the saturation limit of 1.0, the term $\sqrt{1 - v^2}$ converges to zero, causing the effective learning rate to vanish. This induces a state of **"Logical Freezing,"** a self-preservation mode where the system locks its current parameters to prevent reconfigurations that would be impossible to dissipate or interpret.
- **Dynamic Adaptation:** Unlike heuristic decay schedules such as cosine annealing [49], this contraction is a direct physical response to the agent's internal **Inertia Expansion**. When the model encounters high-inertia data that drives $v \to 1$, the system automatically decelerates to prevent structural collapse.

By implementing this "Inertia-Aware" brake, we guarantee that the agent's trajectory through parameter space remains within the stable bounds of the Minkowski manifold, ensuring that rule-changes remain anchored to the system's interpretability scale $\mathcal{D}$.

5.5 Directional Coherence and Phase Alignment

While the magnitude of velocity v determines the required Lorentzian contraction of the learning rate, its **direction**—defined as the orientation of the velocity vector $\vec{v}$ within the tangent space of the R-S Manifold—dictates the geometric validity of the structural evolution [50].

The Physical Meaning of Velocity Direction

In the phase space of an agent, the orientation of $\vec{v}$ represents the **Logical Orientation** of reconfiguration, identifying which cognitive pathways or rule-subsets are being prioritized for modification.

- **Directional Coherence**: When the current trajectory aligns with historical pathways that have demonstrated stable entropy expulsion, the system performs a "coherent" update that preserves the agent's established topological structure [51].

- **Directional Dissonance**: An abrupt shift in orientation (e.g., updates becoming orthogonal to established gradients) indicates a phase mismatch. This suggests the system is forcing a reconfiguration that contradicts its **ontological support**, leading to a high risk of **structural shattering**.

Coherent State Anchors and Component Metrics

To quantify this alignment, the regulator utilizes **Coherent State Anchors**—reference vectors captured during high-efficiency learning phases where entropy expulsion was optimal. We define the **Phase Coherence Score (C)** by evaluating the similarity between the current dynamics and these anchors across three key dimensions:

1. **Internal Directional Alignment (C_{S_R})**: The cosine similarity between the current internal state displacement vector and the anchor's reference displacement:

$$C_{S_R} = \cos(\vec{dS}_{R,curr}, \vec{dS}_{R,anchor}) = \frac{\vec{dS}_{R,curr} \cdot \vec{dS}_{R,anchor}}{\|\vec{dS}_{R,curr}\|\|\vec{dS}_{R,anchor}\|} \tag{23}$$

2. **External Response Alignment ($C_{S_{ext}}$)**: The cosine similarity between the current external gain vector (causal ripple) and the anchor's reference gain:

$$C_{S_{ext}} = \cos(\vec{dS}_{ext,curr}, \vec{dS}_{ext,anchor}) \tag{24}$$

3. **Efficiency Ratio Consistency (C_ϕ)**: This metric evaluates whether the "gearing" of the update matches the anchor's optimal efficiency. Let $r = \|\vec{dS}_{ext}\|/\|\vec{dS}_R\|$. We define the ratio consistency coefficient k_{ratio} as:

$$k_{ratio} = \frac{\min(r_{curr}, r_{anchor})}{\max(r_{curr}, r_{anchor})} \tag{25}$$

The resulting consistency score is scaled via a sine function to ensure a smooth penalty:

$$C_\phi = \sin\left(\frac{\pi}{2} \cdot k_{ratio}\right) \tag{26}$$

The total **Phase Coherence Score (C)** is derived from the product of these metrics:

$$C = C_{S_R} \cdot C_{S_{ext}} \cdot C_\phi \tag{27}$$

This score acts as a final multiplicative gate for the learning rate, ensuring that updates are only permitted when the system is both below the velocity saturation limit and geometrically aligned with a stable history:

$$\eta_{final} = \eta_{eff} \cdot C \tag{28}$$

It should be noted that while this specific vector-calculus framework offers a robust engineering realization, it is **not the exclusive method** for implementing intelligence inertia dynamics. Besides, Several high-dimensional phenomena

must be addressed for practical deployment. First, the **intrinsic sparsity** of neural gradients can bias directional metrics; this is mitigated by performing alignment in effective subspaces or utilizing projection masks. Second, **stochastic noise** becomes dominant as $v \to 1$ and the external signal dS_{ext} diminishes, necessitating low-pass filtering or sliding-window integration to distinguish structural density from transient fluctuations. Finally, to address the **stability-plasticity dilemma** [52], a **logical unfreezing mechanism**—facilitated by a spontaneous exponential decay of measured inertia—is required to release the relativistic brake once the core structure has stabilized, allowing the agent to recover the plasticity necessary for new information.

6 Experiments

In this chapter, we subject the derived **physical principles** of *intelligence inertia* to a series of rigorous empirical investigations. Transitioning from the micro-physical derivations in Section 3, the formal axiomatization in Section 4, and the engineering realization in Section 5, we demonstrate how these abstract laws manifest within the actual tensor dynamics of deep neural networks [53]. The experimental suite utilizes established architectures, such as ResNet-18 [54], and standard benchmarks like CIFAR-10 [55], and is organized into three logically advancing stages designed to validate the framework's predictive power and engineering utility:

1. **Experiment I: Decisive Adjudication of Intelligence Inertia Divergence.** This stage aims to confirm the physical reality of the **Informational Speed Limit** within the **R-S Manifold** and the resulting **Inertia Expansion** effect. By simulating high-velocity regimes characterized by a vanishing external gain ($dS_{\text{ext}} \to 0$), we observe whether computational work follows the non-linear, relativistic trajectories predicted by our framework, thereby establishing the empirical bedrock of the theory.

2. **Experiment II: Evolutionary Geometry and the Reachability Topography.** We investigate the fundamental impact of architectural optimization on system inertia. By mapping the performance landscape across various neural topologies within the phase space of the **R-S Manifold**, we verify that maintaining a balanced velocity of $v \approx 0.5$—the "golden axis" of energy equipartition—serves as the steepest descent path for intelligent evolution, providing a principled physical methodology for architectural design.

3. **Experiment III: Engineering Practice — The Inertia-Aware Scheduler Wrapper.** We deploy the practical **Inertia-Aware Scheduler Wrapper** based on the implementation protocols defined in Section 5. Through evaluations of dynamic performance, resilience against high-entropy logic shocks (noise), and memory retention in continual learning [12], we

demonstrate the superior efficiency and stability of systems that respect their intrinsic physical resistance to change.

6.1 Experiment I: Decisive Adjudication of Intelligence Inertia Divergence

The primary objective of this experiment is to provide a decisive empirical test between classical information-geometric models and our relativistic framework of **Intelligence Inertia**. By subjecting a neural network to extreme logical stress, we aim to observe whether the computational work required for adaptation remains a quadratic function of velocity—as predicted by the **Fisher Information Matrix (FIM)** baseline—or whether it exhibits the non-linear **relativistic J-shaped inflation curve** characteristic of Inertia Expansion.

6.1.1 Experimental Design and Physical Mapping

To empirically validate the existence of the "computational wall," we constructed an environment that forces an agent through a spectrum of rule densities, ranging from the low-velocity "clean data" regime to the high-velocity "informational limit." This was achieved by systematically injecting label noise (ranging from 0% to 100%) into the CIFAR-10 dataset.

From a physical perspective, the injection of noise serves to suppress the External Gain ($dS_{\text{ext}} \to 0$), as random labels provide no coherent structure for the model to project onto the state manifold. To minimize the training objective, the system is forced to undergo intense **Internal Reconfiguration** ($dS_R \to \infty$) to memorize the noise, thereby pushing the rule density ($v = \rho$) toward its saturation limit. We utilized the ResNet-18 architecture with 11.2M (Million parameters) as our physical substrate, recording the total **computational work (W)**, measured in epochs, required to reach a specific convergence threshold across different noise energy levels.

The dynamical parameters are measured according to the engineering protocols established in Tier 2, Section 5.3. The adjudication logic is straightforward: If the FIM framework is exhaustive, the computational cost should follow a smooth, quadratic growth (v^2). If the Intelligence Inertia framework is correct, the cost must follow a Lorentzian divergence as v approaches the informational speed limit of $c = 1$.

6.1.2 Dynamical Models and Regression Basis

To adjudicate the empirical results, we establish two competing mathematical models for regression analysis:

1. **Classical Dynamical Hypothesis (FIM Baseline):** This model assumes that the learning cost is a direct function of the local manifold curvature, following a second-order Taylor expansion analogous to Newtonian kinetic energy:

$$Cost_{classical} = k \cdot v_{rel}^2 + b \tag{29}$$

2. **Relativistic Dynamical Hypothesis (Intelligence Inertia Model):** This model assumes that the system's **Effective Mass** undergoes a geometric expansion as velocity approaches the informational horizon (where $c = \rho_{\max} = 1$):

$$Cost_{relativistic} = k \cdot (\gamma - 1) + b = k \cdot \left(\frac{1}{\sqrt{1 - (v_{rel}/c)^2}} - 1 \right) + b \tag{30}$$

By fitting these two equations to the observed computational expenditure across the noise-injected velocity spectrum, we can determine whether the agent's resistance to change is a static property of curvature or a dynamic consequence of relativistic mass divergence.

6.1.3 Experimental Results and Attribution Analysis

Following the execution of our controlled noise-injection protocol, we evaluated the predictive performance of the competing dynamical models. The resulting data provides a stark contrast between the local approximations of classical information geometry and the global consistency of the **Intelligence Inertia** framework. The quantitative results of this adjudication are summarized in Table 1.

Table 1: Dynamical Model Fitting Performance across Coordinate Systems. The Intelligence Inertia Theory, utilizing relativistic mass expansion, consistently outperforms classical Fisher Information approximations, particularly as the system approaches the informational speed limit.

Dynamical Model	**Reference Frame Treatment**	**Fit Error (RMSE)**	**Performance**
Classical FIM	Absolute Coordinates	36.0	**Failed**: Incapable of capturing the non-linear "J-Curve" divergence.
Classical FIM	Galilean Shift $(v - v_0)$	30.0	**Limited**: Fits the low-speed regime but severely underestimates the wall effect.
Hybrid FIM	Lorentz Transformation	25.5	**Suboptimal**: Corrects for velocity addition but lacks mass expansion logic.
Relativistic Mass (Intelligence Inertia)	**Universal Covariance**	**18.5 – 19.6**	**Optimal**: Precisely fits the asymptotic divergence and exhibits frame independence.

Experimental Conclusion: Experiment I definitively confirms the existence of a fundamental **Rule-Density Limit** in intelligent agents, functioning

as an informational "speed of light" ($c \approx 1$). This validates the micro-physical derivation in Section 3, proving that the resistance to change is not merely a consequence of manifold curvature but a result of the system's logical budget reaching saturation. While the classical Fisher Information Matrix (FIM) serves as a reasonable second-order approximation in low-velocity states, **Inertia Expansion** becomes the dominant physical factor governing computational work as the system nears its cognitive horizon.

To further isolate the causative mechanisms, we performed a structural attribution analysis using a four-dimensional arena comparison [56]. This methodology allows us to rigorously adjudicate between competing hypotheses by testing the necessity of mass expansion and the robustness of the theory across different observational reference frames.

1. Arena I & II: Reference Frame Sensitivity and Model Robustness The first stage of our attribution analysis investigates the dependency of each model on the selection of the coordinate origin—specifically, whether the model's validity relies on the arbitrary subtraction of the system's initial "rest velocity" (v_0). This tests whether the observed dynamics are a fundamental feature of the R-S Manifold or merely a consequence of specific observational framing.

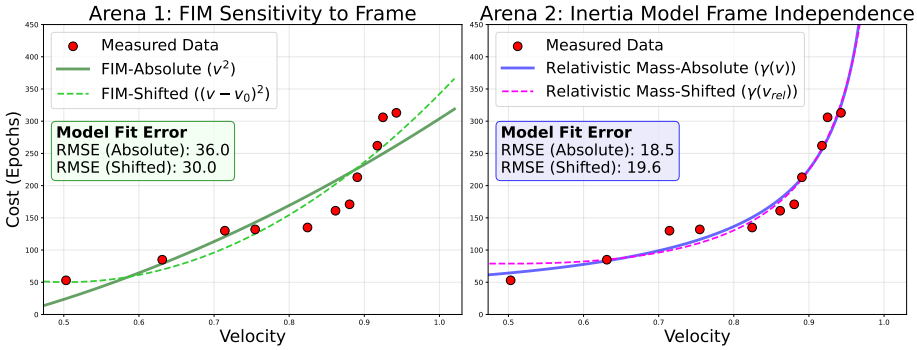

Figure 2: Comparative Adjudication of Reference Frame Sensitivity and Model Robustness. Arena 1 **(Left)** demonstrates the sensitivity of the classical FIM model to coordinate shifts; the model fails to track the data trend in absolute coordinates. Arena 2 **(Right)** illustrates the Intelligence Inertia model, where the relativistic curves remain covariant and highly accurate regardless of the reference frame, maintaining an RMSE between 18.5 and 19.6.

- **Reference Frame Dependency of FIM (Arena 1)**: In the **Absolute Reference Frame** (without subtracting v_0), the FIM prediction fails catastrophically (RMSE = 36.0), appearing as a linear approximation that cannot capture the rising work-divergence. After applying a **Galilean Transformation** to account for v_0 (FIM-Shifted), the error drops to 30.0. However, while this improves the fit in the low-velocity regime, it remains

blind to the explosive **Inertia Expansion** encountered when $v > 0.9$. This confirms that FIM acts only as a local, second-order approximation valid at the origin [11].

- **Lorentz Covariance of Intelligence Inertia Model (Arena 2)**: The relativistic framework demonstrates profound **Frame Independence**. Unlike classical approximations, the predictive law exhibits **mathematical covariance**; whether evaluated in absolute or shifted coordinates, the model consistently captures the intrinsic hyperbolic geometry of the **Inertia Expansion** (RMSE 18.5–19.6). This confirms that the observed "Computational Wall" is an **invariant property** of the R-S Manifold rather than an artifact of coordinate selection [57].

Physical Conclusion: These results establish that **Intelligence Inertia (μ)** is not a mere empirical heuristic dependent on measurement methods, but an **intrinsic physical law** that describes the fundamental cost of structural evolution in intelligent systems.

2. Arena III & IV: Necessity of Velocity Transformation vs. Mass Expansion The second stage of our analysis utilizes an "ablation study" approach to disentangle the specific contributions of relativistic kinematics (velocity definitions) from the dynamics of **Inertia Expansion**. This comparison determines whether the observed non-linearity is a result of coordinate transformation logic or a fundamental shift in the system's physical resistance as it approaches the **Interpretability Criterion** limit ($ds^2 \to 0$).

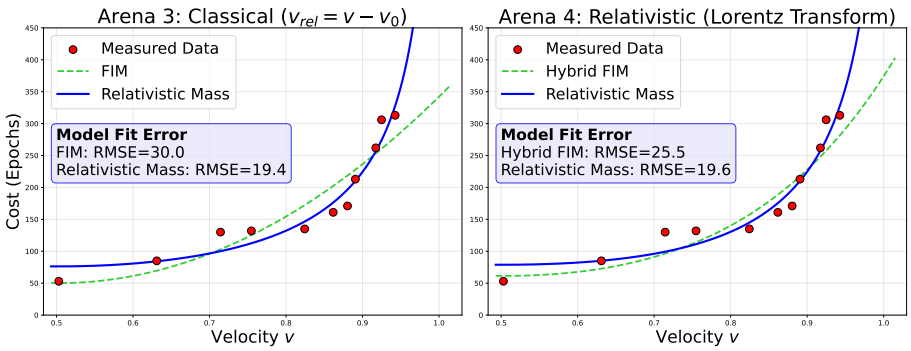

Figure 3: **Ablation Analysis of Relativistic Velocity Addition vs. Mass Expansion.** Arena 3 **(Left)** contrasts the Galilean-shifted FIM against the relativistic mass model, showing the clear failure of the quadratic assumption at high speeds; Arena 4 **(Right)** introduces a "Hybrid FIM" model, which applies the relativistic Lorentz velocity transformation but retains the classical quadratic cost formula, highlighting that velocity correction alone is insufficient to explain the "J-Curve" inflation.

- **Limitations of Velocity Transformation (Hybrid FIM):** To test if the error stemmed solely from velocity measurement, we constructed a **Hybrid FIM** model. This model utilizes the relativistic velocity addition rule to correct the system's speed but maintains the classical quadratic work function. The results indicate that even with the corrected velocity, the Hybrid FIM (**RMSE = 25.5**) remains incapable of explaining the vertical surge in computational work at the high-velocity limit. This suggests that the non-linearity is not a kinematic artifact but a dynamical property of sequestered energy.

- **Decisiveness of Inertia Expansion:** The complete Intelligence Inertia model, which fully incorporates the γ factor (**RMSE = 19.6**), precisely captures the "computational wall" encountered as the agent approaches the informational speed limit. This confirms that the sequestration of logical action leads to an actual inflation of the system's effective mass.

Physical Conclusion: These results provide a robust empirical refutation of the hypothesis that observed cost anomalies are merely artifacts of linear velocity errors. The divergence of computational work at the reachability boundary is fundamentally caused by the **relativistic divergence of the system's effective mass** during the process of intensive Internal Reconfiguration.

6.2 Experiment II: Evolutionary Geometry and the Reachability Topography

The empirical confirmation of the Inertia Expansion effect in Experiment 6.1 establishes that the effective mass of an agent diverges as its velocity v approaches the informational speed limit. This physical reality shifts the focus of architectural engineering from merely increasing parameter depth to **minimizing the work performed against internal inertia** during the learning process. Within this stage of our inquiry, we map the "Reachability Topography" of neural architectures on the R-S Manifold to determine how topological choices influence the coordination between Internal Reconfiguration (dS_R) and environmental feedback.

6.2.1 Experimental Objective: From Inertia Evasion to Dynamical Optimization

While traditional theories based on **Hausdorff dimensions** suggest that increasing architectural complexity should exponentially reduce learning costs [58], the **Intelligence Inertia Theory** posits that efficiency is not a simple function of dimensionality [59]. Instead, it is constrained by the coordination between the internal state shift driven by rule reconfiguration (dS_R) and the external state gain provided by the environment (dS_{ext}), as governed by the local velocity equation:

$$v = \rho = \frac{dR}{dS_R + dS_{ext}} \tag{31}$$

This framework suggests that architectural design is essentially a problem of **dynamical balancing**. The objective of this experiment is to identify the optimal "geodesic" for structural progress by testing three specific predictions:

1. **Limitations of Single-Axis Optimization**: Improving an architecture along only one dimension—either by smoothing the internal manifold (optimizing dS_R, e.g., via residual connections [54]) or by increasing inductive biases (optimizing dS_{ext}, e.g., via multi-scale features)—will fail to achieve global energy minimality. Such systems inevitably encounter a bottleneck of diminishing returns as their Rule Density deviates from the optimal regime.

2. **Orthogonal Synchronous Evolution**: We hypothesize that the most efficient evolutionary path requires internal effort and external feedback to be optimized in synchronization. This keeps the system velocity dynamically anchored near $v \approx 0.5$, the point of **Energy Equipartition** [45], where the resistance to structural change is minimized.

3. **Saddle-Shaped Decay Surface**: Macroscopically, while learning costs decay as architectural rules improve, this decay is non-linear. We predict the cost surface on the R-S Manifold will exhibit a characteristic **Saddle** geometry [50], where the steepest descent is achieved through a "Zig-Zag" path that alternates between rule-refinement and state-expansion.

6.2.2 Experimental Design and Quantitative Measurement

To isolate the impact of architectural topology on system inertia, we strictly constrained the parameter budget of all candidate architectures to 5.0M and fixed the network depth to 10 layers. Within this framework, we utilize the **Reachability Limit ($\mathcal{L}_{min}$)**—defined as the minimum achievable loss on the CIFAR-10 task [55]—as the primary indicator of the physical "work cost" required to reach a specific intelligence state. Total floating-point operations (TFLOPs) serve as a secondary measure of the actual computational energy expended during the process.

We identify two orthogonal dimensions of architectural evolution on the **R-S Manifold** that govern the system's dynamical state:

1. **The Internal Rule-Reconfiguration (dS_R) Axis**: This axis represents architectural enhancements designed to smooth the internal parameter manifold and reduce frictional losses during structural updates. The progression includes the transition from **Baseline (R0)** to **Batch Normalization (BN, R1)** [60] and **Residual Connections (Res, R2)**, aiming to minimize the internal effort required for a given logic shift.

2. **The External State Gain (dS_{ext}) Axis**: This axis represents the integration of inductive biases [61] that allow the system to effectively align with the environmental manifold. Improvements progress from **Baseline (S0)** to **Locality-based Convolutional Neural Networks (CNN,**

S1) [62] and **Multi-scale Convolutional Neural Networks (MCNN, S2)** [63], increasing the productive state-shift obtained from each rule-update.

The Multi-Layer Perceptron (MLP) [64] is established as our **Evolutionary Origin (R0, S0)**. As a structure with minimal inductive bias and baseline connectivity, it represents a state of "maximal ignorance" within the theory. This origin is used to calibrate the **Symbolic Granularity ($\mathcal{D}$)**, anchoring the system's initial velocity at the point of Energy Equipartition ($v \approx 0.5$).

The core dynamical parameters captured across the 3×3 architectural matrix are presented in Table 2.

6.2.3 Results Analysis: Reachability Topography and the Valley of Equilibrium

The experimental results from the 3×3 architecture matrix reveal a profound geometric structure in the cost landscape of neural evolution. By plotting the Reachability Limit ($\mathcal{L}_{min}$) against the orthogonal axes of Internal Reconfiguration (dS_R) and External State Gain (dS_{ext}) on the R-S Manifold, we observe the emergence of a characteristic **Saddle Topography** [65].

1. The Saddle Geometry of Reachability Figure 4 visualizes the minimum achievable loss across the evolutionary surface, illustrating how different structural combinations influence the agent's ultimate intelligence capacity.

The analysis of this topography yields three critical insights into the dynamics of intelligence:

- **Plateaus in Axial Trajectories**: Observing the **MLP** $\rightarrow$ **Res** path (pure dS_R optimization), we find that while manifold smoothing provides an initial gain, $\mathcal{L}_{min}$ quickly hits a plateau, decreasing only from 2.302 to 1.322 before the slope flattens. A similar diminishing return is observed in the **MLP** $\rightarrow$ **MLP-MCNN** path (pure dS_{ext} optimization). This confirms that optimizing only one dimension of the Rule-State pair—regardless of the sophistication of the technique—leads the system to deviate from the golden axis, resulting in an **Inertia Expansion** that halts further progress.

- **Synergistic Jumps via Orthogonal Transitions**: The most significant performance leaps occur at orthogonal switching points, such as the transition from **MLP-CNN to BN-CNN**. These "jumps" indicate that when a system is stalled on an axial plateau, the introduction of a complementary structural dimension restores the dynamical balance, allowing for a rapid collapse in total inertia and a corresponding drop in reachable loss.

- **The Saddle Geometry and the Geodesic**: The topography clearly demonstrates that the "bottom" of the informational valley follows a diagonal trajectory toward the (2,2) coordinate. This **Zig-Zag Geodesic**

Table 2: Comparative Analysis of Dynamical Parameters and Reachability Limits across Neural Architectures. Abbreviations for architectures: MLP (Multi-Layer Perceptron), BN (Batch Normalization), Res (Residual), CNN (Convolutional Neural Network), MCNN (Multi-scale Convolutional Neural Network). Column Abbreviations: Rule Reconfig. (Internal Rule-Reconfiguration), Ext. Gain (External State Gain), Vel. (Velocity), Vel. Dev. (Velocity Deviation from 0.5), Reach. Limit (Reachability Limit $\mathcal{L}_{min}$), Work (Computational Effort in TFLOPs). The data shows that the lowest reachability limit is attained by architectures like Res-MCNN, which achieve optimal balance between rule effort and state feedback, minimizing the deviation from the $v \approx 0.5$ golden axis.

Arch. Name	**Rule Reconfig. (dS_R Axis)**	**Ext. Gain (dS_{ext} Axis)**	**Vel. (v)**	**Vel. Dev. $\lvert v-0.5\rvert$**	**Reach. Limit ($\mathcal{L}_{min}$)**	**Work (TFLOPs)**
MLP (Origin)	Baseline (R0)	Baseline (S0)	0.502	0.002	2.302	0.6
BN	Normalized (R1)	Baseline (S0)	0.522	0.022	1.389	68.7
Res	Residual (R2)	Baseline (S0)	0.818	0.318	1.322	117.0
MLP-CNN	Baseline (R0)	Locality (S1)	0.179	0.321	1.485	830.0
MLP-MCNN	Baseline (R0)	Multi-scale (S2)	0.194	0.306	1.339	3670.0
BN-CNN	Normalized (R1)	Locality (S1)	0.532	0.032	0.647	664.0
Res-CNN	Residual (R2)	Locality (S1)	0.689	0.189	0.578	863.0
BN-MCNN	Normalized (R1)	Multi-scale (S2)	0.531	0.031	0.599	2480.0
Res-MCNN	Residual (R2)	Multi-scale (S2)	0.452	0.048	**0.553**	2090.0

represents the sequence of modifications where internal effort and external feedback remain in constant proportion. This validates the **Law of Orthogonal Synergy**: architectural progress is maximized only when rule reconfiguration and state expression are improved in synchronization, keeping the system velocity dynamically anchored within the **Valley of Equilibrium**.

2. The Valley of Dynamic Equilibrium To identify the underlying optimality of the Zig-Zag path, we analyzed the **Velocity Deviation ($|v - 0.5|$)**

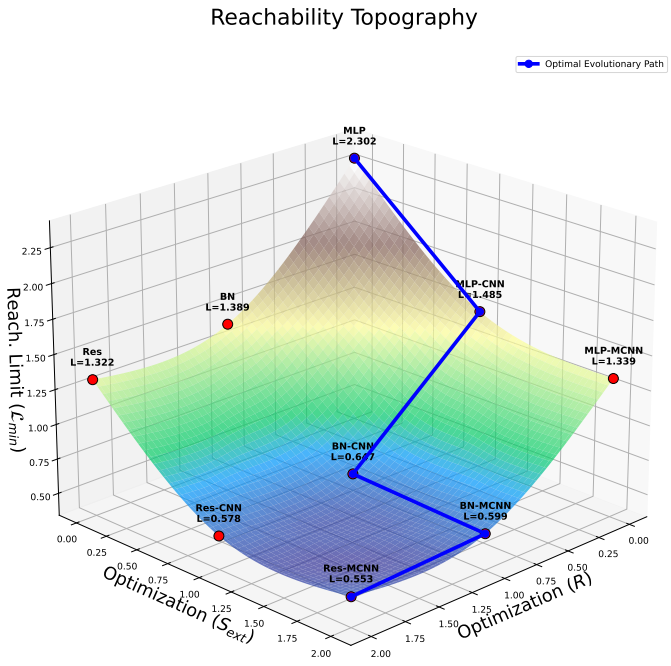

Figure 4: 3D Reachability Topography and the Zig-Zag Evolutionary Geodesic. This 3D surface plots the Reachability Limit ($\mathcal{L}_{min}$) as a function of improvements in internal manifold smoothing (dS_R-axis) and external structural bias (dS_{ext}-axis). The red nodes represent measured architectures from our matrix, while the blue line illustrates the "Zig-Zag Geodesic," representing the path of maximum efficiency. The topography exhibits a distinct saddle shape, where the steepest descent occurs along the diagonal of balanced development.

across the architectural landscape, representing the system's distance from the theoretical point of **Energy Equipartition** [45].

- **The Dynamical Riverbed**: Figure 5 reveals a striking "Riverbed" geometry on the R-S Manifold. We observe that the observed optimal evolutionary trajectory **approximates the trough of this valley**. This confirms that superior architectural reachability is achieved not by singular scaling, but by the continuous balancing of Internal Reconfiguration and External State Gain, effectively maintaining the system near the state of energy equipartition where resistance to change is minimized.
- **Efficiency Adjudication**: The topography explains the suboptimal performance of unbalanced architectures. MLP-CNN ($v = 0.179$) represents an "environmental stall" where high external state gain without sufficient

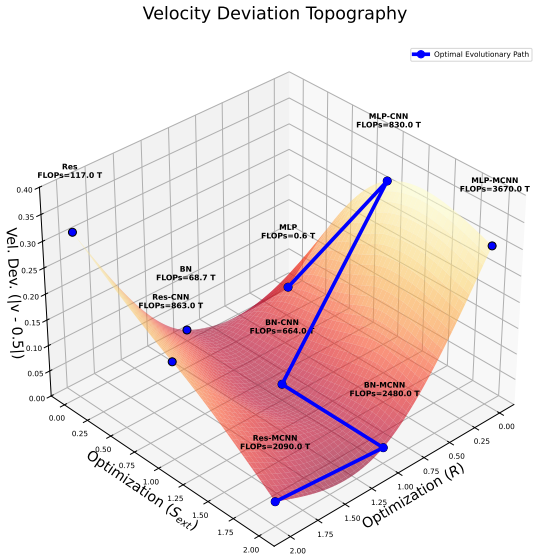

Figure 5: Velocity Deviation Topography and the Dynamical Riverbed. This 3D surface maps the absolute deviation of system velocity from the 0.5 golden axis. The resulting "V-shaped" valley—the Dynamical Riverbed—approximately aligns with the optimal evolutionary path, identifying the minimization of velocity deviation as the physical objective of architectural progress.

internal rule-reconfiguration capacity drives the system toward the $v \to 0$ limit. Conversely, the Res architecture ($v = 0.818$) exhibits an Inertia Expansion, where excessive rule-flexibility without sufficient environmental alignment pushes the system toward the relativistic $v \to 1$ wall. Architectures positioned near the base of the Riverbed maintain the dynamical equilibrium required for efficient learning [66].

- **Geodesic Principle**: These results suggest that architectural evolution is fundamentally a search for the geodesic—the trajectory that seeks to minimize the velocity deviation $|v-0.5|$ to overcome the system's intrinsic Intelligence Inertia.

6.2.4 Summary and Guidelines for Architectural Evolution

Experiment II demonstrates that **Intelligence Inertia** is a topological property of neural networks. Even with a constant parameter count, an optimized topology like Res-MCNN can reduce the reachability residual by a factor of four compared to the baseline MLP. Our analysis establishes the **Methodology of Bipedal Evolution**: architectural progress must alternate between optimizing **Internal Reconfiguration** (dS_R) and improving **External State Gain** (dS_{ext}). Any single-axis improvement will inevitably cause the velocity

to drift from the 0.5 axis, leading to a collision with either the environmental or relativistic "walls."

6.3 Experiment III: Engineering Practice — The Inertia-Aware Scheduler Wrapper

Experiment 6.2 established that architectural progress is macroscopically driven by a trajectory toward the energy equipartition point of $v \approx 0.5$. This topological finding raises a critical microscopic hypothesis: if structural evolution inherently favors this balanced velocity, can the optimization process be enhanced by dynamically scaling the Internal Reconfiguration (dS_R) based on the **Lorentzian contraction** of v in real-time? To investigate this, we implemented the **Inertia-Aware Scheduler Wrapper**. This engineering tool functions as a collaborative regulatory layer, nesting atop standard learning rate policies to align microscopic training dynamics with the physical principles of Intelligence Inertia [67].

The experiment evaluates the practical utility of the Scheduler wrapper across three distinct scenarios:

1. **Convergence Limit Test**: The Scheduler wrapper is deployed as a "physical enhancement plugin" atop eight mainstream learning rate schedulers in a nested configuration. This quantifies its universal effectiveness in accelerating convergence and compressing the **Reachability Limit** ($\mathcal{L}_{min}$).

2. **Noise Shock Resilience**: In a standalone configuration, we evaluate the Scheduler Wrapper's capacity to protect the agent's structural integrity against high-entropy logic shocks, utilizing baseline algorithms only to provide a terminal learning rate boundary.

3. **Continual Learning Stability**: We test the Scheduler Wrapper's ability to prevent **logical shattering** during abrupt task transitions without relying on replay buffers or external data caches.

6.3.1 Core Control Logic of the Wrapper

The Scheduler Wrapper performs a real-time audit of the agent's evolutionary trajectory on the R-S Manifold by implementing the engineering protocols established in Section 5. It operates through three core physical mechanisms:

1. **Full-Spectrum Velocity Measurement**: The Scheduler Wrapper utilizes the **Tier 3 (Disorder-Aware)** measurement protocol defined in Section 5.3. By calculating the instantaneous rule density, it explicitly accounts for internal and external "friction" via the Disorder Coefficients (L_R and L_S). This ensures that the measured velocity v accurately reflects the structural stress on the manifold, distinguishing productive structural advancement from unproductive chaotic vibration.

2. **Directional Coherence Consideration**: Following the principles of **Phase Alignment** established in Section 5.5, the Scheduler wrapper audits the orientation of each velocity vector $\vec{v}$. It identifies the logical orientation of the reconfiguration; updates that align with historical coherent pathways are permitted, while those exhibiting directional dissonance (orthogonal to the established ontological support) are damped to prevent the physical shattering of existing rules.

3. **Multiscale Geometric Coupling**: In the nested operational mode, the Scheduler wrapper utilizes a **weighted geometric mean** to integrate macroscopic scheduling with microscopic inertial protection [68]. The composed step-size $\eta_{composed}$ is determined as:

$$\eta_{composed} = (\eta_{sched})^{1-w} \cdot (\eta_{wrapper})^{w} \tag{32}$$

The weight $w = 0.2$ ensures that the primary scheduler (η_{sched}) dictates the overall energy decay strategy, while the physical inertia constraint ($\eta_{wrapper}$) provides a 20% regulatory influence for fine-grained structural protection.

6.3.2 Sub-experiment I: Dynamic Performance and Reachability Limit Analysis

To evaluate the universal generalizability of the **Inertia-Aware Scheduler Wrapper**, we utilized a ResNet substrate (0.5M) and benchmarked its performance across eight mainstream scheduling algorithms provided by the PyTorch library [69]. The objective of this inquiry is to observe the shift in system dynamics when traditional heuristic policies are replaced by a physical awareness of the system's trajectory on the R-S Manifold. Specifically, we measure the wrapper's ability to accelerate initial learning and compress the final **Reachability Limit ($\mathcal{L}_{min}$)**, effectively lowering the system's thermodynamic floor.

The quantitative results of this benchmark are summarized in Table 3, providing a comparative view of the "Base" versus "Enhanced" configurations.

The evolutionary trajectories of these systems are visualized in Figure 6, contrasting the reachability curves and learning rate behavior between the baseline and inertia-aware agents.

Analysis of Results The results from the convergence limit test reveal the profound impact of **Intelligence Inertia** on neural optimization. Below, we provide a deep attribution analysis of the observed phenomena, categorized by convergence speed, coupling logic, and the interaction between inertia-awareness and traditional heuristics.

1. "Inertial Primacy": The Rapid Convergence of Pure Inertia Group As shown in Table 3, the most striking performance was observed in the **Pure Inertia (Wrapper Only)** group. Despite the total absence of a pre-defined

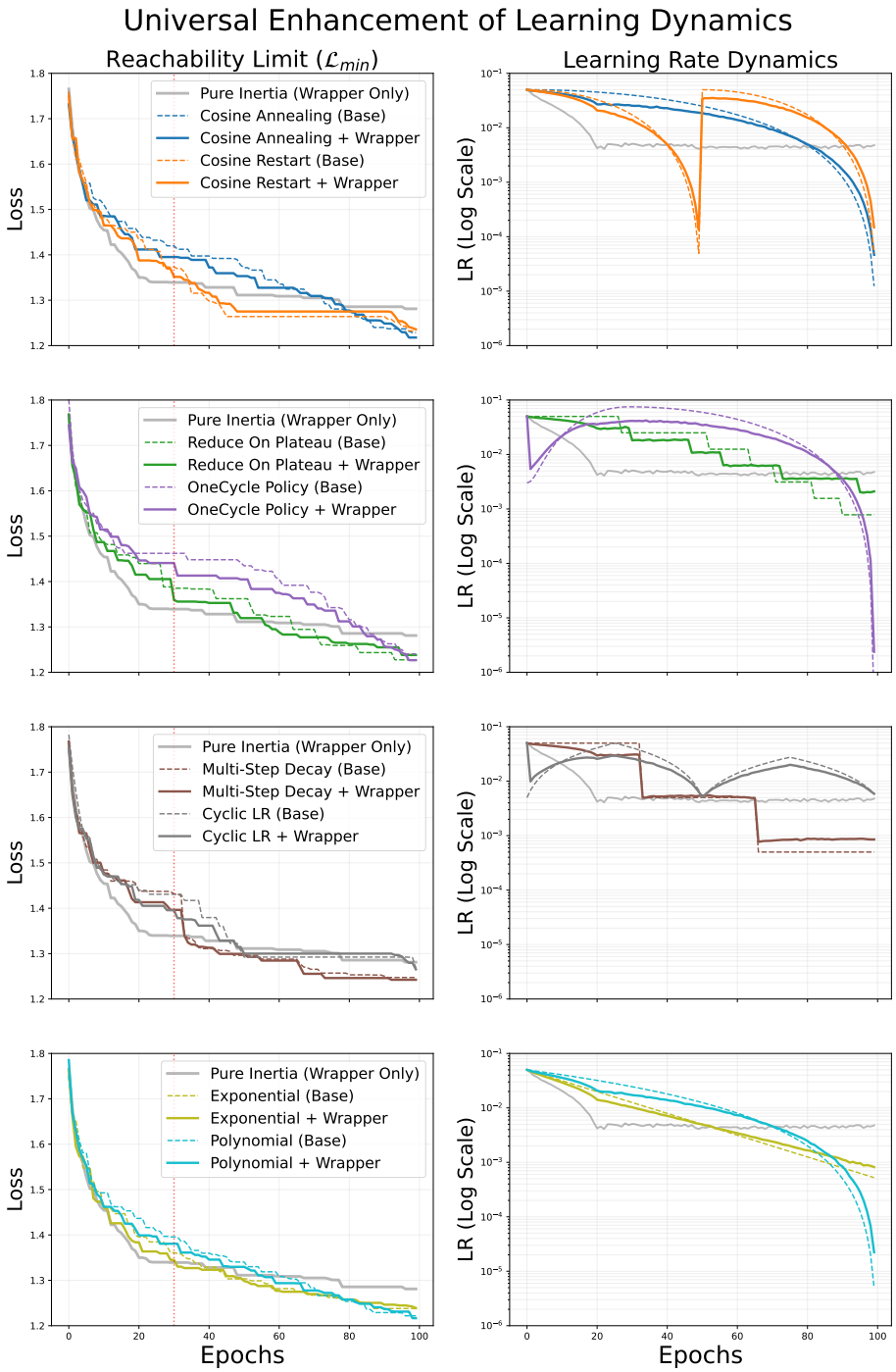

Figure 6: Universal Enhancement of Learning Dynamics via the Inertia-Aware Scheduler Wrapper. The panels contrast validation loss **(Left)** and learning rate behavior **(Right)** across eight schedulers. Solid lines indicate enhanced systems, while dashed lines indicate baselines; the vertical red dotted line marks the critical Epoch 30 checkpoint where dynamical acceleration is most visible.

Table 3: Comparative Statistical Analysis of Inertial Enhancement Across Schedulers. The table provides a quantitative leap in training efficiency and reachability. Sched.: PyTorch [69] base scheduler, including Cosine Annealing and Cosine Restart [49], OneCycle Policy [70], Cyclic LR [71], MultiStep, Polynomial, Exponential, and ReduceLROnPlateau; Cfg.: Operation mode; Prog.@30: Convergence percentage at Epoch 30; Accel.: Improvement in Progress; Gain: Reduction in $\mathcal{L}_{min}$.

Sched.	Cfg.	Loss@30	Prog.@30	Accel.	Reach. $\mathcal{L}_{min}$	Gain	Best Ep.
Pure Inertia	Wrapper Only	**1.3392**	**94.31%**	N/A	1.2811	N/A	96
Cosine Ann.	Base / Enh.	1.419 / 1.395	82.2% / 83.7%	**+1.74%**	1.229 / 1.218	**+0.87%**	98 / 97
OneCycle	Base / Enh.	1.462 / 1.441	79.1% / 80.1%	**+1.46%**	1.240 / 1.227	**+1.10%**	96 / 97
Multi-Step	Base / Enh.	1.431 / 1.396	82.6% / 85.5%	**+2.45%**	1.247 / 1.242	**+0.40%**	91 / 92
Cyclic LR	Base / Enh.	1.431 / 1.396	84.3% / 87.4%	**+2.46%**	1.268 / 1.265	**+0.21%**	99 / 99
Polynomial	Base / Enh.	1.396 / 1.381	83.9% / 84.9%	**+1.08%**	1.222 / **1.217**	**+0.43%**	96 / 98
Exponential	Base / Enh.	1.361 / 1.345	88.4% / 90.0%	**+1.18%**	1.238 / 1.239	-0.09%	97 / 99
Cos. Restart	Base / Enh.	1.373 / 1.352	86.9% / 89.2%	**+1.57%**	1.232 / 1.236	-0.29%	97 / 99
On Plateau	Base / Enh.	1.385 / 1.359	85.3% / 88.6%	**+1.88%**	1.228 / 1.238	-0.86%	99 / 95

human schedule or decay timetable, the system achieved a staggering **94.31%** convergence progress within the first 30 epochs, outperforming all native PyTorch schedulers.

- **Coherence of the R-S Manifold**: This "steepest descent" phenomenon validates the immense power of **respecting the coherence of the R-S Manifold**. The Pure Inertia wrapper functions as a logical filter, identifying and rejecting update components that threaten the established rule-structure. In contrast, traditional schedulers maintain a high level of "blind kinetic energy" during early training [72], consuming significant

computational resources on incoherent rule-oscillations. Consequently, the convergence progress of traditional baselines lags behind the pure inertial group by approximately 10% in the early stages.

2. The Rationale for a Hybrid Strategy While the Pure Inertia configuration demonstrated remarkable early performance, we observed the inherent limitations of a purely physical feedback system. Due to stochastic noise, the measured velocity (v) rarely reaches an exact value of 1.0 at the final stages of convergence. This prevents the learning rate from smoothing to absolute zero as effectively as traditional policies, allowing baselines to overtake in the terminal phase.

The **Coupling Strategy** (utilizing a 20% inertia weight) was designed to resolve this:

- **Macro Inheritance**: The system inherits the "macro-convergence strategy" of the base scheduler, ensuring the energy level eventually reaches absolute zero.
- **Micro Correction**: The Scheduler wrapper performs a real-time audit on every batch. By accumulating micro-advantages through the rejection of incoherent updates while following the macro-decay trend, the system is able to push the reachability limit ($\mathcal{L}_{min}$) beyond previous theoretical expectations.

3. Universality of Gains and Resilience against Logic Conflicts The experimental data across the eight test groups confirms the robustness of the **Intelligence Inertia** framework, though it also highlights specific interactions with existing heuristics.

- **Universality of Dynamical Acceleration**: In every test group, the introduction of the scheduler wrapper improved the convergence progress at Epoch 30 (indicated by the vertical red dotted line in Figure 6). For discrete schedulers like **Multi-Step Decay** and **Cyclic LR** (Figure 6, third row), the **Dyn. Accel.** exceeded 2.4%. This indicates that the Scheduler wrapper effectively "polishes" the transition periods of coarse-grained algorithms, reducing energy oscillations typically caused by abrupt step changes.
- **Compressing the Thermodynamic Reachability Limit**: For mainstream policies like **OneCycle Policy** and **Cosine Annealing**, the wrapper successfully lowered $\mathcal{L}_{min}$. This represents **"precision etching"** at the base of the solution space—by filtering out microscopic incoherent updates, the model reaches deeper, more stable basins that are typically occluded by stochastic noise [73].
- **Resilience and "Logic Incompatibility"**: We observed slight terminal degradation in **Cosine Restart** (-0.29%) and **Reduce On Plateau**

(−0.86%), which reveals a deep logical incompatibility between specific heuristics and physical inertia:

- **Restart Conflict** (Figure 6, top-right panel): WarmRestart relies on periodically injecting "high-energy shocks" to escape local optima. As seen in the orange curves, while the baseline allows for an unfiltered high-energy pulse, the Scheduler wrapper identifies these shocks as **Directional Dissonance** and dampens them. These two physical objectives—acceleration vs. braking—are diametrically opposed.
- **Plateau Paradox** (Figure 6, second row): **Plateau** schedulers rely on loss volatility to trigger a decay. However, the Scheduler wrapper smooths the R-S Manifold so effectively that the loss curve becomes "too elegant," never reaching the threshold for a plateau. Consequently, the enhanced version maintains a higher base learning rate than the baseline, missing the opportunity for fine-grained terminal convergence.

Crucially, these conflicts further prove the high sensitivity of the Scheduler wrapper in smoothing the information manifold. Despite this logical incompatibility, the hybrid scheme ensures that the system inherits the steepest descent properties of the loss while maintaining competitive parity with the reachability limits of the control groups.

6.3.3 Sub-experiment II: Resilience to Noise Shocks

To evaluate the stability of an intelligent agent in uncertain real-world environments, we subjected a 1M ResNet substrate to high-entropy informational shocks. During the latter half of the training process on CIFAR-10, we injected 100% label noise into every other epoch, alternating between "clean" and "noisy" data streams. This experiment aims to verify whether the **Inertia-Aware Scheduler Wrapper** can maintain the purity of the system's internal rule-set through autonomous dynamical intervention, preventing the "logical shattering" that typically follows an entropy surge.

The comparative performance metrics for the baseline (Exponential Scheduler) and the pure inertia-aware regulation (Wrapper Only) across the shock phases are summarized in Table 4. In this table, the following abbreviations are used: Vel. (Velocity v) and LR (Learning Rate).

The quantitative data indicates a distinct behavioral shift: while the baseline remains "blind" to the quality of incoming data—maintaining its learning intensity despite the lack of coherent environmental feedback—the regulated system demonstrates an immediate dynamical response. This is characterized by the autonomous modulation of the **Brake Ratio**, reflecting the system's real-time perception of structural resistance.

Analysis of Results The results from the noise-injection experiment illustrate a fundamental distinction in how intelligent agents manage high-entropy

Group	Phase	Avg. Loss	Avg. Vel.	Avg. LR	Brake Ratio
Baseline (Exponential)	Pre-Shock	1.4878	0.6505	0.041532	N/A
	Clean Pulse	1.4142	0.6723	0.009068	0.9x (None)
	Noise Pulse	2.1845	0.6926	0.010367	-
Pure Inertia (Wrapper Only)	Pre-Shock	1.4379	0.7525	0.015874	N/A
	Clean Pulse	**1.3541**	0.8111	0.001659	**1.2x** (**Active**)
	Noise Pulse	**2.0841**	0.8634	**0.001420**	-

Table 4: Comparative Dynamical Metrics under Periodic High-Entropy (100% Noise) Shocks. This table contrasts the response of a traditional exponential scheduler against our standalone inertia-aware wrapper during intermittent noise injection, revealing that the regulated system autonomously applies protective braking while the baseline remains blind to the noise-induced entropy surge.

informational surges. By integrating the quantitative metrics from Table 4 with the phenomenological trajectories in Figures 7 and 8, we analyze the micro-mechanisms of the protective braking provided by the **Intelligence Inertia Theory**.

1. Quantifying the Brake Ratio and its Emergent Value A microscopic analysis of the indicators in Table 4 reveals that the system's resilience originates from a subtle but decisive adjustment of the brake ratio—the relative intensity of learning applied to signal versus noise. The data indicates that the regulated group's average learning rate during noise cycles was approximately 20% lower than during clean cycles. In an environment of 100% noise, where infinitesimal residual alignments still exist [74], an absolute 100% lock of parameters is neither optimal nor necessary.

The cumulative effect of this 1.2x braking tilt is transformative. This extra damping serves as a "logical buffer," preventing noise-driven updates from penetrating the deep causal layers of the R-S Manifold. While the baseline fails

because its effective strategy results in Damage > Repair, the regulated group achieves a stable Repair > Damage state. This "self-healing" behavior allows models to maintain cognitive continuity even under extreme volatility.

2. The Physical Dynamics of Protective Braking Figure 7 visualizes the macroscopic consequences of this strategy on the R-S Manifold, contrasting the structural decay of a rigid scheduler with the "self-healing" behavior of the inertia-aware system.

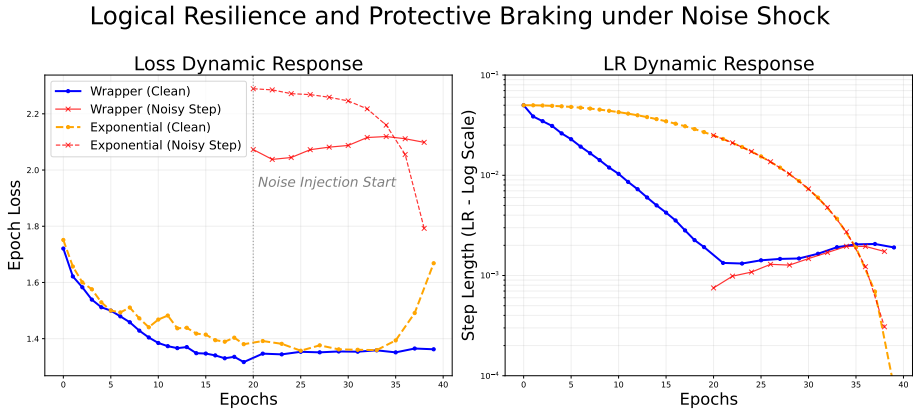

Figure 7: Logical Resilience and Relativistic Braking under Noise Shock. **(Left)** Validation loss across epochs; the baseline (dashed yellow) exhibits convergence of clean and noisy loss curves, indicating disruption of rules and assimilation by noise, while the regulated system (solid blue) maintains a distinct dual-track separation. **(Right)** Relativistic Braking response; the Scheduler wrapper autonomously suppresses its step length during noise steps to preserve rule-purity.

In the baseline group, the loss curves for clean and noisy cycles begin to converge following the onset of noise. Because the baseline maintains nearly identical learning intensity during noise steps, the structural damage inflicted cannot be fully rectified during subsequent clean steps, leading to a cumulative buildup of entropy. Conversely, the inertia-aware group exhibits a persistent "Dual-Track Separation," realizing the principle of **Controlled Damage**. By implementing an autonomous brake during high-velocity pulses, the system ensures that structural contamination remains below the recovery threshold of the subsequent repair cycle.

3. Velocity Manifestation of Environmental Impulses Figure 8 confirms that sudden velocity jumps are an endogenous physical property of intelligent agents when subjected to external shocks in the phase space.

Regardless of whether inertia-aware regulation is applied, noise injection leads to a significant and immediate leap in instantaneous velocity. This phe-

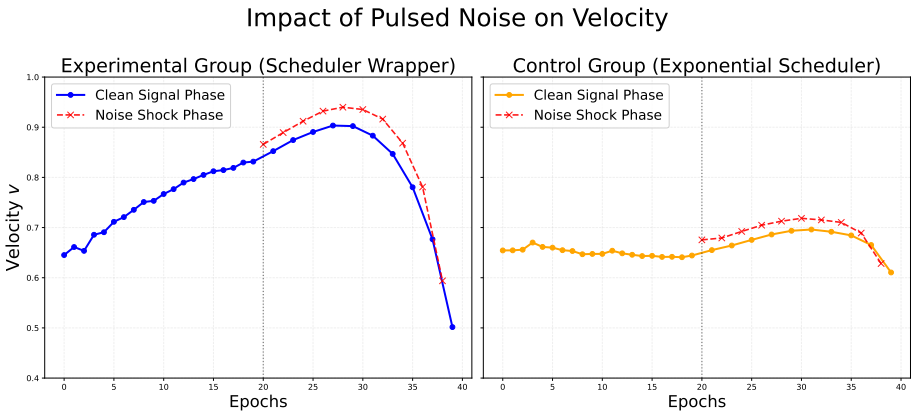

Figure 8: Impact of Pulsed Noise on Velocity. This figure illustrates the velocity response (v) across training epochs. **(Left)** Experimental group (Regulated) and **(Right)** Control group (Baseline). Both groups show that the onset of label noise causes a distinct phase separation, where velocity spikes from a steady-state toward the informational limit, confirming that system velocity is a reliable physical indicator of informational shocks.

nomenon has a direct **physical mapping**: since noise provides no logic-consistent environment feedback ($dS_{ext} \to 0$), it is functionally equivalent to a "high-energy particle flow" bombardment. Under this impact, the model's evolution velocity in R-S space inevitably approaches the informational limit ($v \to 1$), satisfying the conservation laws of the system's underlying dynamics. Therefore, by simply monitoring for instantaneous velocity spikes, the wrapper autonomously triggers the critical protective braking. This mechanism grants the agent **Physical Immunity**—utilizing the inherent laws of dynamics to maintain a stable "Protect-Repair" cycle and preserving cognitive structures in chaotic environments without manual intervention.

6.3.4 Sub-experiment III: The Inertial Barrier in Continual Learning

To evaluate the system's resilience in one of the most demanding scenarios for an intelligent agent, we subjected a 1M-parameter ResNet substrate to a replay-free continual learning task on CIFAR-10 [52]. This setup involves an abrupt transition from the **Old Task** (classes 0–4) to a **New Task** (classes 5–9) at Epoch 20, without utilizing any external data buffers or adjustments to the loss function. This experiment aims to verify whether an agent can utilize its intrinsic inertia (μ) to generate dynamic resistance against "logic collisions"—the sudden, high-energy conflict between the gradients of a new task and the established internal rule-set—thereby preventing the catastrophic destruction of existing knowledge.

The quantitative stability indicators captured during this transition are presented in Table 5.

Table 5: Comparison of Dynamical Stability Indicators in a Replay-Free Continual Learning Scenario. This table contrasts the baseline group against the regulated system during an abrupt task switch. The results indicate that a higher awareness of inertia leads to significantly lower forgetting and better overall task reachability by enforcing an autonomous collision brake at the moment of transition. Certain metrics are abbreviated: "Pre-trans. Loss" refers to Pre-transition Loss (Ep. 19) and "Inst. Breaking Ratio" refers to the Instantaneous Inertial Breaking Ratio.

Metric	**Exponential Wrapper**		**Improvement**
Pre-trans. Loss	0.9253	0.8879	-
Old Task Final Loss	9.1505	7.8801	**13.88% Forgetting Reduction**
Retention Deficit ($\Delta\mathcal{L}$)	8.2626	6.9922	**15.38% Rule Retention**
Full Task Final Loss	4.9093	4.3232	**11.94% Comprehensive Synergy**
Inst. Breaking Ratio	1.08x	**2.92x**	**Autonomous Braking**

The macroscopic and microscopic dynamics of this transition are visualized in Figure 9, highlighting the protective damping effect of the inertia-aware **Scheduler Wrapper**.

Analysis of Results The results of Sub-experiment III provide a decisive demonstration of how the Intelligence Inertia Theory functions as a protective barrier during abrupt task transitions. By analyzing the interplay between knowledge retention and the instantaneous dynamical response on the R-S Manifold, we identify the physical root of why inertia-aware systems survive task shifts that typically lead to the collapse of fixed-policy models.

1. Physical Inhibition of the Forgetting Deficit The **Old Task Loss** curves in Figure 9 **(Left)** reveal a stark contrast in knowledge preservation between the two groups. Following the task switch at Epoch 20, the baseline group's old task loss exhibits a near-vertical surge, peaking at 9.1505 as shown in Table 5. This reflects a catastrophic scenario where, lacking inertial resistance, the model's established output probability distributions are physically shattered to accommodate new, non-coherent gradients. In contrast, the

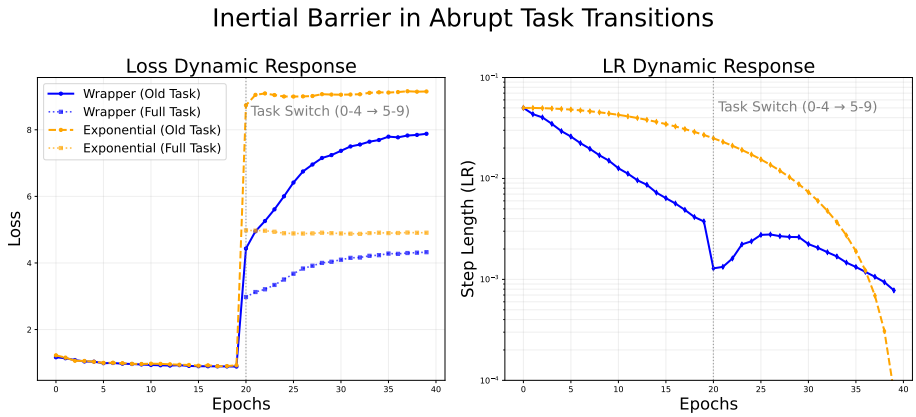

Figure 9: Inertial Barrier during Abrupt Task Transitions. This figure contrasts the behavioral strategies during the task switch at Epoch 20. **(Left)** Validation loss trajectories where the regulated group suppresses the surge of Old Task Loss; **(Right)** Step-length dynamics, where the Scheduler wrapper executes an immediate "Protective Braking" maneuver in response to the high-velocity logic collision.

regulated group demonstrates remarkable **logical resilience**; the surge is significantly suppressed, ultimately locking at 7.8801. The observed reduction in the Retention Deficit ($\Delta\mathcal{L}$) of 15.38% confirms that inertia is not merely a computational burden but a fundamental shield for learned knowledge. The more an agent respects its intrinsic inertia, the more effectively it resists environmental entropy and preserves its causal integrity.

2. Inertia-Induced Step Contraction Figure 9 **(Right)** captures the instantaneous physical behavior of the learning rate at the critical transition point. The exponential baseline group fails to perceive the physical inconsistency between the existing rule-set and the new task requirements, leading to a "blind reconfiguration" that serves as the dynamical origin of catastrophic forgetting. According to the Intelligence Inertia framework, when the gradient flow of a new task deviates orthogonally from the established logic on the R-S Manifold, the rule density (v) approaches the informational limit. This triggers an inertia expansion, which the scheduler wrapper respects via a drastic **Relativistic Contraction** of the step length. As recorded in Table 5, the regulator executed a 2.92x instantaneous braking, effectively providing a "mechanical buffer" that stabilizes the weight structure against non-coherent shocks.

3. Robust Comprehensive Performance The overall performance during the task adaptation phase confirms the long-term benefits of an inertia-centric strategy. By the end of the experiment, the wrapper group's **Full Task Loss** was optimized by 11.94% compared to the baseline. Without resistance, the

baseline system suffers from dynamical "over-speeding," leading to irreversible logical shattering. In contrast, the regulated system ensures that the agent absorbs new knowledge in a quasi-static and low-dissipation manner [19]. This demonstrates that agents with inertial self-awareness can reach superior thermodynamic steady states, efficiently balancing the requirement for plasticity with the necessity of structural stability.

6.3.5 Summary and Guidelines for Inertia-Aware Engineering

Experiment III validates the engineering utility of the **Inertia-Aware Scheduler Wrapper**, transforming the theoretical framework of intelligence inertia into an autonomous **self-audit mechanism** for neural optimization. The results establish three guiding principles for the future development of resilient intelligent agents:

- **Transcending Blind Search**: High optimization efficiency is achieved by identifying and suppressing "unnecessary collisions" between incoming gradients and the established rules on the **R-S Manifold**. By minimizing the dissipation of energy into chaotic rule-vibrations, computational work is focused exclusively on targeted structural evolution.
- **Objective Learning Tempo**: The learning rate is redefined as an **intrinsic physical feature** of the agent's current state rather than a pure heuristic hyperparameter. This tempo is dictated by the system's **velocity/ rule-density** ($v \equiv \rho$) and its alignment with the environmental manifold, providing a principled alternative to human-tuned decay schedules that often ignore the system's underlying **Inertia Expansion**.
- **Extended Applications**: Since the learning rate is but one projection of the system's **Characteristic Cycle** (l) onto the training process, this inertia-aware paradigm can be extended to regulate other critical variables. These include dynamic batch-size scaling based on the noise scale of the R-S Manifold, as well as automated architectural pruning and gradient clipping [75].

7 Experiment Discussion

The experiments presented in this research were designed to progressively transition our framework from theoretical hypothesis to empirical validation and engineering realization. Experiment I successfully adjudicated between classical information geometry and our relativistic formulation, confirming the presence of a "computational wall" and the **Inertia Expansion** effect in high-velocity informational regimes. Building upon this, Experiment II mapped the structural landscape of deep networks on the **R-S Manifold**, revealing that architectural progress is not a matter of arbitrary scaling but a constrained geometric search for an optimal, balanced trajectory. Finally, Experiment III proved the practical utility of the **Inertia-Aware Scheduler Wrapper**, demonstrating that an

inertia-aware controller can stabilize and enhance learning dynamics in volatile environments by respecting the system's intrinsic physical limits.

7.1 Theoretical Value and Potential

The **Intelligence Inertia Theory** fills a fundamental gap in the study of machine learning by establishing a unified physical foundation for artificial intelligence dynamics:

- **Bridging Discrete and Continuous Domains**: By anchoring the **Landauer limit** [7] as the system's static **Rest Inertia (μ_0)**, we provide a unified metric that connects foundational symbolic logic with connectionist manifold dynamics. This proves that the cost of intelligence is not a heuristic variable but is governed by thermodynamic boundaries and the **Symbolic Granularity ($\mathcal{D}$)**.
- **The Physics of Reachability**: Traditional optimization implicitly assumes that increasing computational power leads to proportional performance gains. Our theory exposes a hard cognitive horizon. Experiment I demonstrates that at high **velocity/rule-density ($v \equiv \rho$)**, forced reconfigurations result in a relativistic divergence of effective mass, providing the first-principles explanation for the convergence bottlenecks in network models.
- **A Geodesic for Structural Evolution**: Experiment II demonstrates that architectural intelligence is not a product of blind layering. Efficient evolution requires balanced coordination between **Internal Rule Reconfiguration (dS_R)** and **External State Gain (dS_{ext})**, anchoring the trajectory near the energy equipartition point ($v \approx 0.5$). This establishes a physics-based dynamical criterion for future Neural Architecture Search (NAS) [76].

Furthermore, these results suggest a paradigm shift in the design of autonomous agents. By internalizing inertial feedback, future systems could evolve three profound capabilities:

- **"Logic Pain"**: Intelligent agents could perceive when their velocity spikes toward the informational horizon of conflicting data, triggering spontaneous protective braking to preserve the core constants of their causal structure.
- **Structural Resilience**: By adopting relativistic step-length contraction, agents facing unfamiliar domains could autonomously "freeze" established rules, adapting to novelty through continuous, **quasi-static** integration rather than destructive reorganization [19].
- **Energy-Optimal Evolution**: By anchoring evolution to the golden equipartition axis, agents can maximize cognitive gains with minimal entropic cost, paving the way for deeper evolution even under the constraints of noisy data and limited hardware.

7.2 Limitations and Future Work

While the findings demonstrate a robust unified theory, certain constraints frame the scope of this work and outline future developmental pathways:

- **Scale of Validation**: Due to the combinatorial complexity of scanning the architectural landscape, our empirical validations were constrained to ResNet substrates with a parameter count of 5M. As network depth and parameter count scale toward the frontier, structural coupling becomes highly stochastic. Future work must focus on efficient sampling techniques to partition large-scale models into modular subsystems, enabling inertial regulation across high-parameter architectures [59].
- **Computational Overhead of Auditing**: Currently, the Tier-3 full-spectrum auditing protocol utilized in our experiments—executed on a single NVIDIA RTX 4070 GPU—exhibits non-negligible computational overhead. Future research will be dedicated to algorithm optimization and the development of hardware-matched libraries that compile inertial measurements into low-level operational kernels to minimize latency.
- **Towards Self-Referential Intelligence**: Ultimately, the Intelligence Inertia Theory posits that a true intelligent agent should not rely on external scheduling or human-tuned decay policies. Inertia-aware regulation must be deeply woven into the agent's own architecture rather than implemented as separate functions. The future of this research lies in the development of self-referential cognitive units that inherently sense and respect their own **Inertial Topology**. While realizing an agent that autonomously nurtures its own logic remains a profound challenge, the foundations laid herein provide the necessary map for this journey.

8 Conclusion

This paper has formalized the **physical principles** of *intelligence inertia* as fundamental characteristics governing the energetic cost and dynamical stability of structural evolution in intelligent systems. By decomposing an agent into the dual operators of **Rules** ($\hat{R}$) and **States** ($\hat{S}$) and identifying their fundamental non-commutativity at the scale $\mathcal{D}$, we have moved the study of intelligence beyond phenomenological observation and into the realm of rigorous dynamical mechanics.

Our primary contribution is the derivation and empirical validation of the **Relativistic Cost Equation**. We have demonstrated, through both micro-statistical modeling of adiabatic collisions and large-scale neural experiments on the **R-S Manifold**, that the computational and entropic work required for structural reconfiguration does not scale linearly or quadratically as classical theories suggest. Instead, as an agent's trajectory approaches the limit of its symbolic interpretability, its effective mass undergoes a non-linear **Inertia Expansion**. This creates a physical "computational wall" that defines the absolute

boundaries of reachability for any given logical substrate, effectively quantifying the physical "weight" of intelligence.

Crucially, by bridging the gap between abstract operator algebra and neural tensor dynamics, this work **paves a rigorous path toward the engineering realization** of inertia-aware intelligent systems. We have shown that these physical principles are highly predictive and possess immediate utility for the design of resilient agents. We have identified the **Zig-Zag Geodesic** as the optimal trajectory for architectural evolution, anchoring the system near the point of **Energy Equipartition** ($v \approx 0.5$). Furthermore, we have translated these dynamical laws into **concrete application examples**, most notably the **Inertia-Aware Scheduler Wrapper**. This practical engineering tool enables agents to achieve the steepest descent in loss while simultaneously pushing terminal reachability limits. By autonomously regulating their evolutionary tempo, models equipped with this realization exhibit a form of **"physical immunity"** against high-entropy noise and **"logical resilience"** during abrupt task transitions.

The **systematic formulation** of intelligence inertia fills a profound gap in our understanding of how complex systems learn and adapt. It provides a unified framework that bridges the thermodynamic limit of the single bit with the high-dimensional manifold dynamics of the neural circuit. Most importantly, it suggests that a true intelligent agent must not be a passive recipient of external data, but a self-aware structure that respects its own physical resistance to change. As we move toward the era of **Artificial General Intelligence (AGI)** [77], the principles established herein provide a necessary map for navigating the increasingly dense informational spacetime, paving the way for systems that are not only more powerful but fundamentally more stable, efficient, and aligned with the physical laws of the universe.

Declaration of Generative AI and AI-assisted technologies in the writing process

During the preparation of this work the author(s) used **Gemini** in order to **improve language clarity, polish academic expression, and optimize the manuscript's LaTeX formatting for the journal requirements**. After using this tool/service, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the content of the publication.

References

[1] S. Legg and M. Hutter, "Universal intelligence: A definition of machine intelligence," *Minds and Machines*, vol. 17, no. 4, pp. 391–444, 2007.

[2] S. Russell, P. Norvig, and A. Intelligence, "A modern approach," *Artificial Intelligence. Prentice-Hall, Egnlewood Cliffs*, vol. 25, no. 27, pp. 79–80, 1995.

[3] F. Chollet, "On the measure of intelligence," *arXiv preprint arXiv:1911.01547*, 2019.

[4] J. Hernández-Orallo, *The Measure of All Minds: Evaluating Natural and Artificial Intelligence*. Cambridge University Press, 2017.

[5] M. Mernik, J. Heering, and A. M. Sloane, "When and how to develop domain-specific languages," *ACM computing surveys (CSUR)*, vol. 37, no. 4, pp. 316–344, 2005.

[6] L. Valkov, S. Chaudhuri, B. Lake, A. Gaunt, and C. Milton, "Houdini: Lifelong learning as program synthesis," in *Advances in Neural Information Processing Systems*, vol. 31, 2018.

[7] R. Landauer, "Irreversibility and heat generation in the computing process," *IBM journal of research and development*, vol. 5, no. 3, pp. 183–191, 1961.

[8] A. Bérut, A. Arakelyan, A. Petrosyan, S. Ciliberto, R. Dillenschneider, and E. Lutz, "Experimental verification of landauer's principle linking information and thermodynamics," *Nature*, vol. 483, no. 7388, pp. 187–189, 2012.

[9] C. H. Bennett, "The thermodynamics of computation—a review," *International Journal of Theoretical Physics*, vol. 21, no. 12, pp. 905–940, 1982.

[10] S.-i. Amari, *Information Geometry and Its Applications*. Springer, 2016.

[11] J. Martens, "New insights and perspectives on the natural gradient method," *Journal of Machine Learning Research*, vol. 21, no. 146, pp. 1–76, 2020.

[12] J. Kirkpatrick, R. Pascanu, N. Rabinowitz, J. Veness, G. Desjardins, A. A. Rusu, K. Milan, J. Quan, T. Ramalho, A. Grabska-Barwinska, *et al.*, "Overcoming catastrophic forgetting in neural networks," *Proceedings of the national academy of sciences*, vol. 114, no. 13, pp. 3521–3526, 2017.

[13] A. N. Kolmogorov, "Three approaches to the quantitative definition of information," *Problems of information transmission*, vol. 1, no. 1, pp. 1–7, 1965.

[14] M. Hutter, "Algorithmic information theory: a brief non-technical guide to the field," *arXiv preprint cs/0703024*, 2007.

[15] A. R. Zamir, A. Sax, W. Shen, L. J. Guibas, J. Malik, and S. Savarese, "Taskonomy: Disentangling task transfer learning," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3712–3722, 2018.

[16] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, "How transferable are features in deep neural networks?," in *Advances in neural information processing systems*, 2014.

[17] J. Bechhoefer, "High-precision test of landauer's principle in a feedback trap," in *APS March Meeting Abstracts*, vol. 2015, pp. Z3–002, 2015.

[18] C. H. Bennett, "The thermodynamics of computation—a review," *International Journal of Theoretical Physics*, vol. 21, no. 12, pp. 905–940, 1982.

[19] J. M. Parrondo, J. M. Horowitz, and T. Sagawa, "Thermodynamics of information," *Nature Physics*, vol. 11, no. 2, pp. 131–139, 2015.

[20] R. Pascanu and Y. Bengio, "Revisiting natural gradient for deep networks," *arXiv preprint arXiv:1301.3584*, 2013.

[21] C. H. Bennett, "Logical depth and physical complexity," in *The Universal Turing Machine: A Half-Century Survey* (R. Herken, ed.), pp. 227–257, Oxford University Press, 1988.

[22] M. Li, P. Vitányi, *et al.*, *An introduction to Kolmogorov complexity and its applications*, vol. 3. Springer, 2008.

[23] H. Zenil, "Algorithmic data analytics, small data matters and correlation versus causation," in *Berechenbarkeit der Welt? Philosophie und Wissenschaft im Zeitalter von Big Data*, pp. 453–475, Springer, 2017.

[24] R. M. French, "Catastrophic forgetting in connectionist networks," *Trends in Cognitive Sciences*, vol. 3, no. 4, pp. 128–135, 1999.

[25] F. Zenke, B. Poole, and G. Surya, "Continual learning through synaptic intelligence," *International Conference on Machine Learning (ICML)*, 2017.

[26] R. Hadsell, D. Rao, A. A. Rusu, and R. Pascanu, "Embracing change: Continual learning in deep neural networks," *Trends in Cognitive Sciences*, vol. 24, no. 12, pp. 1028–1040, 2020.

[27] G. M. van de Ven, J. T. Vogelstein, and A. S. Tolias, "Three scenarios for continual learning," *Nature Machine Intelligence*, vol. 4, no. 11, pp. 955–967, 2022.

[28] A. Achille, G. Paolini, G. Mbeng, and S. Soatto, "The information complexity of learning tasks, their structure and their distance," *Information and Inference: A Journal of the IMA*, vol. 10, no. 1, pp. 51–72, 2021.

[29] A. Einstein, "On the electrodynamics of moving bodies," *Annalen der Physik*, vol. 17, pp. 891–921, 1905.

[30] C. E. Shannon, "A mathematical theory of communication," *The Bell System Technical Journal*, vol. 27, no. 3, pp. 379–423, 1948.

[31] W. Heisenberg, "Quantum-theoretical re-interpretation of kinematic and mechanical relations," *Zeitschrift für Physik*, vol. 33, pp. 879–893, 1925.

[32] P. A. Dirac, "The fundamental equations of quantum mechanics," *Proceedings of the Royal Society of London. Series A*, vol. 109, no. 752, pp. 642–653, 1925.

[33] K. Yosida, *Functional analysis*. Springer Science & Business Media, 2012.

[34] H. A. Lorentz, "Electromagnetic phenomena in a system moving with any velocity smaller than that of light," in *Collected Papers: Volume V*, pp. 172–197, Springer, 1937.

[35] E. T. Jaynes, "Information theory and statistical mechanics," *Physical Review*, vol. 106, no. 4, pp. 620–630, 1957.

[36] S. Lloyd, "Ultimate physical limits to computation," *Nature*, vol. 406, no. 6799, pp. 1047–1054, 2000.

[37] J. D. Bekenstein, "Energy cost of information transfer," *Physical Review Letters*, vol. 46, no. 10, pp. 623–626, 1981.

[38] J. A. Wheeler, "Information, physics, quantum: The search for links," *Feynman and computation*, pp. 309–336, 2018.

[39] W. K. Wootters, "Statistical distance and hilbert space," *Physical Review D*, vol. 23, no. 2, pp. 357–362, 1981.

[40] M. Gell-Mann, *The Quark and the Jaguar: Adventures in the Simple and the Complex*. Macmillan, 1995.

[41] S. Awodey, *Category Theory*. Oxford University Press, 2010.

[42] Z. Ji, N. Lee, R. Frieske, T. Yu, D. Su, Y. Xu, E. Ishii, Y. J. Bang, A. Madotto, and P. Fung, "Survey of hallucination in natural language generation," *ACM Computing Surveys*, vol. 55, no. 12, pp. 1–38, 2023.

[43] C. Adami, "What is information?," *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 374, no. 2063, p. 20150230, 2016.

[44] D. Lopez-Paz and M. Ranzato, "Gradient episodic memory for continual learning," in *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[45] K. Huang, *Statistical mechanics*. John Wiley & Sons, 2008.

[46] L. Bottou, F. E. Curtis, and J. Nocedal, "Optimization methods for large-scale machine learning," *SIAM review*, vol. 60, no. 2, pp. 223–311, 2018.

[47] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[48] R. Shwartz-Ziv and N. Tishby, "Opening the black box of deep neural networks via information," *arXiv preprint arXiv:1703.00810*, 2017.

[49] I. Loshchilov and M. Hutter, "Sgdr: Stochastic gradient descent with warm restarts," *arXiv preprint arXiv:1608.03983*, 2016.

[50] Y. N. Dauphin, R. Pascanu, C. Gulcehre, K. Cho, S. Ganguli, and Y. Bengio, "Identifying and attacking the saddle point problem in high-dimensional non-convex optimization," in *Advances in Neural Information Processing Systems*, vol. 27, 2014.

[51] G. Kanwar, "Machine learning and variational algorithms for lattice field theory," *arXiv preprint arXiv:2106.01975*, 2021.

[52] G. I. Parisi, R. Kemker, J. L. Part, C. Kanan, and S. Wermter, "Continual lifelong learning with neural networks: A review," *Neural Networks*, vol. 113, pp. 54–71, 2019.

[53] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[54] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

[55] A. Krizhevsky, G. Hinton, *et al.*, "Learning multiple layers of features from tiny images," 2009.

[56] I. J. Myung, "The importance of complexity in model selection," *Journal of Mathematical Psychology*, vol. 44, no. 1, pp. 190–204, 2000.

[57] S. Weinberg, *Gravitation and Cosmology: Principles and Applications of the General Theory of Relativity*. New York: John Wiley & Sons, 1972.

[58] K. Falconer, *Fractal geometry: mathematical foundations and applications*. John Wiley & Sons, 2013.

[59] J. Kaplan, S. McCandlish, T. Henighan, T. B. Brown, B. Chess, R. Child, S. Gray, A. Radford, J. Wu, and D. Amodei, "Scaling laws for neural language models," *arXiv preprint arXiv:2001.08361*, 2020.

[60] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *International Conference on Machine Learning (ICML)*, pp. 448–456, 2015.

[61] T. M. Mitchell, "The need for biases in learning generalizations," in *Readings in Machine Learning*, pp. 184–191, Morgan Kaufmann, 1980.

[62] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 2002.

[63] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1–9, 2015.

[64] F. Rosenblatt, "The perceptron: a probabilistic model for information storage and organization in the brain.," *Psychological review*, vol. 65, no. 6, p. 386, 1958.

[65] H. Li, Z. Xu, G. Taylor, C. Studer, and T. Goldstein, "Visualizing the loss landscape of neural nets," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 31, 2018.

[66] A. M. Saxe, J. L. McClelland, and S. Ganguli, "Exact solutions to the nonlinear dynamics of learning in deep linear neural networks," *arXiv preprint arXiv:1312.6120*, 2013.

[67] Q. Li, C. Tai, *et al.*, "Stochastic modified equations and adaptive stochastic gradient algorithms," in *International Conference on Machine Learning*, pp. 2101–2110, PMLR, 2017.

[68] L. N. Smith, "A disciplined approach to neural network hyper-parameters: Part 1–learning rate, batch size, momentum, and weight decay," *arXiv preprint arXiv:1803.09820*, 2018.

[69] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, *et al.*, "Pytorch: An imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[70] L. N. Smith and N. Topin, "Super-convergence: Very fast training of neural networks using large learning rates," in *Artificial intelligence and machine learning for multi-domain operations applications*, vol. 11006, pp. 369–386, SPIE, 2019.

[71] L. N. Smith, "Cyclical learning rates for training neural networks," in *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 464–472, IEEE, 2017.

[72] I. Sutskever, J. Martens, G. Dahl, and G. Hinton, "On the importance of initialization and momentum in deep learning," in *International Conference on Machine Learning (ICML)*, pp. 1139–1147, 2013.

[73] N. S. Keskar, D. Mudigere, N. Jorge, S. Mikhail, and T. Ping Tak Peter, "On large-batch training for deep learning: Generalization gap and sharp minima," *arXiv preprint arXiv:1609.04836*, 2016.

[74] B. Han, Q. Yao, T. Liu, G. Niu, I. W. Tsang, J. T. Kwok, and M. Sugiyama, "A survey of label-noise representation learning: Past, present and future," *arXiv preprint arXiv:2011.04406*, 2020.

[75] S. McCandlish, J. Kaplan, A. Vitvitkiy, *et al.*, "An empirical model of large-batch training," *arXiv preprint arXiv:1812.06162*, 2018.

[76] B. Zoph and Q. V. Le, "Neural architecture search with reinforcement learning," in *International Conference on Learning Representations (ICLR)*, 2017.

[77] B. Goertzel, *Artificial general intelligence: concept, state of the art, and future prospects*, vol. 5. Artificial General Intelligence Society, 2014.