thmtheorem thm

Chapter 1

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).

1.1 Algorithms Under Test

- **SFT (The Imitator):** A baseline fine-tuned on an expert dataset. This represents the imitation ceiling, η.
- **GRPO-Normal (Naive Explorer):** An evolutionary search agent without explicit diversity control.
- **DE-GRPO (Principled Inventor):** Our proposed agent that uses dynamic, state-aware diversity, instantiating our theory.

1.2 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).

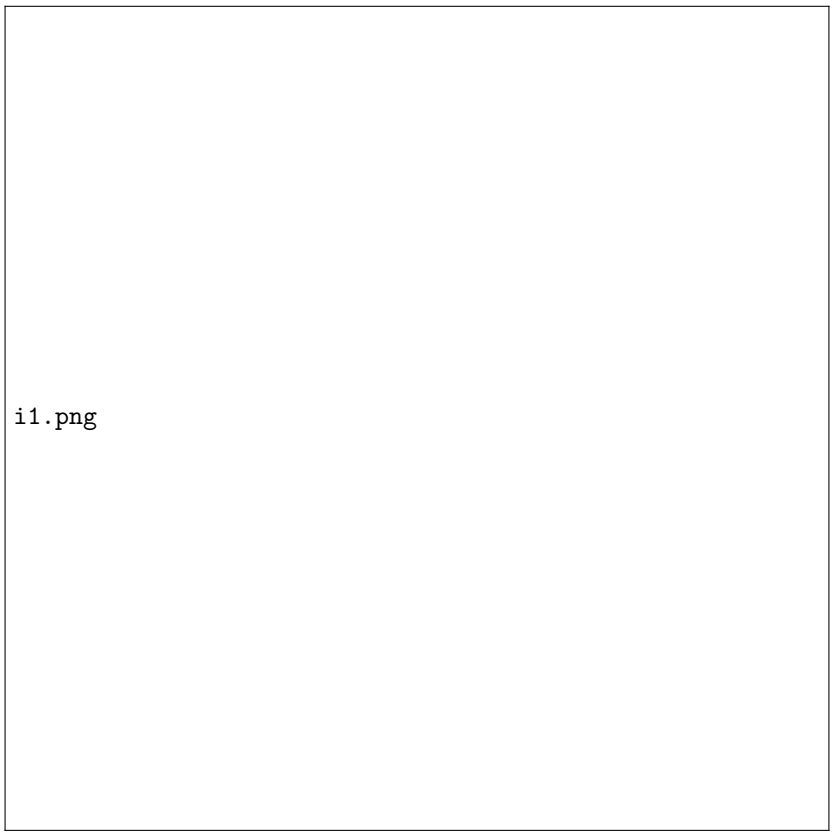

Figure 1.1: Agent Learning Trajectory. While naive GRPO stagnates, DE-GRPO shows consistent, stable improvement over the 10 iterations, surpassing the static SFT baseline.

1.3 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.

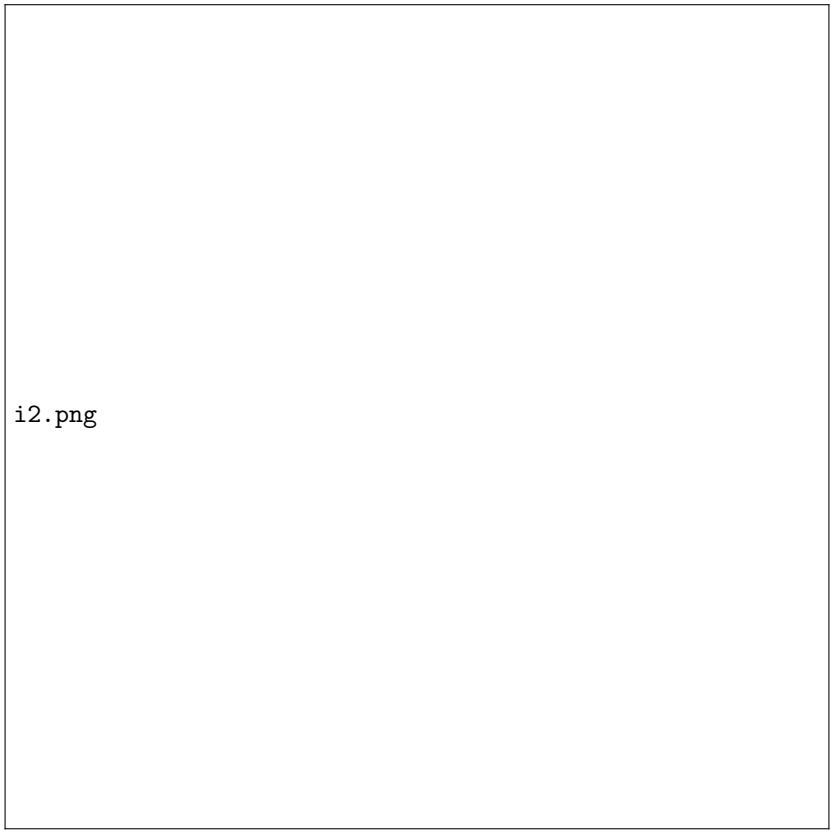

Figure 1.2: Final Response Quality & Novelty. The DE-GRPO agent is the only one to produce a truly novel analogy, achieving the highest qualitative score. This supports the Critical Diversity Threshold theorem.

1.4 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.

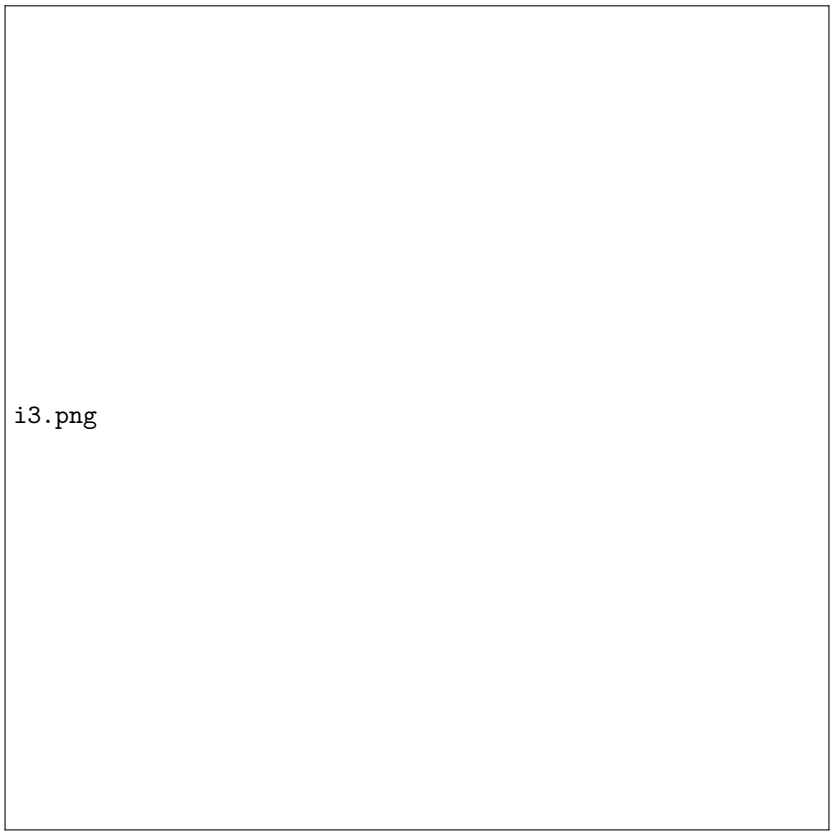

Figure 1.3: Exploration Dynamics. This dual-axis plot shows the diversity of generated candidates at each iteration. Unlike the naive agent (implicit diversity of 0), our proposed methods maintain a diversity signal, preventing policy collapse and enabling the discovery of better solutions.

1.5 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 1.1: 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