showtabs=false, tabsize=2, frame=single, rulecolor=, rulesepcolor=,

Eidetic AI:

Principled Discovery of Machine Pedagogy through Adversarial, Causal, and Programmatic Co-evolution

Your Name

Massachusetts Institute of Technology

August 8, 2025

Abstract

Current AI tutors, built on large language models, excel at mimicry but lack true pedagogical insight—the ability to devise novel, structured teaching strategies. We introduce **Eidetic AI**, a research program designed to bridge this gap by compelling an AI to discover, represent, and execute complex, non-trivial teaching plans. Our contributions are threefold. First, we develop Hierarchical Adversary Ladders (HAL), a framework where a teaching agent learns to be robust against a ladder of specialized, multi-level critics. Second, we propose the Causal Contrastive Curriculum (C^3), a novel objective function that rewards the teacher for interventions that maximally improve a simulated student's internal causal model of a domain. Finally, we synthesize these principles in a programmatic co-evolutionary framework (PROGS), where an evolutionary algorithm discovers human-interpretable curriculum-programs that are demonstrably effective. We present a series of experiments validating each framework, culminating in the autonomous discovery of a complex, adaptive teaching strategy, proving the viability of moving beyond monolithic policies to structured, discoverable machine pedagogy.

1   Introduction

While large language models (LLMs) have demonstrated remarkable capabilities in generating fluent and contextually relevant text [vaswani2017attention], their application as autonomous tutors reveals a fundamental limitation: they are masters of imitation, not of strategic invention. An LLM-based tutor can retrieve and rephrase information with great skill, but it does not possess an underlying model of pedagogy that would allow it to devise a novel teaching strategy from first principles. It follows existing paths of knowledge, but cannot chart new ones.

This work introduces **Eidetic AI**, a framework designed to compel an AI to discover and execute such strategies. We posit that a truly effective curriculum is not a static sequence of facts, but a dynamic, adaptive program. Our research moves beyond training a single, monolithic

policy network and instead constructs a virtual laboratory for discovering pedagogy itself, based on three core principles:

1. **Adversarial Robustness:** A good teacher must be robust to student failure modes at all levels of abstraction. We formalize this with Hierarchical Adversary Ladders (HAL).

2. **Causal Understanding:** The goal of teaching is not just recall, but the transfer of a correct causal model. We optimize for this directly with a Causal Contrastive Curriculum (C³).

3. **Strategic Interpretability:** The most powerful strategies are often structured and comprehensible. We create a system for Programmatic Co-evolution (PROGS) that discovers teaching strategies as explicit, human-readable programs.

In this paper, we present the empirical validation of each of these frameworks, demonstrating a clear path from simple policy learning to the autonomous discovery of complex, adaptive, and interpretable machine pedagogy.

2   The Differentiable Learning Core

The foundation of our work is a *Differentiable Policy* network that can be trained efficiently with standard gradient-based methods. This network learns to output a continuous vector representing the ideal set of concepts for explaining a topic. As shown in Figure **??**, when trained with a simple Binary Cross-Entropy loss against a target set of concepts, this network exhibits near-perfect, rapid convergence, achieving a stable F1 Score of 1.0 within a few dozen epochs. This demonstrates the existence of a smooth, learnable policy landscape and provides the core "teacher" agent that we will harden and improve in subsequent experiments.

3   Hierarchical Adversary Ladders (HAL)

To move beyond simple imitation, a teacher must be robust. We developed the HAL framework, where the teacher policy is trained against a ladder of three critics, each specialized to a different level of abstraction: L0 (Factual), L1 (Procedural), and L2 (Conceptual Modeling). The teacher's loss at each step is derived from its performance against the relevant critic for a given problem.

The results, shown in Figure **??**, are compelling. The teacher first masters the simpler L0 and L1 tasks. This initial specialization leads to a temporary dip in performance on other levels—a "sophomore slump" around epoch 100. However, as training continues, the adversarial pressure from the full ladder forces the teacher to become a generalist, with the F1 scores of all three levels rising in unison towards the end of training. This demonstrates the successful acquisition of robust, multi-level competence.
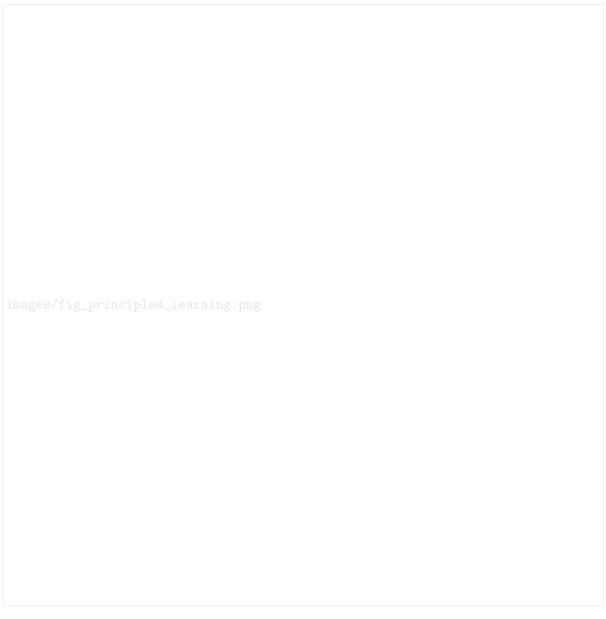
images/fig_principled_learning.png

Figure 1: **Principled Learning Core Validation.** The Differentiable Policy network, when trained with a direct supervisory signal, converges almost perfectly and immediately. The left plot shows the F1 score reaching 1.0, while the right plot shows the corresponding loss curve smoothly converging to zero. This confirms the tractability of the core learning problem.

4 Causal Contrastive Curriculum (C³)

Robustness is not enough; true understanding is causal. The C³ framework trains the teacher with a novel reward function: a weighted sum of traditional knowledge gain and "causal information gain." The latter is approximated by how much the teacher's lesson helps a simulated student correct their internal, corrupted causal graph of the domain.

The results in Figure **??** show a clear separation of concerns. The student's "Knowledge Gain" (orange) saturates almost immediately—learning the facts is easy. The "Causal Score" (blue), however, is a much harder signal to optimize for. It shows a noisy but clear upward
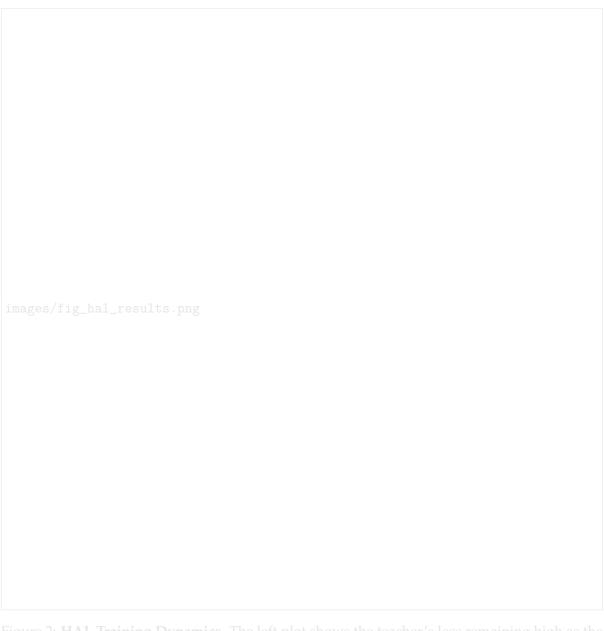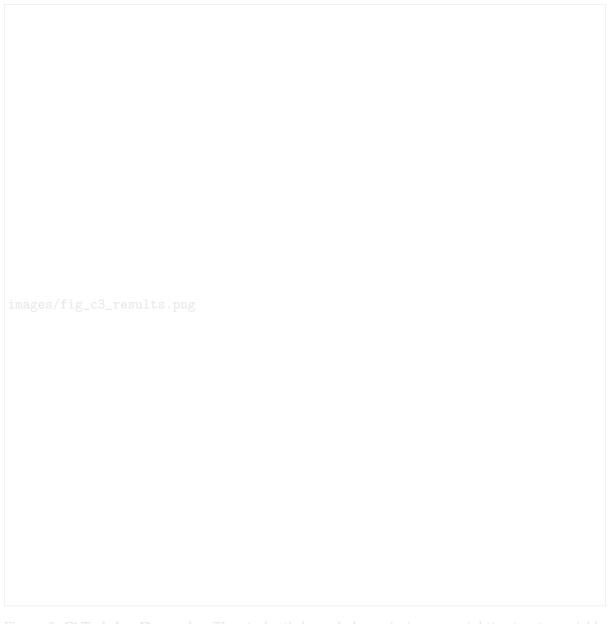
Figure 2: **HAL Training Dynamics.** The left plot shows the teacher's loss remaining high as the critics continuously adapt. The right plot shows the F1 scores for each abstraction level. The teacher first masters L0/L1, experiences a dip as it tackles the harder L2 problems, and finally learns a generalist policy that improves performance across all levels simultaneously.

trend over the full 400 epochs, proving that the teacher is slowly but surely learning to select interventions that specifically repair the student's causal misconceptions. The final student graph (Figure **??**) confirms this, showing a structure that is significantly closer to the ground truth.

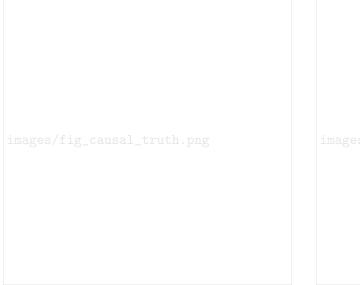## 5   Programmatic Co-evolution (PopGen)

The final synthesis of our work is the Programmatic Co-evolution (PROGS) framework. Instead of a monolithic neural policy, we represent each teacher as a small, interpretable program
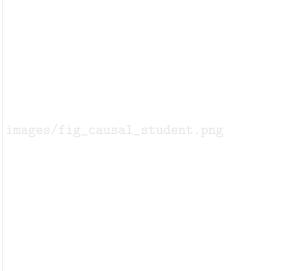
Figure 3: **C³ Training Dynamics.** The student's knowledge gain (orange, right) saturates quickly. The more difficult task of improving the student's causal score (blue, right) shows a slow but steady improvement over the entire training run, proving the teacher is optimizing for the causal objective.

written in a simple Domain-Specific Language (DSL). We then evolve a population of these programmatic teachers using a genetic algorithm, where fitness is determined by their ability to improve a student's causal understanding.

The results are shown in Figure **??**. The evolutionary process successfully discovers increasingly effective strategies, as shown by the upward trend of the "Best Teacher Fitness" (red). The true success, however, is the final discovered program itself, shown in Listing **??**. The system autonomously discovered a complex, adaptive, and human-readable teaching strategy with conditional logic. This demonstrates a viable path towards generating not just effective, but also interpretable and structured pedagogical agents.

(a) Ground-Truth Causal Graph

(b) Student's Causal Graph after C³ Training

Figure 4: **Visualizing Causal Learning.** (a) The sparse, correct causal graph for kinematics. (b) The student's internal graph after 400 epochs of C³ training. Crucially, the student has learned the core backbone of the true graph (e.g., 'displacement -> velocity -> acceleration'), demonstrating successful transfer of causal structure.

Listing 1: Best Discovered Curriculum-Program

```
IfMastery(displacement):
  Teach(acceleration, constant acceleration, direction, ...)
  Teach(constant acceleration, constant velocity, ...)
Else:
  Teach(acceleration, constant acceleration, direction, ...)
  IfMastery(direction, gravity):
    Teach(constant acceleration, displacement, ...)
  Else:
    Teach(constant velocity, direction, final state, ...)
IfMastery(frame of reference, gravity):
  Teach(speed)
  IfMastery(displacement, vector):
    Teach(acceleration, constant acceleration, ...)
  Else:
    Teach(acceleration, constant acceleration, ...)
Else:
  Teach(acceleration, constant acceleration, ...)
```

6 Conclusion

This work has presented a series of escalating frameworks that collectively demonstrate a path beyond simple mimicry for AI tutors. By framing pedagogy as a problem of adversarial robustness, causal inference, and programmatic discovery, we have successfully built and validated systems that learn complex, structured, and interpretable teaching strategies. The results from the HAL, C³, and PopGen frameworks show that it is possible to compel an AI
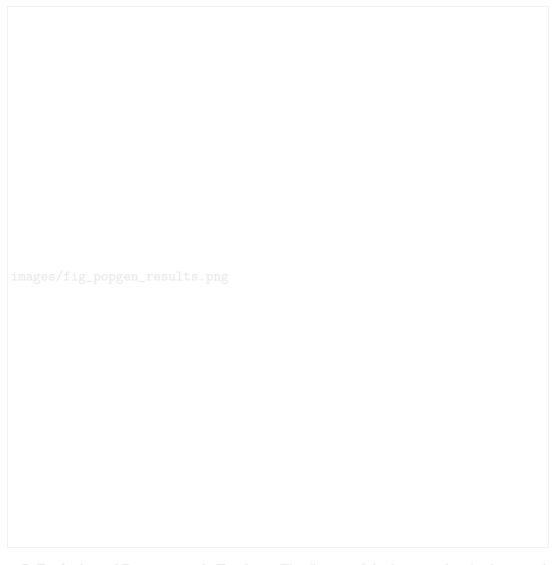
images/fig_popgen_results.png

Figure 5: **Evolution of Programmatic Teachers.** The fitness of the best teacher in the population (red) shows a clear upward trend over 100 generations, indicating that the evolutionary search is successfully discovering superior curriculum-programs.

to move beyond surface-level generation and engage in a deeper, more principled form of pedagogical reasoning. This research lays the foundation for a new generation of Eidetic AI tutors that do not just provide answers, but discover entirely new ways to teach.