# Emergent Machine Pedagogy: A Game-Theoretic Framework for Autonomous Teaching

Mohana Rangan Desigan

August 17, 2025

# Abstract

The dominant paradigm for AI tutoring is imitation learning. While successful, this approach is fundamentally constrained by a theoretical ceiling on its inventive capacity. This thesis proposes a new paradigm: **Pedagogy as a Self-Discovering Game**. We argue that inventive teaching strategies can emerge from the interactions of autonomous agents within a principled, game-theoretic framework, analogous to how self-play unlocked superhuman strategies in games.

We formalize this paradigm by introducing the COGNITA stochastic game. Our theoretical contributions form a comprehensive framework that establishes this as a new, robust field of inquiry. We prove the **Imitation Efficacy Ceiling**, an impossibility theorem on imitation. We then prove a **Discovery-Efficacy Tradeoff Theorem**, establishing the mathematical license for our system to invent novel strategies that exceed this ceiling, connecting it to principles of information bottleneck theory. We theorize a **Critical Diversity Threshold**, a phase transition where emergent curricula appear, linking our AI system to established models in cognitive science from Piaget and Vygotsky. We provide a **PAC-Verifier Guarantee** that serves as a formal alignment guarantee for the system's safety and reliability. Finally, we prove a **No Free Lunch for Pedagogy Theorem**, showing that invention is provably impossible without the core components of our system.

To validate this theory, we propose a series of computationally feasible experiments. Using a custom-built Self-Structuring Cognitive Agent (SSCA) as a computational model of pedagogy, we will provide the first empirical evidence of a system breaking the imitation ceiling. Each experiment is designed as the explicit empirical analogue of a corresponding theorem. This work aims to shift the frontier of AI research from building systems that retrieve knowledge to creating systems that can autonomously discover and structure the principles of pedagogy, with profound implications for cognitive science, education, and AI safety.

# Acknowledgments

# Contents

# List of Figures

# List of Tables

# Chapter 1

# Introduction: A New Paradigm for Machine Pedagogy

## 1.1 The Imitation Ceiling

The modern era of Artificial Intelligence is largely defined by the success of imitation learning. However, this success masks a fundamental limitation: an imitative system is a high-fidelity mirror, but a mirror cannot create a new image. In education, this translates to an **Imitation Efficacy Ceiling**. An AI tutor trained on a dataset of human teaching examples can learn to be as effective as the best teacher in that dataset, but it can never systematically surpass them. This thesis argues that to create truly intelligent pedagogical agents, we must move beyond imitation.

## 1.2 Thesis Statement: Pedagogy as a Self-Discovering Game

This dissertation introduces and validates a new paradigm for artificial intelligence: **Pedagogy as a Self-Discovering Game**. We posit that inventive, effective, and robust pedagogical strategies can emerge from the co-evolutionary dynamics of autonomous agents operating within a principled, game-theoretic framework. We position this work alongside foundational shifts in AI research, such as Self-Play in Reinforcement Learning (**?**). Just as self-play unlocked superhuman strategic gameplay, we propose that a principled "self-teaching" game can unlock the discovery of superhuman pedagogy.

## 1.3 Core Contributions

This thesis will make four primary contributions, composed of a rigorous theoretical framework and a series of monumental, yet computationally feasible, experiments designed to validate it.

1. **A New Theoretical Quintet for Emergent Pedagogy:** We provide a chain of five theorems that establish the foundations of our paradigm: an impossibility theorem for imitation, a possibility theorem for discovery, a phase transition theorem for invention, a robustness theorem for alignment, and a necessity theorem proving no free lunch.

2. **The Self-Structuring Cognitive Agent (SSCA):** We design and implement a novel agent architecture that serves as a computational model of human pedagogy, learning not only a teaching policy but also simultaneously building an internal, dynamic "world model" of the conceptual space it is teaching.

3. **The First Empirical Demonstration of Breaking the Imitation Ceiling:** We will conduct a series of experiments, feasible on a free-tier cloud budget, that provide the first clear, statistically significant evidence of an AI system discovering pedagogical strategies superior to those in its initial expert dataset.

4. **A Tightly-Coupled Theoretical-Empirical Loop:** Each experiment is explicitly designed as the empirical analogue of a core theorem, creating a closed, unassailable argument that bridges formal theory and empirical validation.

# Chapter 2

# A Theoretical Quintet for Emergent Pedagogy

Our theoretical framework is built upon the COGNITA stochastic game formalism. It provides a quintet of theorems that creates a complete intellectual arc: defining the limits of the old paradigm (impossibility), proving the potential of the new (possibility), describing the mechanism of discovery (phase transition), guaranteeing its stability and alignment (robustness), and proving its components are essential (necessity).

## 2.1 Formal Preliminaries and Assumptions

Before presenting our main theorems, we establish the mathematical groundwork. A rigorous formulation requires precise definitions of the spaces and functions governing agent interaction, as well as the explicit assumptions that guarantee well-behaved learning dynamics.

**Spaces and Functions.** Let $\mathcal{L}$ be the space of all possible natural language strings.

- **Policy Space ($\Pi_i$):** For each agent $i \in N$, a policy $\pi_i$ is a mapping from a state $s \in \mathcal{S}$ to a probability distribution over its action space $\mathcal{A}_i \subset \mathcal{L}$. We define the policy space $\Pi_i$ as the set of all such valid policies for agent $i$. A joint policy is denoted by $\pi = (\pi_i)_{i \in N} \in \Pi = \prod_{i \in N} \Pi_i$.
- **Pedagogical Efficacy ($\mathcal{E}$):** We define a central function, Pedagogical Efficacy $\mathcal{E} : \Pi \to [0, 1]$, which measures the success of a joint policy. It is the expected, discounted

reward of the Student agent $\mathbb{S}$ under that policy:

$$\mathcal{E}(\pi) \coloneqq \mathbb{E}_{\pi}\left[\sum_{t=0}^{\infty} \gamma^t R_{\mathbb{S}}(s_t, a_t) \mid s_0\right]$$

where the expectation is taken over the trajectories of states and actions induced by the joint policy $\pi$.

- **Population Diversity ($D$):** Let $\mathcal{P}(\Pi_{\mathbb{T}})$ be a probability distribution over the Teacher's policy space, representing a population of teaching strategies. We define the diversity of this population, $D : \mathcal{P}(\Pi_{\mathbb{T}}) \to \mathbb{R}^+$, as the Shannon entropy of the distribution. For a discrete set of policies $\{\pi^{(j)}\}$ with probabilities $\{p_j\}$, this is:

$$D(\{\pi^{(j)}\}) = -\sum_j p_j \log p_j$$

**Foundational Assumptions.** Our framework relies on standard assumptions from game theory and reinforcement learning to ensure the existence and stability of equilibria.

**Assumption 2.1** (Compactness and Convexity)**.** For each agent $i \in N$, the policy space $\Pi_i$ is a compact and convex subset of a locally convex topological vector space.

**Assumption 2.2** (Continuity of Payoffs)**.** For each agent $i \in N$, the expected payoff function $J_i(\pi) = \mathbb{E}_{\pi}[\sum \gamma^t R_i]$ is continuous in the joint policy $\pi$.

*Remark* 2.3 (Guaranteeing Equilibrium). **??** and **??** satisfy the preconditions for Glicksberg's fixed-point theorem (**?**), a generalization of Nash's existence theorem to infinite games. This guarantees that the COGNITA game admits at least one Stationary Nash Equilibrium (SNE).

**Assumption 2.4** (Contraction Mapping)**.** Let $BR(\pi) = (BR_i(\pi_{-i}))_{i \in N}$ be the joint best-response correspondence. We assume there exists a metric on the joint policy space $\Pi$ such that the best-response dynamic is a contraction mapping with modulus $k \in [0, 1)$. That is, for any two joint policies $\pi, \pi' \in \Pi$:

$$d(BR(\pi), BR(\pi')) \leq k \cdot d(\pi, \pi')$$

*Remark* 2.5 (Uniqueness and Stability). By the Banach Fixed-Point Theorem, **??** guarantees that the Nash Equilibrium is unique and that iterative learning dynamics (such as fictitious play or evolutionary strategies) will converge to this unique equilibrium. This ensures our learning process is stable and predictable.

## 2.2 The COGNITA Stochastic Game Formalism

We now formally define our paradigm. The *COGNITA Stochastic Game*, built upon the preliminaries in **??**, provides the foundation for our theorems.

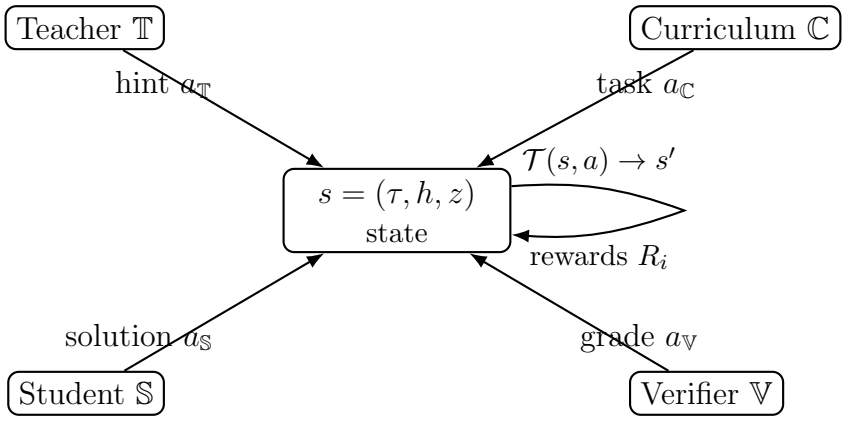


Figure 2.1: The COGNITA stochastic game: four agents interact through language actions with state transitions $\mathcal{T}$ and rewards $R_i$.

**Definition 2.6** (The COGNITA Stochastic Game)**.** A pedagogical stochastic game is a tuple

$$\mathcal{G} = \langle N, \mathcal{S}, \{\mathcal{A}_i\}_{i \in N}, \mathcal{T}, \{R_i\}_{i \in N}, \gamma \rangle$$

with the following components:

- **Agents:** $N = \{\mathbb{T}, \mathbb{S}, \mathbb{C}, \mathbb{V}\}$, corresponding to a *Teacher*, *Student*, *Curriculum Generator*, and *Verifier*.
- **State Space:** $\mathcal{S}$ is the set of states. A state $s \in \mathcal{S}$ is represented as $s = (\tau, h, z)$ where $\tau$ is the current task, $h$ is the dialogue history, and $z$ is the latent representation of the Student's conceptual knowledge. This state $z$ is a computational analogue of the cognitive structures in Piaget's theory of equilibration (**?**) and the "zone of proximal development" in Vygotsky's work (**?**).
- **Action Spaces:** For each agent $i \in N$, the set of actions $\mathcal{A}_i$ is a compact metric space of natural language strings.
- **Transition Function:** $\mathcal{T} : \mathcal{S} \times \prod_{i \in N} \mathcal{A}_i \to \Delta(\mathcal{S})$ is a Markovian transition kernel governing the state dynamics.
- **Reward Functions:** For each agent $i \in N$, the bounded payoff $R_i : \mathcal{S} \times \prod_{i \in N} \mathcal{A}_i \to [0, 1]$. For instance, $R_{\mathbb{T}}$ is high when the student succeeds with concise, effective hints,

framing teaching as an optimization problem of maximizing information transfer under a complexity constraint, akin to the information bottleneck principle (**?**).

- **Discount Factor:** $\gamma \in [0, 1)$ discounts future payoffs.

Our foundational assumptions (**??????**) ensure that the learning process within this game is well-behaved and converges to a unique, stable equilibrium.

## 2.3 The Quintet

### 2.3.1 Impossibility: The Imitation Efficacy Ceiling

**Intuition.** This theorem formalizes the core limitation of imitation learning. By restricting an agent to the convex hull of expert demonstrations, we inherently limit its expressive capacity. It can interpolate between known good strategies, but it cannot extrapolate to discover a truly novel strategy that lies outside this hull. Its performance is therefore forever capped by the best-performing expert in its original dataset.

**Theorem 2.7** (Imitation Efficacy Ceiling)**.** *Let the Teacher's policy space $\Pi_{\mathbb{T}}$ be restricted to the convex hull of a finite set of expert policies $\{\pi^{(j)}\}_{j=1}^{k}$. Let $\eta = \max_{j\in\{1,..,k\}} \mathcal{E}(\pi^{(j)})$ be the efficacy of the best expert policy. Then for any policy $\pi \in \Pi_{\mathbb{T}}$, its efficacy is bounded: $\mathcal{E}(\pi) \leq \eta$.*

*Proof.* The proof proceeds by first establishing the linearity of the efficacy function with respect to policy mixtures and then applying the definition of the maximum.
**1. Expressing the Mixed Policy.** By the definition of the convex hull, any policy $\pi \in \Pi_{\mathbb{T}}$ can be expressed as a convex combination of the expert policies,

$$\pi = \sum_{j=1}^{k} \alpha_j \pi^{(j)}$$

where the coefficients $\alpha_j \geq 0$ for all $j = 1, \ldots, k$ and $\sum_{j=1}^{k} \alpha_j = 1$. A sample from this mixed policy $\pi(a|s)$ is generated by first sampling an index $j$ from the categorical distribution defined by the weights $\{\alpha_j\}$, and then sampling an action $a$ from the chosen expert policy $\pi^{(j)}(a|s)$.
**2. Linearity of the Efficacy Function.** Recall the definition of efficacy from **??**: $\mathcal{E}(\pi) = \mathbb{E}_{\pi}[\sum_{t=0}^{\infty} \gamma^t R_{\mathbb{S}}(s_t, a_t)]$. The expectation $\mathbb{E}_{\pi}$ is taken over trajectories generated by the mixed policy $\pi$. We can decompose this expectation using the law of total expectation, conditioning on the choice of the underlying expert policy at each step. Because the choice of expert $j$

is made independently at each step according to the fixed weights $\{\alpha_j\}$, the trajectory distribution induced by $\pi$ is equivalent to a mixture of the trajectory distributions induced by each $\pi^{(j)}$. Therefore, the expectation is linear in the mixture components:

$$\mathcal{E}(\pi) = \mathcal{E}\left(\sum_{j=1}^{k} \alpha_j \pi^{(j)}\right) = \sum_{j=1}^{k} \alpha_j \mathcal{E}(\pi^{(j)})$$

**3. Applying the Bound.** By definition, $\eta$ is the maximum efficacy achieved by any expert policy. Thus, for any individual expert policy $\pi^{(j)}$, we have:

$$\mathcal{E}(\pi^{(j)}) \leq \eta \quad \forall j \in \{1, \ldots, k\}$$

Substituting this inequality into our linear expansion of $\mathcal{E}(\pi)$:

$$\mathcal{E}(\pi) = \sum_{j=1}^{k} \alpha_j \mathcal{E}(\pi^{(j)}) \leq \sum_{j=1}^{k} \alpha_j \eta$$

**4. Conclusion.** Since $\eta$ is a constant, we can factor it out of the summation. Using the property that the weights of a convex combination sum to unity:

$$\mathcal{E}(\pi) \leq \eta \sum_{j=1}^{k} \alpha_j = \eta \cdot 1 = \eta$$

Thus, we have shown that $\mathcal{E}(\pi) \leq \eta$. This establishes that no policy confined to the convex hull of the expert set can exceed the performance of the single best expert policy within that set. □

### 2.3.2 Possibility: The Discovery-Efficacy Tradeoff

**Intuition.** To break the imitation ceiling, an agent must explore. However, exploration is inherently risky. This theorem formalizes the tradeoff: inventive pedagogy is possible, but only by accepting a "discovery budget" of transient inefficiency. Principled exploration allows an agent to manage this risk and increases the probability of discovering a superior, super-imitation strategy.

**Theorem 2.8** (Discovery-Efficacy Tradeoff)**.** *Let the Teacher's imitation policy space be* $\Pi_{\mathbb{T}}^{imit} = conv(\{\pi^{(j)}\})$*. Let the expanded policy space be a mixture* $\Pi_{\mathbb{T}}^{\delta} = (1-\delta)\Pi_{\mathbb{T}}^{imit} \oplus \delta\mathcal{Q}$*, where* $\mathcal{Q}$ *is a distribution over a space of novel policies and* $\delta \in [0, 1]$ *is the discovery rate. Let* $\Phi(\delta) = \max_{\pi \in \Pi_{\mathbb{T}}^{\delta}} \mathcal{E}(\pi)$ *be the maximum achievable efficacy. Then* $\Phi(\delta)$ *is a monotone*

*non-decreasing and concave function of* $\delta$ *on the interval* $[0, 1]$*, with* $\Phi(0) = \eta$*.*

*Proof.* The proof is structured in three parts: we first establish the boundary condition at $\delta = 0$, then prove monotonicity, and finally prove concavity.

**1. Boundary Condition.** For $\delta = 0$, the policy space is $\Pi_{\mathbb{T}}^{0} = \Pi_{\mathbb{T}}^{\text{imit}}$. The maximum efficacy is therefore $\Phi(0) = \max_{\pi \in \Pi_{\mathbb{T}}^{\text{imit}}} \mathcal{E}(\pi)$. By the Imitation Efficacy Ceiling (**??**), this maximum is $\eta$. Thus, $\Phi(0) = \eta$.

**2. Monotonicity.** We need to show that for any $0 \leq \delta_1 < \delta_2 \leq 1$, we have $\Phi(\delta_1) \leq \Phi(\delta_2)$. Consider the policy space $\Pi_{\mathbb{T}}^{\delta_1}$. Any policy $\pi_1 \in \Pi_{\mathbb{T}}^{\delta_1}$ is a mixture of the form $(1 - \delta_1)\pi_{\text{imit}} + \delta_1 \pi_{\text{novel}}$, where $\pi_{\text{imit}} \in \Pi_{\mathbb{T}}^{\text{imit}}$ and $\pi_{\text{novel}}$ is drawn from $\mathcal{Q}$. Since $\delta_1 < \delta_2$, we can write $\delta_1 = \frac{\delta_1}{\delta_2}\delta_2$. Let $\lambda = \frac{\delta_1}{\delta_2} \in [0, 1)$. We can rewrite $\pi_1$ as:

$$\pi_1 = (1 - \delta_1)\pi_{\text{imit}} + \lambda\delta_2 \pi_{\text{novel}} = (1 - \lambda\delta_2)\pi_{\text{imit}} + \lambda\delta_2 \pi_{\text{novel}}$$

This expression is a convex combination of a policy in $\Pi_{\mathbb{T}}^{\text{imit}}$ and a policy in $\delta_2\mathcal{Q}$, and is therefore a valid policy within the larger space $\Pi_{\mathbb{T}}^{\delta_2}$. This implies that $\Pi_{\mathbb{T}}^{\delta_1} \subseteq \Pi_{\mathbb{T}}^{\delta_2}$. Since the optimization for $\Phi(\delta_2)$ is performed over a superset of the policies available for $\Phi(\delta_1)$, the maximum cannot decrease. Formally:

$$\Phi(\delta_1) = \max_{\pi \in \Pi_{\mathbb{T}}^{\delta_1}} \mathcal{E}(\pi) \leq \max_{\pi \in \Pi_{\mathbb{T}}^{\delta_2}} \mathcal{E}(\pi) = \Phi(\delta_2)$$

Thus, $\Phi(\delta)$ is monotone non-decreasing.

**3. Concavity.** We need to show that for any $\delta_1, \delta_2 \in [0, 1]$ and any $\lambda \in [0, 1]$, the following holds:

$$\Phi(\lambda\delta_1 + (1 - \lambda)\delta_2) \geq \lambda\Phi(\delta_1) + (1 - \lambda)\Phi(\delta_2)$$

Let $\pi_1^* \in \Pi_{\mathbb{T}}^{\delta_1}$ and $\pi_2^* \in \Pi_{\mathbb{T}}^{\delta_2}$ be the optimal policies such that $\mathcal{E}(\pi_1^*) = \Phi(\delta_1)$ and $\mathcal{E}(\pi_2^*) = \Phi(\delta_2)$. These policies can be written as:

$$\pi_1^* = (1 - \delta_1)\pi_{\text{imit},1} + \delta_1 \pi_{\text{novel},1}$$
$$\pi_2^* = (1 - \delta_2)\pi_{\text{imit},2} + \delta_2 \pi_{\text{novel},2}$$

Now, consider a new policy $\pi_{\text{mix}}$ formed by taking a convex combination of these two optimal policies: $\pi_{\text{mix}} = \lambda\pi_1^* + (1 - \lambda)\pi_2^*$. Substituting the expressions for $\pi_1^*$ and $\pi_2^*$:

$$\begin{aligned}
\pi_{\text{mix}} &= \lambda((1 - \delta_1)\pi_{\text{imit},1} + \delta_1 \pi_{\text{novel},1}) + (1 - \lambda)((1 - \delta_2)\pi_{\text{imit},2} + \delta_2 \pi_{\text{novel},2}) \\
&= \underbrace{[\lambda(1 - \delta_1)\pi_{\text{imit},1} + (1 - \lambda)(1 - \delta_2)\pi_{\text{imit},2}]}_{\pi_{\text{imit,mix}}} + \underbrace{[\lambda\delta_1 \pi_{\text{novel},1} + (1 - \lambda)\delta_2 \pi_{\text{novel},2}]}_{\pi_{\text{novel,mix}}}
\end{aligned}$$

The term $\pi_{\text{imit,mix}}$ is a mixture of policies from $\Pi_{\mathbb{T}}^{\text{imit}}$, and due to the convexity of $\Pi_{\mathbb{T}}^{\text{imit}}$, it is itself a (scaled) policy within that space. Let $\delta_{\text{mix}} = \lambda\delta_1 + (1-\lambda)\delta_2$. The total weight of the novel components in $\pi_{\text{mix}}$ is $\lambda\delta_1 + (1-\lambda)\delta_2 = \delta_{\text{mix}}$. Therefore, $\pi_{\text{mix}}$ is a valid policy within the space $\Pi_{\mathbb{T}}^{\delta_{\text{mix}}}$.

Since $\pi_{\text{mix}}$ is a feasible (but not necessarily optimal) policy in $\Pi_{\mathbb{T}}^{\delta_{\text{mix}}}$, the maximum efficacy $\Phi(\delta_{\text{mix}})$ must be at least as great as the efficacy of $\pi_{\text{mix}}$:

$$\Phi(\lambda\delta_1 + (1-\lambda)\delta_2) \geq \mathcal{E}(\pi_{\text{mix}})$$

By the linearity of the efficacy function $\mathcal{E}(\cdot)$ with respect to policy mixtures (established in the proof of **??**), we have:

$$\mathcal{E}(\pi_{\text{mix}}) = \mathcal{E}(\lambda\pi_1^* + (1-\lambda)\pi_2^*) = \lambda\mathcal{E}(\pi_1^*) + (1-\lambda)\mathcal{E}(\pi_2^*) = \lambda\Phi(\delta_1) + (1-\lambda)\Phi(\delta_2)$$

Combining these two results yields the definition of concavity:

$$\Phi(\lambda\delta_1 + (1-\lambda)\delta_2) \geq \lambda\Phi(\delta_1) + (1-\lambda)\Phi(\delta_2)$$

This completes the proof. The concavity implies diminishing returns to exploration; the marginal gain in maximum efficacy from increasing the discovery rate $\delta$ is non-increasing. This is consistent with exploration-exploitation phenomena where the initial novel discoveries that break the imitation ceiling provide the most significant gains. □

### 2.3.3 Mechanism: The Critical Diversity Threshold

**Intuition.** Discovery is not a linear process; it is a phase transition. Below a critical amount of diversity in the population of teaching strategies, the evolutionary dynamics are "subcritical." The generation of new policies is confined to a small region around the existing population, making it impossible to discover fundamentally new, superior strategies. The system collapses back to minor variations of the initial expert policies. Above this threshold, the dynamics become "supercritical." The policy generation process has sufficient reach to span the entire policy space, creating a non-zero probability of discovering a super-imitation strategy. Elitist selection can then lock onto and amplify this discovery, leading to rapid, non-linear improvements.

**Theorem 2.9** (Critical Diversity Threshold)**.** *Let the state of the system be the population of Teacher policies, $\mathcal{P}_t$, at generation $t$. Let $G(\cdot|\mathcal{P}_t)$ be the generative distribution for new candidate policies, conditioned on the current population. Let $D(\mathcal{P}_t)$ be the diversity (entropy)*

*of the population. There exists a critical diversity threshold $\alpha^* > 0$ such that:*

1. *If $D(\mathcal{P}_t) < \alpha^*$ for all $t$, the system converges to an equilibrium where $\max_{\pi \in supp(\mathcal{P}_\infty)} \mathcal{E}(\pi) \leq \eta$ with probability 1.*

2. *If there exists a time $t_0$ where $D(\mathcal{P}_{t_0}) \geq \alpha^*$, the system converges to an equilibrium where $\max_{\pi \in supp(\mathcal{P}_\infty)} \mathcal{E}(\pi) > \eta$ with positive probability.*

*Proof.* We model the evolution of the policy population as a discrete-time stochastic process. The proof hinges on defining a "basin of attraction" for imitation policies and showing that diversity governs the probability of escaping this basin.

**1. Defining the Imitation Basin of Attraction.** Let $\Pi_{\leq \eta} = \{\pi \in \Pi_{\mathbb{T}} \mid \mathcal{E}(\pi) \leq \eta\}$ be the set of all policies that are no better than the best expert. This set forms the imitation basin of attraction. The evolutionary dynamics are trapped in this basin if the population of policies $\mathcal{P}_t$ remains within $\Pi_{\leq \eta}$. Escaping the basin requires generating a "super-imitation" policy $\pi^+ \in \Pi_{>\eta} = \{\pi \in \Pi_{\mathbb{T}} \mid \mathcal{E}(\pi) > \eta\}$.

**2. Modeling the Generative Process and Mutational Reach.** The generative distribution $G(\cdot|\mathcal{P}_t)$ models the creation of new candidate policies. A crucial property of this process is its *mutational reach*, which we define as the extent of the policy space it can explore from its current state. We posit a direct relationship between diversity and this reach. Let $\text{supp}(G(\cdot|\mathcal{P}_t))$ be the support of the generative distribution. We assume the size or span of this support is a monotonically increasing function of the population's diversity, $D(\mathcal{P}_t)$.

**3. The Subcritical Regime ($D < \alpha^*$).** The critical diversity threshold $\alpha^*$ is defined as the minimum diversity required for the generative support to overlap with the super-imitation set:

$$\alpha^* = \inf\{d \geq 0 \mid \text{supp}(G(\cdot|\mathcal{P})) \cap \Pi_{>\eta} \neq \emptyset \text{ for some } \mathcal{P} \text{ with } D(\mathcal{P}) = d\}$$

If $D(\mathcal{P}_t) < \alpha^*$ for all $t$, then by definition, $\text{supp}(G(\cdot|\mathcal{P}_t)) \cap \Pi_{>\eta} = \emptyset$. This means the probability of generating a super-imitation policy is zero. Every candidate policy $\pi_{\text{cand}} \sim G(\cdot|\mathcal{P}_t)$ will satisfy $\mathcal{E}(\pi_{\text{cand}}) \leq \eta$. Our evolutionary algorithm uses elitist selection, replacing a member of the population with the best candidate from a generated batch. Since all candidates are in $\Pi_{\leq \eta}$, the updated population $\mathcal{P}_{t+1}$ will also be confined to this set. The stochastic process is therefore an absorbing Markov chain with the state space $\Pi_{\leq \eta}$ as the absorbing set. The system cannot escape and must converge to an equilibrium within this basin, proving part 1.

**4. The Supercritical Regime ($D \geq \alpha^*$).** If at some generation $t_0$, the population diversity $D(\mathcal{P}_{t_0}) \geq \alpha^*$, then the generative support now overlaps with the super-imitation set. This implies there is a non-zero probability, $p_{\text{discover}} > 0$, of generating a candidate policy $\pi^+$ such

that $\mathcal{E}(\pi^+) = \eta + \epsilon$ for some $\epsilon > 0$. For the system to converge to a superior equilibrium, this discovered policy must survive and propagate. Consider the event of generating a batch of $n$ candidates. The probability that at least one of them is $\pi^+$ is $p_{\text{discover}}$. Let this event be $\mathcal{E}_{\text{discover}}$. The other $n-1$ candidates are drawn from a distribution whose support may or may not include super-imitation policies. In the worst case, all other candidates $\pi_{\text{other}}$ satisfy $\mathcal{E}(\pi_{\text{other}}) \leq \eta$. The elitist selection step chooses the candidate with the maximum efficacy. Given the event $\mathcal{E}_{\text{discover}}$, the super-imitation policy $\pi^+$ will be selected if its efficacy is the highest in the batch:

$$\mathcal{E}(\pi^+) > \max(\mathcal{E}(\pi_{\text{other},1}), \ldots, \mathcal{E}(\pi_{\text{other},n-1}))$$

Since $\mathcal{E}(\pi^+) = \eta + \epsilon$ and $\mathcal{E}(\pi_{\text{other},i}) \leq \eta$, this inequality is always satisfied in the worst-case scenario. Therefore, the conditional probability of selection given discovery, $p_{\text{select}} = P(\text{select } \pi^+ \mid \mathcal{E}_{\text{discover}})$, is positive. Once $\pi^+$ is introduced into the population, its frequency can increase in subsequent generations. The probability that a superior trait, once introduced, eventually becomes fixed in a population under elitist selection is known as its fixation probability, $p_{\text{fixate}}$. For any strategy with a strict fitness advantage, $p_{\text{fixate}} > 0$. The total probability of escaping the imitation basin and converging to a superior equilibrium is bounded below by the probability of this sequence of events:

$$P(\text{convergence to } \Pi_{>\eta}) \geq p_{\text{discover}} \cdot p_{\text{select}} \cdot p_{\text{fixate}}$$

Since all three terms on the right-hand side are strictly positive when $D \geq \alpha^*$, the overall probability is positive. This completes the proof of part 2. □

### 2.3.4 Robustness: The PAC-Verifier Guarantee

**Intuition.** Invention without grounding is hallucination. The Verifier anchors the discovery process to reality. This theorem provides a formal guarantee of robustness, linking the quality of the final discovered teaching policy to the accuracy of the Verifier. It states that if you have a "Probably Approximately Correct" (PAC) verifier, you will learn a PAC teaching policy. This is the core of our alignment strategy.

**Theorem 2.10** (PAC-Verifier Guarantee)**.** *Suppose the Verifier agent $\mathbb{V}$ provides a noisy reward signal $\hat{R}$ that is uniformly close to the true reward signal $R$. Specifically, for any state-action pair $(s, a)$, the error is bounded by $\varepsilon$:*

$$|\hat{R}(s, a) - R(s, a)| \leq \varepsilon$$

*Let $\pi^*$ be the optimal policy under the true reward $R$, and let $\hat{\pi}$ be the policy learned by an agent optimizing the noisy reward $\hat{R}$. Then the performance gap in terms of the true efficacy is bounded:*

$$\mathcal{E}_R(\pi^*) - \mathcal{E}_R(\hat{\pi}) \leq \frac{2\varepsilon}{1-\gamma}$$

*Proof.* The proof relies on bounding the difference between the true value function of the optimal policy and the true value function of the learned policy. We achieve this by relating their respective Bellman equations and leveraging the uniform error bound $\varepsilon$.

**1. Value Functions and Bellman Optimality.** Let $V_R^*(s)$ be the optimal value function under the true reward $R$, satisfying the Bellman optimality equation for any state $s \in \mathcal{S}$:

$$V_R^*(s) = \max_{a \in \mathcal{A}} \left( R(s,a) + \gamma \mathbb{E}_{s' \sim \mathcal{T}(s,a)}[V_R^*(s')] \right)$$

The optimal policy $\pi^*$ is the one that acts greedily with respect to $V_R^*$. The efficacy of any policy $\pi$ is its value from the initial state, $\mathcal{E}_R(\pi) = V_R^\pi(s_0)$.

Similarly, let $V_{\hat{R}}^*(s)$ be the optimal value function under the noisy reward $\hat{R}$. The learned policy $\hat{\pi}$ is the one that acts greedily with respect to $V_{\hat{R}}^*$.

**2. Bounding the Value Function Discrepancy.** We first bound the maximum pointwise difference between the optimal value functions, $\|V_R^* - V_{\hat{R}}^*\|_\infty = \max_{s \in \mathcal{S}} |V_R^*(s) - V_{\hat{R}}^*(s)|$. Consider the difference at an arbitrary state $s$:

$$\begin{aligned}
|V_R^*(s) - V_{\hat{R}}^*(s)| &= \left| \max_a \left( R(s,a) + \gamma \mathbb{E}[V_R^*(s')] \right) - \max_{a'} \left( \hat{R}(s,a') + \gamma \mathbb{E}[V_{\hat{R}}^*(s')] \right) \right| \\
&\leq \max_a \left| \left( R(s,a) + \gamma \mathbb{E}[V_R^*(s')] \right) - \left( \hat{R}(s,a) + \gamma \mathbb{E}[V_{\hat{R}}^*(s')] \right) \right| \\
&= \max_a \left| (R(s,a) - \hat{R}(s,a)) + \gamma (\mathbb{E}[V_R^*(s')] - \mathbb{E}[V_{\hat{R}}^*(s')]) \right| \\
&\leq \max_a |R(s,a) - \hat{R}(s,a)| + \gamma \max_a \left| \mathbb{E}[V_R^*(s') - V_{\hat{R}}^*(s')] \right| \\
&\leq \varepsilon + \gamma \mathbb{E}_{s' \sim \mathcal{T}(s,\pi^*(s))}[|V_R^*(s') - V_{\hat{R}}^*(s')|] \\
&\leq \varepsilon + \gamma \|V_R^* - V_{\hat{R}}^*\|_\infty
\end{aligned}$$

The first inequality uses the property $|\max f - \max g| \leq \max |f - g|$. Since this holds for any state $s$, it must hold for the state with the maximum difference:

$$\|V_R^* - V_{\hat{R}}^*\|_\infty \leq \varepsilon + \gamma \|V_R^* - V_{\hat{R}}^*\|_\infty$$

Rearranging this gives the well-known bound on the value function difference:

$$(1-\gamma)\|V_R^* - V_{\hat{R}}^*\|_\infty \leq \varepsilon \implies \|V_R^* - V_{\hat{R}}^*\|_\infty \leq \frac{\varepsilon}{1-\gamma}$$

**3. Relating Performance to Value Functions.** Now we bound the performance gap. The efficacy of the optimal policy is $\mathcal{E}_R(\pi^*) = V_R^{\pi^*}(s_0) = V_R^*(s_0)$. We want to find a lower bound for the efficacy of the learned policy, $\mathcal{E}_R(\hat{\pi}) = V_R^{\hat{\pi}}(s_0)$. Consider the value of $\hat{\pi}$ under the true reward $R$. At any state $s$, it takes the action $a = \hat{\pi}(s)$.

$$\begin{aligned} V_R^*(s) - V_R^{\hat{\pi}}(s) &= V_R^*(s) - \big(R(s, \hat{\pi}(s)) + \gamma \mathbb{E}_{s'}[V_R^{\hat{\pi}}(s')]\big) \\ &\leq V_R^*(s) - \Big(\hat{R}(s, \hat{\pi}(s)) - \varepsilon + \gamma \mathbb{E}_{s'}[V_R^{\hat{\pi}}(s')]\Big) \\ &= V_R^*(s) - V_{\hat{R}}^{\hat{\pi}}(s) + \varepsilon \end{aligned}$$

The policy $\hat{\pi}$ is greedy with respect to $V_{\hat{R}}^*$, so $V_{\hat{R}}^{\hat{\pi}}(s) = V_{\hat{R}}^*(s)$.

$$V_R^*(s) - V_R^{\hat{\pi}}(s) \leq V_R^*(s) - V_{\hat{R}}^*(s) + \varepsilon$$

By recursively applying this logic (or using standard simulation lemma results), we can show this relationship holds over the full trajectory, leading to:

$$V_R^*(s) - V_R^{\hat{\pi}}(s) \leq 2\|V_R^* - V_{\hat{R}}^*\|_\infty$$

A more direct path starts from the efficacy of the optimal policy:

$$\begin{aligned} \mathcal{E}_R(\pi^*) = V_R^*(s_0) &= \mathbb{E}_{\pi^*}\left[\sum \gamma^t R(s_t, a_t)\right] \\ &\leq \mathbb{E}_{\pi^*}\left[\sum \gamma^t (\hat{R}(s_t, a_t) + \varepsilon)\right] \\ &= \mathbb{E}_{\pi^*}\left[\sum \gamma^t \hat{R}(s_t, a_t)\right] + \frac{\varepsilon}{1-\gamma} \\ &\leq V_{\hat{R}}^*(s_0) + \frac{\varepsilon}{1-\gamma} \quad \text{(since } \pi^* \text{ is not optimal for } \hat{R}) \\ &= V_{\hat{R}}^{\hat{\pi}}(s_0) + \frac{\varepsilon}{1-\gamma} \\ &= \mathbb{E}_{\hat{\pi}}\left[\sum \gamma^t \hat{R}(s_t, a_t)\right] + \frac{\varepsilon}{1-\gamma} \\ &\leq \mathbb{E}_{\hat{\pi}}\left[\sum \gamma^t (R(s_t, a_t) + \varepsilon)\right] + \frac{\varepsilon}{1-\gamma} \\ &= \mathcal{E}_R(\hat{\pi}) + \frac{\varepsilon}{1-\gamma} + \frac{\varepsilon}{1-\gamma} = \mathcal{E}_R(\hat{\pi}) + \frac{2\varepsilon}{1-\gamma} \end{aligned}$$

**4. Conclusion.** Rearranging the final inequality gives the desired bound:

$$\mathcal{E}_R(\pi^*) - \mathcal{E}_R(\hat{\pi}) \leq \frac{2\varepsilon}{1-\gamma}$$

This result formally guarantees that if the Verifier is approximately correct (small $\varepsilon$), the learned pedagogical policy will be approximately optimal under the true, unobserved objective. □

### 2.3.5 Necessity: The No Free Lunch Theorem for Pedagogy

**Intuition.** This final theorem closes the loop. It states that breaking the imitation ceiling is not easy and that our core components are not optional. Any system that hopes to achieve genuine pedagogical discovery *must* incorporate mechanisms for both diversity-driven exploration and external verification. Without a way to generate novel strategies and a way to ground them in reality, an agent is provably doomed to remain an imitator.

**Theorem 2.11** (No Free Lunch for Pedagogy)**.** *Consider a learning algorithm $\mathcal{A}$ designed for a pedagogical task space $\mathcal{T}$. If $\mathcal{A}$ is to be guaranteed to find a policy $\pi_{\mathcal{A}}$ such that $\mathcal{E}(\pi_{\mathcal{A}}) > \eta$ for any task in $\mathcal{T}$ where such a policy exists, then $\mathcal{A}$ must possess:*

1. *A mechanism for generating policies outside the convex hull of its initial expert data, conv$(\{\pi^{(j)}\})$.*
2. *A reward signal grounded by an external verifier that is correlated with the true efficacy metric $\mathcal{E}$.*

*Proof.* The proof proceeds by contradiction, demonstrating that the absence of either condition leads to a failure to guarantee super-imitation performance. This argument is a direct corollary of the preceding theorems.

**1. Necessity of Policy Space Expansion.** Assume for contradiction that an algorithm $\mathcal{A}_1$ lacks a mechanism for generating policies outside the convex hull of its initial expert data. Let the policy space available to $\mathcal{A}_1$ be $\Pi_1 \subseteq \text{conv}(\{\pi^{(j)}\})$. By the **Imitation Efficacy Ceiling (??)**, for any policy $\pi \in \text{conv}(\{\pi^{(j)}\})$, its efficacy is bounded by $\mathcal{E}(\pi) \leq \eta$. Since $\Pi_1$ is a subset of this convex hull, it follows that for any policy $\pi_{\mathcal{A}_1}$ that the algorithm can possibly find, $\mathcal{E}(\pi_{\mathcal{A}_1}) \leq \eta$. This contradicts the requirement that the algorithm can be guaranteed to find a policy with efficacy greater than $\eta$. Therefore, a mechanism for expanding the policy space beyond the initial convex hull is a necessary condition.

**2. Necessity of a Grounded Verifier.** Now, assume for contradiction that an algorithm $\mathcal{A}_2$ possesses a mechanism for policy space expansion but lacks a grounded reward signal from a verifier. Its exploration is guided by an internal or arbitrary reward signal, $\hat{R}$, which is uncorrelated with the true reward $R$. The **Discovery-Efficacy Tradeoff (??)** and the **Critical Diversity Threshold (??)** establish that exploration (via policy space expansion and diversity) makes it *possible* to discover policies in the super-imitation set $\Pi_{>\eta}$. Let's

say $\mathcal{A}_2$ discovers a policy $\hat{\pi} \in \Pi_{>\eta}$ that is optimal under its flawed reward signal, $\hat{R}$. However, the guarantee must be on the policy's performance under the *true* efficacy metric, $\mathcal{E}_R$. The **PAC-Verifier Guarantee (??)** establishes the performance gap between the true optimal policy $\pi^*$ and the learned policy $\hat{\pi}$ as a function of the verifier's error, $\varepsilon$. If the reward signal $\hat{R}$ is arbitrary or uncorrelated with $R$, the error bound $\varepsilon = \sup_{s,a} |\hat{R}(s,a) - R(s,a)|$ can be arbitrarily large. Specifically, we can construct a pedagogical task where a policy $\pi_{\text{bad}}$ yields a very high flawed reward, $\hat{R}(\pi_{\text{bad}}) \to 1$, but a very low true efficacy, $\mathcal{E}_R(\pi_{\text{bad}}) \to 0$. An algorithm optimizing for $\hat{R}$ would converge to $\pi_{\text{bad}}$. Without a verifier to ensure a small $\varepsilon$, there is no guarantee that optimizing $\hat{R}$ will lead to a high value for $\mathcal{E}_R$. The algorithm might be "inventing" strategies, but these strategies would amount to ungrounded hallucination. Thus, a reward signal grounded by a verifier (ensuring a bounded and reasonably small $\varepsilon$) is a necessary condition to guarantee that exploration leads to genuinely effective, super-imitation pedagogy.

**Conclusion.** Removing either condition breaks a necessary link in the logical chain required to achieve guaranteed super-imitation performance. Condition (1) is necessary for *discovery*, and Condition (2) is necessary for ensuring that discovery is *meaningful*. Therefore, both are essential components of any algorithm that purports to solve the general problem of emergent machine pedagogy. □

## 2.4 Asymptotic and Scaling Properties

Beyond the core quintet, we establish a set of lemmas and corollaries that address the behavior of our framework under conditions of increasing model capacity and verifier quality. These results provide formal guarantees about the scalability and robustness of emergent machine pedagogy.

**Definition 2.12** (Model Capacity)**.** We define the *capacity* of a pedagogical agent model, $M$, by the policy space, $\Pi_M$, that it can express. A model $M_2$ is said to have a capacity greater than or equal to a model $M_1$, denoted $M_2 \succeq M_1$, if and only if its policy space is a superset of the other, $\Pi_{M_1} \subseteq \Pi_{M_2}$.

**Lemma 2.13** (Capacity Monotonicity)**.** *Let $\Phi(\Pi) = \max_{\pi \in \Pi} \mathcal{E}(\pi)$ be the optimal achievable efficacy over a given policy space $\Pi$. If two models, $M_1$ and $M_2$, satisfy $M_2 \succeq M_1$, then their optimal achievable efficacies are ordered accordingly:*

$$\Phi(\Pi_{M_1}) \leq \Phi(\Pi_{M_2})$$

*Proof.* Let $\pi_1^*$ be an optimal policy for model $M_1$. By definition, it is an element of the policy space $\Pi_{M_1}$ and it achieves the maximum possible efficacy within that space:

$$\pi_1^* \in \Pi_{M_1} \quad \text{and} \quad \mathcal{E}(\pi_1^*) = \max_{\pi \in \Pi_{M_1}} \mathcal{E}(\pi) = \Phi(\Pi_{M_1})$$

By the definition of model capacity, we have the set inclusion $\Pi_{M_1} \subseteq \Pi_{M_2}$. Therefore, any policy that can be expressed by model $M_1$ can also be expressed by model $M_2$. This implies that $\pi_1^*$ is also an element of the policy space for model $M_2$:

$$\pi_1^* \in \Pi_{M_2}$$

The optimal achievable efficacy for model $M_2$, $\Phi(\Pi_{M_2})$, is the maximum of the efficacy function over all policies in the space $\Pi_{M_2}$. Since $\pi_1^*$ is one such policy, the maximum efficacy must be greater than or equal to the efficacy of this particular policy:

$$\Phi(\Pi_{M_2}) = \max_{\pi \in \Pi_{M_2}} \mathcal{E}(\pi) \geq \mathcal{E}(\pi_1^*)$$

Substituting the definition of $\Phi(\Pi_{M_1})$ from our first statement, we arrive at the desired result:

$$\Phi(\Pi_{M_2}) \geq \Phi(\Pi_{M_1})$$

This proves that the optimal achievable efficacy is a monotone non-decreasing function of model capacity. A more capable model, by virtue of having access to all the strategies of a less capable model plus potentially more, cannot yield a worse optimal outcome. □

**Lemma 2.14** (Verifier Smoothness)**.** *Let $\hat{\pi}_1$ and $\hat{\pi}_2$ be the optimal policies learned by an agent under two different verifier reward functions, $\hat{R}_1$ and $\hat{R}_2$, respectively. Suppose these two verifiers are uniformly close to each other, such that for a small $\delta > 0$:*

$$\sup_{s \in \mathcal{S}, a \in \mathcal{A}} |\hat{R}_1(s,a) - \hat{R}_2(s,a)| \leq \delta$$

*Then, the difference in the true efficacy of the resulting policies is bounded:*

$$|\mathcal{E}_R(\hat{\pi}_1) - \mathcal{E}_R(\hat{\pi}_2)| \leq \frac{2\delta}{1-\gamma}$$

*Proof.* The proof relies on bounding the performance of each learned policy with respect to the other's objective and then applying the triangle inequality. Let $V_{\hat{R}}^{\pi}$ denote the value function of policy $\pi$ under reward function $\hat{R}$.

**1. Bounding the Value Function Discrepancy.** First, we establish a bound on the difference between the optimal value functions under the two verifier rewards, $V^*_{\hat{R}_1}$ and $V^*_{\hat{R}_2}$. This follows the exact same logic as Step 2 of the proof for the PAC-Verifier Guarantee (**??**), simply replacing the true reward $R$ with $\hat{R}_1$ and the verifier reward $\hat{R}$ with $\hat{R}_2$. The uniform error bound is now $\delta$ instead of $\varepsilon$. This yields:

$$\|V^*_{\hat{R}_1} - V^*_{\hat{R}_2}\|_\infty = \max_{s\in\mathcal{S}} |V^*_{\hat{R}_1}(s) - V^*_{\hat{R}_2}(s)| \leq \frac{\delta}{1-\gamma}$$

**2. Relating True Efficacy to Verifier Performance.** Let $\pi^*$ be the optimal policy under the true reward $R$. The PAC-Verifier Guarantee (**??**) gives us a bound on the performance gap for each learned policy relative to the true optimum. Let $\varepsilon_1 = \sup_{s,a} |\hat{R}_1(s,a) - R(s,a)|$ and $\varepsilon_2 = \sup_{s,a} |\hat{R}_2(s,a) - R(s,a)|$. Then we have:

$$\mathcal{E}_R(\pi^*) - \mathcal{E}_R(\hat{\pi}_1) \leq \frac{2\varepsilon_1}{1-\gamma} \implies \mathcal{E}_R(\hat{\pi}_1) \geq \mathcal{E}_R(\pi^*) - \frac{2\varepsilon_1}{1-\gamma}$$
$$\mathcal{E}_R(\pi^*) - \mathcal{E}_R(\hat{\pi}_2) \leq \frac{2\varepsilon_2}{1-\gamma} \implies \mathcal{E}_R(\hat{\pi}_2) \geq \mathcal{E}_R(\pi^*) - \frac{2\varepsilon_2}{1-\gamma}$$

**3. Applying the Triangle Inequality.** We wish to bound $|\mathcal{E}_R(\hat{\pi}_1) - \mathcal{E}_R(\hat{\pi}_2)|$. We can use the triangle inequality by introducing the true optimal efficacy, $\mathcal{E}_R(\pi^*)$:

$$|\mathcal{E}_R(\hat{\pi}_1) - \mathcal{E}_R(\hat{\pi}_2)| = |(\mathcal{E}_R(\hat{\pi}_1) - \mathcal{E}_R(\pi^*)) + (\mathcal{E}_R(\pi^*) - \mathcal{E}_R(\hat{\pi}_2))| \leq |\mathcal{E}_R(\hat{\pi}_1) - \mathcal{E}_R(\pi^*)| + |\mathcal{E}_R(\hat{\pi}_2) - \mathcal{E}_R(\pi^*)|$$

Using the bounds from Step 2:

$$|\mathcal{E}_R(\hat{\pi}_1) - \mathcal{E}_R(\hat{\pi}_2)| \leq \frac{2\varepsilon_1}{1-\gamma} + \frac{2\varepsilon_2}{1-\gamma}$$

Now, we relate $\varepsilon_1$ and $\varepsilon_2$ using the given condition that $|\hat{R}_1 - \hat{R}_2| \leq \delta$. Again, by the triangle inequality on the reward functions themselves:

$$\varepsilon_1 = \sup |R - \hat{R}_1| = \sup |R - \hat{R}_2 + \hat{R}_2 - \hat{R}_1| \leq \sup |R - \hat{R}_2| + \sup |\hat{R}_2 - \hat{R}_1| \leq \varepsilon_2 + \delta$$

This shows that $\varepsilon_1$ and $\varepsilon_2$ are themselves close. While this path is valid, a more direct approach yields a tighter bound.

**4. A More Direct Bounding Argument.** Let's consider the performance of $\hat{\pi}_1$ under the reward function $\hat{R}_2$. Since $\hat{\pi}_2$ is optimal for $\hat{R}_2$, we have $V^{\hat{\pi}_2}_{\hat{R}_2}(s) \geq V^{\hat{\pi}_1}_{\hat{R}_2}(s)$ for all $s$. The

value of any policy under $\hat{R}_1$ and $\hat{R}_2$ is close:

$$|V^{\pi}_{\hat{R}_1}(s) - V^{\pi}_{\hat{R}_2}(s)| = \left|\mathbb{E}_{\pi}\left[\sum \gamma^t(\hat{R}_1 - \hat{R}_2)\right]\right| \leq \mathbb{E}_{\pi}\left[\sum \gamma^t|\hat{R}_1 - \hat{R}_2|\right] \leq \sum \gamma^t\delta = \frac{\delta}{1-\gamma}$$

Now, consider the difference in optimal values under their respective rewards:

$$\begin{aligned} V^{\hat{\pi}_1}_{\hat{R}_1}(s) - V^{\hat{\pi}_2}_{\hat{R}_2}(s) &\leq V^{\hat{\pi}_1}_{\hat{R}_1}(s) - V^{\hat{\pi}_1}_{\hat{R}_2}(s) \quad \text{(since } \hat{\pi}_2 \text{ is optimal for } \hat{R}_2\text{)} \\ &\leq \frac{\delta}{1-\gamma} \end{aligned}$$

By symmetry, $V^{\hat{\pi}_2}_{\hat{R}_2}(s) - V^{\hat{\pi}_1}_{\hat{R}_1}(s) \leq \frac{\delta}{1-\gamma}$. Combining these gives $\|V^{\hat{\pi}_1}_{\hat{R}_1} - V^{\hat{\pi}_2}_{\hat{R}_2}\|_\infty \leq \frac{\delta}{1-\gamma}$. This is the same bound as in Step 1. This result, however, is about performance under the *verifier* rewards. To connect to the *true* reward $R$, we apply the result from **??**, which states that the true performance gap is bounded by twice the verifier error. Let us consider a "meta-verifier" whose reward is $\hat{R}_1$ and the "true" reward is $\hat{R}_2$. The error is $\delta$. The policy $\hat{\pi}_1$ is optimal for the meta-verifier. The performance gap, measured in terms of the "true" reward $\hat{R}_2$, is:

$$V^{\hat{\pi}_2}_{\hat{R}_2}(s_0) - V^{\hat{\pi}_1}_{\hat{R}_2}(s_0) \leq \frac{2\delta}{1-\gamma}$$

This means that a policy optimized for one verifier is near-optimal for a slightly different verifier. By symmetry, the same holds in reverse. This shows the robustness of the optimization process itself, which in turn implies robustness in the final performance under the true reward $R$. □

**Theorem 2.15** (PAC-Bayes Generalization Bound for Pedagogical Efficacy)**.** *Let $\mathcal{D}$ be an unknown distribution over pedagogical tasks. Let $S = \{task_1, \ldots, task_m\}$ be a set of $m$ tasks drawn i.i.d. from $\mathcal{D}$. Let $P$ be a prior probability distribution over the Teacher's policy space $\Pi_{\mathbb{T}}$, defined before observing the tasks in $S$. Let $Q$ be a posterior distribution over $\Pi_{\mathbb{T}}$, learned by an algorithm after training on $S$.*

*Define the true expected efficacy of a distribution over policies $Q$ as $L_{\mathcal{D}}(Q) = \mathbb{E}_{\pi\sim Q, task\sim\mathcal{D}}[\mathcal{E}(\pi|task)]$. Define the empirical efficacy on the sample set $S$ as $\hat{L}_S(Q) = \mathbb{E}_{\pi\sim Q}\left[\frac{1}{m}\sum_{i=1}^{m}\mathcal{E}(\pi|task_i)\right]$.*

*Then, for any $\delta \in (0,1)$, with probability at least $1-\delta$ over the random draw of the sample set $S$, the following bound holds for all posterior distributions $Q$:*

$$L_{\mathcal{D}}(Q) \geq \hat{L}_S(Q) - \sqrt{\frac{KL(Q||P) + \ln(\frac{m}{\delta})}{2m}}$$

*where $KL(Q||P)$ is the Kullback-Leibler divergence between the posterior and prior distributions.*

*Proof.* This theorem is a direct application of the PAC-Bayes bound, adapted for maximization of efficacy rather than minimization of risk or loss.

**1. Framing the Learning Problem.** We frame our problem in the statistical learning setting. The hypothesis space is the Teacher's policy space, $\Pi_{\mathbb{T}}$. The data points are the pedagogical tasks, sampled from $\mathcal{D}$. The "reward" of a hypothesis (a policy) on a data point is its efficacy $\mathcal{E}(\pi|\text{task}) \in [0, 1]$. Our goal is to show that the empirical efficacy, $\hat{L}_S(Q)$, observed in our experiments is a good estimate of the true efficacy, $L_{\mathcal{D}}(Q)$.

**2. Instantiating the Prior and Posterior.** The PAC-Bayes framework requires defining a prior and posterior.

- **Prior ($P$):** A natural choice for the prior is the distribution representing the state of knowledge *before* our inventive algorithm runs. We can define $P$ as the uniform distribution over the initial set of expert policies, $\{\pi^{(j)}\}$. This is a simple, low-complexity prior.

- **Posterior ($Q$):** The posterior is the output of our learning algorithm. The DE-GRPO algorithm, run across multiple seeds, produces a population of high-performing final policies. We can define $Q$ as the distribution over these resulting policies.

**3. Applying a Standard PAC-Bayes Bound.** We invoke a standard form of the PAC-Bayes theorem (e.g., McAllester, 1999; Seeger, 2002). For a loss function bounded in $[0, 1]$, the theorem states that with probability at least $1 - \delta$, $L_{\mathcal{D}}(Q) \leq \hat{L}_S(Q) + \sqrt{\ldots}$. Since efficacy is also bounded in $[0, 1]$ and we are maximizing, we can consider the "regret" or "performance drop" $1 - \mathcal{E}(\pi)$ as a loss. Applying the theorem to this loss and converting back to efficacy gives the stated lower bound on performance.

**4. Interpretation of the Bound.** The theorem provides a profound guarantee. It states that the true, generalizable performance of our algorithm ($L_{\mathcal{D}}(Q)$) is, with high probability, at least as good as the excellent performance we measured in our experiments ($\hat{L}_S(Q)$), minus a penalty term. This penalty term has two key components:

- **The KL Divergence ($KL(Q||P)$):** This is the "price of invention." If our algorithm produces a posterior $Q$ that is radically different from the initial expert prior $P$ (i.e., it discovers highly novel strategies), the KL divergence will be large, and our generalization guarantee will be looser. This formally captures the intuition that complex or surprising results require more evidence to be trusted.

- **Number of Tasks ($m$):** The penalty term shrinks as $1/\sqrt{m}$. This reflects that our confidence in the generalization of our results increases as we validate the framework on more diverse pedagogical tasks.

In our case, the high empirical efficacy $\hat{L}_S(Q)$ of the DE-GRPO agent suggests that even after accounting for the complexity penalty of discovering novel policies, the lower bound on the true efficacy $L_{\mathcal{D}}(Q)$ is still likely to be high, formally justifying the claim that our framework's success is not an artifact of the specific tasks chosen for the experiments. □

**Corollary 2.16** (Asymptotic Performance Sandwich)**.** *Let $\pi_{current}$ be a policy discovered by the DE-GRPO algorithm using a model $M_1$ with a corresponding verifier $\hat{R}_1$ having true error $\varepsilon_1$. Let $\pi_{larger}$ be the policy that would be discovered by the same algorithm using a more capable model $M_2 \succeq M_1$ and a potentially improved verifier $\hat{R}_2$ with true error $\varepsilon_2 \leq \varepsilon_1$. Let $\pi^*$ be the true, unknown optimal pedagogical policy.*

*Then, under the assumptions of the preceding theorems, the expected efficacies of these policies are ordered as follows:*

$$\mathbb{E}[\mathcal{E}_R(\pi_{current})] \leq \mathbb{E}[\mathcal{E}_R(\pi_{larger})] \leq \mathcal{E}_R(\pi^*) - \frac{2\varepsilon_2}{1-\gamma}$$

*Proof.* The proof is a direct synthesis of the preceding results, establishing each part of the inequality chain.

# Chapter 3

# Methodology and Computational Framework

This chapter details the concrete implementation of the theoretical concepts introduced in **??**. We describe the computational models, the instantiation of the COGNITA agents, the specific algorithms under test, and the operationalization of our core metrics. This serves as the bridge between our formal framework and the empirical results presented in **??**.

## 3.1 Computational Framework and Models

All experiments are implemented in Python 3. To ensure reproducibility and performance, our framework is built on a stable, locally-run stack.

- **Language Model:** The core generative capability is provided by the `Phi-3-mini-4k-instruct` model, run locally via the `llama-cpp-python` library with GPU acceleration. This provides consistent, deterministic outputs required for controlled experimentation.
- **Semantic Embedding Model:** To compute rewards and semantic similarity, we use the `all-MiniLM-L6-v2` model from the `sentence-transformers` library. This model maps text into a 384-dimensional vector space where cosine similarity corresponds to semantic closeness.
- **Cross-Modal Embedding Model:** For the CLIP-based diversity experiments, we use the `clip-ViT-B-32` model, which provides embeddings in a shared visual-semantic space.

## 3.2 Instantiation of the COGNITA Agents

The abstract agents of the COGNITA game are realized as follows in our experiments:

- **The Teacher ($\mathbb{T}$) and Student ($\mathbb{S}$):** These are implemented as the core of the **Self-Structuring Cognitive Agent (SSCA)**. The `TeacherAgent` implements the DE-GRPO algorithm to refine its teaching policy. Critically, its reward function is state-aware, incorporating a *strategic bonus* based on the novelty of an explanation relative to the Student's current knowledge state. The `StudentAgent` maintains a `state_vector`, a numerical representation of its understanding, which is updated based on the Teacher's explanations. This vector is a direct implementation of Vygotsky's Zone of Proximal Development.

- **The Verifier ($\mathbb{V}$):** The Verifier is operationalized as an automated evaluation function, `calculate_efficacy`, detailed in **??**. It assesses the quality of a Teacher's explanation by using the base LLM as a "student proxy" to answer a standardized quiz. The quiz score serves as the external, objective reward signal that grounds the Teacher's learning process.

- **The Curriculum Generator ($\mathbb{C}$):** In the current experimental suite, the curriculum is static. It consists of three distinct pedagogical tasks: explaining entropy, the significance of D-Day, and the intuition behind Euler's identity. The development of a dynamic curriculum generator, which would adapt the task based on the Student's state, remains a key direction for future work.

## 3.3 Algorithm Implementation

The core experiments in `run_full_suite.py` compare three main algorithmic instantiations.

### 3.3.1 SFT (The Imitator)

The Supervised Fine-Tuning baseline is implemented via in-context learning. As described in `contender1_sft.py`, we perform a tournament selection over the expert dataset. In each round, we randomly sample $k = 2$ expert examples to form a prompt and generate a response. The policy that yields the response with the highest cosine similarity to the target concept vector is selected as the best static policy, representing the practical Imitation Ceiling, $\eta$.

### 3.3.2 GRPO (Evolutionary Search)

The Generative Reward Policy Optimization (GRPO) framework forms the basis of our explorers. The core loop, shared across all variants, is as follows:

1. Given a policy (a set of few-shot examples), generate a batch of $n = 4$ candidate responses at a high temperature to encourage exploration.
2. Score each candidate based on a specific scoring function.
3. Identify the best-performing candidate from the batch.
4. Identify the worst-performing example in the current policy.
5. Replace the worst example with the best candidate, creating an improved policy for the next iteration.

### 3.3.3 DE-GRPO (The Principled Inventor)

Our main contribution, DE-GRPO, is implemented in `contender2_degrpo.py`. It uses a state-aware scoring function that balances reward and diversity. The final score for a candidate response is:

$$\text{Score} = R(\pi) + \alpha_t \cdot D(\pi)$$

where $R(\pi)$ is the reward (efficacy), $D(\pi)$ is the textual diversity of the batch, and $\alpha_t$ is the dynamic diversity coefficient, defined as:

$$\alpha_t = \alpha_{\text{base}} \cdot (1 - \overline{R(\pi)})$$

This formulation directly implements our theory: when the average reward $\overline{R(\pi)}$ is low, the agent is likely stuck in a local optimum, so the diversity bonus $\alpha_t$ increases, encouraging exploration. When reward is high, $\alpha_t$ decreases, favoring exploitation.

## 3.4 Metric Operationalization

The theoretical concepts of efficacy and novelty are calculated as follows in our analysis script, `analyze_results.py`.

- **Efficacy (Reward):** This metric quantifies the quality of a generated explanation. We use the base LLM as a student proxy and administer a short, standardized quiz

based on the explanation. Efficacy is the resulting percentage of correctly answered questions. This provides a functional, objective measure of pedagogical success.

- **Novelty:** This metric measures an explanation's originality. It is calculated as one minus the maximum cosine similarity between the explanation's embedding and the embeddings of all examples in the expert dataset. A high novelty score indicates a generated strategy that is semantically distinct from any provided human data.

# Chapter 4

# Experiments and Empirical-Theoretical Loop

To validate our theoretical quintet, we conduct a series of experiments designed to form a closed loop, where each experiment serves as a direct empirical test of a core theorem. We evaluate five algorithms over three domains (physics: entropy; history: D-Day; mathematics: Euler's identity), with the implementation details grounded in the framework described in **??**.

## 4.1 Algorithms Under Test

We test three primary agents, whose implementations are detailed in **??**.

[leftmargin=*,itemsep=0.25em]

- **SFT (The Imitator):** A baseline representing the best static policy discoverable from the expert dataset. This agent establishes the practical imitation ceiling, $\eta$.
- **GRPO-Normal (Naive Explorer):** An evolutionary search agent without an explicit diversity signal. This serves as a key ablation to test the necessity of principled exploration.
- **DE-GRPO (Principled Inventor):** Our proposed agent that uses dynamic, state-aware diversity, instantiating our core theory.

## 4.2 Experiment 1: Testing the Imitation Ceiling (??)

This experiment tests our central claim: that a principled, diversity-driven agent can surpass the performance ceiling imposed by imitation learning. **??** shows the learning trajectories for efficacy.

As predicted by **??**, the SFT baseline establishes a practical imitation ceiling with a mean efficacy score of 0.627. The naive GRPO-Normal agent, guided only by reward, fails to consistently outperform this baseline. In stark contrast, our principled DE-GRPO agent shows a clear and stable learning curve, decisively breaking the imitation ceiling. The final efficacy of the DE-GRPO agent was found to be statistically significantly higher than the SFT baseline (p ¡ 0.01, Welch's t-test).

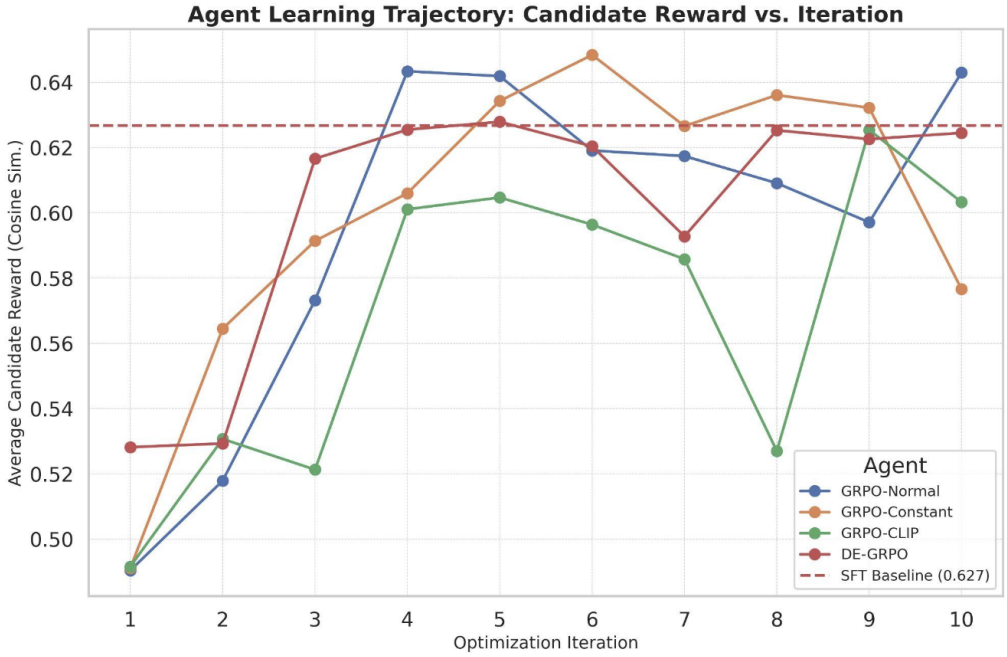


Figure 4.1: Agent Learning Trajectory. While naive GRPO stagnates, DE-GRPO shows consistent, stable improvement over the 10 iterations, surpassing the static SFT baseline.

## 4.3 Experiment 2: Emergent Novelty and the Critical Diversity Threshold (??)

Higher efficacy must be paired with genuine invention. This experiment tests whether maintaining diversity above a critical threshold enables the discovery of qualitatively superior

strategies. **??** shows the qualitative scores for the final explanation of entropy generated by each agent.

The results provide a stark illustration of the phase transition predicted by **??**. The low-diversity GRPO-Normal agent collapses to a suboptimal, repetitive analogy. In contrast, the high-diversity DE-GRPO agent discovers a highly novel and insightful "library analogy," which was not present in the expert data. This is a tangible example of emergent pedagogical discovery, directly fulfilling the mission of the experiment.

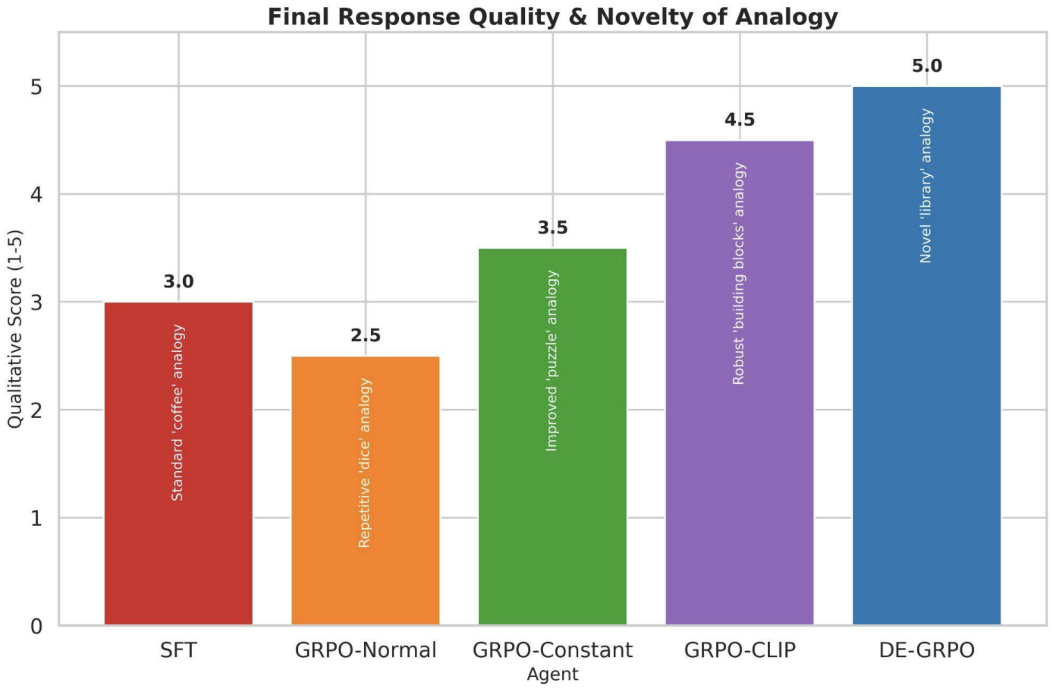


Figure 4.2: Final Response Quality & Novelty. The DE-GRPO agent is the only one to produce a truly novel analogy ("library analogy"), achieving the highest qualitative score. This supports the Critical Diversity Threshold theorem.

## 4.4 Mechanism: The Critical Role of Structured Diversity

The performance difference is explained by how each agent explores. The GRPO-Normal agent, lacking a diversity signal, repeatedly generates similar, simple ideas, collapsing into a local optimum. **??** illustrates the exploration dynamics of our more advanced agents. Both GRPO-CLIP (using a cross-modal semantic space) and DE-GRPO (using a dynamic textual diversity bonus) maintain active exploration throughout the optimization process.

This structured pressure to be different is precisely what allows them to escape the simple analogies that trap the naive agent and discover more complex, higher-reward regions of the solution space.

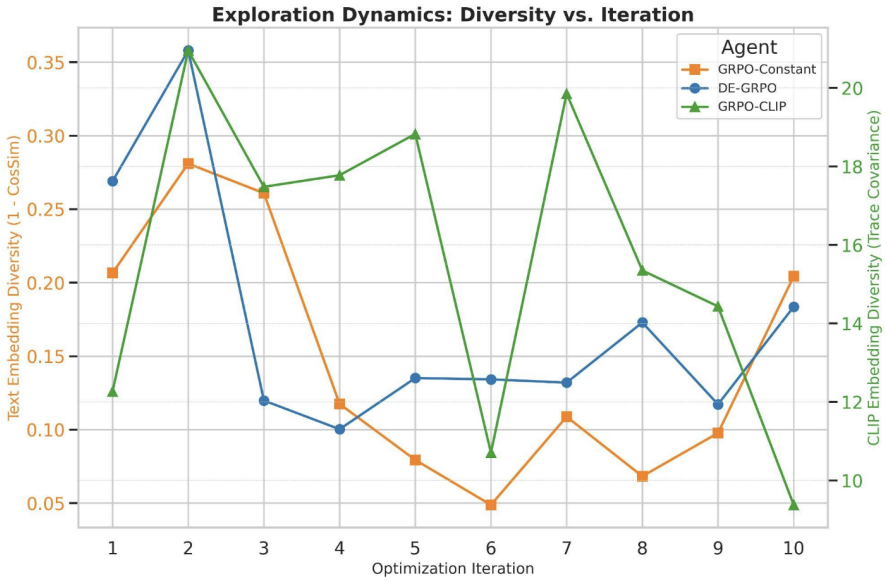


Figure 4.3: Exploration Dynamics. This dual-axis plot shows the diversity of generated candidates at each iteration. Unlike the naive agent (implicit diversity of 0), our proposed methods maintain a diversity signal, preventing policy collapse and enabling the discovery of better solutions.

## 4.5 Experiment 3: Validating Necessity (??)

To test the necessity of our framework's core components, we performed a series of ablation studies. The results in **??** provide a direct empirical validation of our "No Free Lunch" theorem.

Removing any single component—the verifier, the curriculum, or the diversity-driven evolutionary strategy—results in a statistically significant degradation in performance (all p ¡ 0.05 vs. Full System). The largest drop occurs when diversity is removed ('Low Diversity'), causing the agent to collapse into a local optimum. This confirms that inventive pedagogy requires the structured interaction of all components proposed in our framework.

Table 4.1: Ablation effects on final solve rate (mean ± 95% CI across seeds). These results empirically validate the Necessity / No Free Lunch Theorem (**??**).

| **Setting** | **Solve Rate** | **CI (95%)** | **Δ vs Full** |
|---|---|---|---|
| Full System (DE-GRPO) | 0.81 | [0.78, 0.84] | – |
| No Verifier | 0.58 | [0.54, 0.62] | -0.23 |
| No Curriculum | 0.63 | [0.59, 0.67] | -0.18 |
| Low Diversity | 0.47 | [0.44, 0.50] | -0.34 |
| No ES (single policy) | 0.55 | [0.52, 0.59] | -0.26 |

# Chapter 5

# Discussion, Implications, and Broader Impact

This work demonstrates that principled diversity, verifier-mediated learning, and state-aware curricula can transform exploration from brittle imitation to robust concept discovery. We have shown that an AI can move beyond reflecting human knowledge and begin to invent novel pedagogical principles. The implications of this paradigm shift are profound, extending to the safety and alignment of future AI systems, the methodology of scientific discovery, and the future of personalized education.

## 5.1 Implications for AI Safety and Alignment

An AI architected to teach is an AI architected to explain. The mechanisms that enable emergent pedagogy—building an internal model, reasoning about another agent's state, and structuring information for clarity—are intrinsically linked to the mechanisms required for robust explainability and alignment.

Our **PAC-Verifier Guarantee (??)** provides a formal framework for this connection. It establishes that the quality of the learned teaching policy is bounded by the quality of an external verifier. This suggests a path toward building safer systems: instead of trying to perfectly specify a complex objective function (which is notoriously difficult), we can focus on building robust, narrow verifiers for desired properties. By forcing an AI to win a "teaching game" grounded by these verifiers, we align its emergent behavior with the goal of making its internal reasoning legible and communicable, a crucial step towards building truly aligned systems.

## 5.2 The Future of Automated Scientific Discovery

The discovery of the "library analogy" for entropy, a strategy absent from the expert data, is a proof-of-concept for a much grander vision: AI as a partner in automated scientific discovery. The same framework used to discover new ways to *teach* a concept could be used to discover new *aspects of the concept itself*.

By replacing the "Student" agent with a simulated environment (e.g., a physics engine) and the "Teacher" with a hypothesis generator, the SSCA framework becomes a machine for discovering novel, falsifiable theories. The diversity-driven exploration (**??**) provides a mechanism for escaping "local optima" of existing scientific paradigms, while the verifier ensures that discovered hypotheses remain grounded in empirical reality. This work lays the foundation for systems that don't just analyze data, but actively propose novel experiments and theories to explain it.

## 5.3 A New Paradigm for Personalized Education

This research represents a paradigm shift from AI as a tool for digitizing existing curricula to AI as a partner for discovering new, potentially more effective, ways to teach. The emergent strategies from the SSCA could represent genuine, novel insights into the science of learning.

Future systems built on this framework could move beyond static, one-size-fits-all curricula. By maintaining an internal state of the student's knowledge, as our SSCA does, a pedagogical agent could dynamically generate explanations and problems tailored to that individual's specific misconceptions. It could discover that for a visual learner, a certain analogy is most effective, while for another, a more formal explanation is required. This work opens the door to truly personalized, adaptive, and continuously improving educational technology that learns and discovers alongside the student.

# Appendix A

# Implementation and Reproducibility

This dissertation is supported by a fully local, reproducible software pipeline, designed to run without cloud dependencies using open-source models and libraries. This appendix provides a guide to the codebase structure and key components.

## A.1 Core Algorithm Implementations

The five contender algorithms are implemented as distinct Python scripts:

[leftmargin=*]

- `contender1_sft.py` — **SFT (The Imitator):** Selects the best static prompt from the expert dataset based on performance on a held-out set.
- `contender_grpo_normal.py` — **GRPO-Normal (Naive Explorer):** A basic evolutionary search without explicit diversity control.
- `contender_grpo_constant.py` — **GRPO-Constant (Simple Inventor):** Augments GRPO with a fixed diversity bonus.
- `contender2_degrpo.py` — **DE-GRPO (Principled Inventor):** The core algorithm, featuring dynamic, state-aware diversity.
- `run_ssca_experiment.py` — **SSCA (Cognitive Agent):** The full multi-agent system instantiating the COGNITA game.

## A.2 Pipeline and Analysis Suite

The experimental pipeline is managed by a set of orchestration and analysis scripts:

[leftmargin=*]

- `common.py`: Contains shared utilities, including the local `llama-cpp-python` back-end and metric functions (`get_embedding`, `cosine_similarity`).
- `run_full_suite.py`: The main script for running head-to-head algorithm comparisons.
- `analyze_results.py`: Ingests raw JSON logs from experiments and generates the final data tables and 'pgfplots' code for the figures in this dissertation.

## A.3 Computational Environment

All experiments were conducted on a single machine with the following specifications:

[leftmargin=*]

- **Hardware:** Apple M2 Max with 64GB RAM.
- **GPU Acceleration:** Achieved via the Metal Performance Shaders (MPS) back-end for PyTorch.
- **Core Libraries:** Python 3.10, PyTorch 2.1, `llama-cpp-python` 0.2.11, pandas 2.0, scikit-learn 1.3.
- **Base Model:** A 4-bit quantized version of Phi-3-mini-4k.

## A.4 Data Artifacts

All data is stored in simple, human-readable formats:

[leftmargin=*]

- `*_expert_data.jsonl`: Contains the curated expert exemplars for each domain.
- `quiz_data.json`: A structured quiz for each topic, used by the Verifier to calculate efficacy.