POLICY OPTIMIZATION PREFERS THE PATH OF LEAST RESISTANCE

Anonymous authors
Paper under double-blind review

ABSTRACT

Policy optimization (PO) algorithms are used to refine Large Language Models (LLMs) for complex, multi-step reasoning. Current state-of-the-art pipelines enforce a strict think-then-answer format to elicit chain-of-thought (CoT); however, the behavior of PO when these rigid constraints are relaxed into an open-ended CoT structure remains an under-studied question. We investigate this gap with an extensive suite of controlled experiments and identify a powerful principle: *policy optimization consistently follows the path of least resistance*. When afforded the flexibility to interleave reasoning and response, policy optimization consistently learns to discard explicit reasoning, causing the policy to degenerate to a direct <answer>-only format. This outcome holds true across a rigorous evaluation suite spanning 5 model families (4B-24B), 3 reasoning domains (math, code, logic), and 3 distinct PO algorithms (GRPO, DAPO, REINFORCE++). We find that this collapse in format is persistent even when the complex <think><answer> format is assigned up to 8x larger reward weights. We formalize this principle through a series of controlled reward decomposition experiments, demonstrating a clear hierarchy: PO systematically optimizes for the simplest reward component first, a preference that holds even when faced with mutually exclusive choices or strong incentives for more complex behaviors. Finally, we show that successful convergence on the high-reward shortcut is not a low-effort drift but is driven by the optimization process that requires the KL-regularized policy to have sufficient freedom to make a significant shift from its initial prior. Our findings reveal that granting policies the freedom to diverge is a double-edged sword: while necessary for discovering high-reward shortcuts, it also creates a powerful incentive to game the simplest aspects of the reward function, posing a critical challenge for reward hacking under alignment.

1 INTRODUCTION

Policy Optimization Yu et al. (2025); Shao et al. (2024); Liu et al. (2025c); Yue et al. (2025); Liu et al. (2025a) has emerged as the principal tool for refining Large Language Models Team et al. (2025); Qwen et al. (2025); Jiang et al. (2023); Grattafiori et al. (2024) towards complex, multi-step reasoning. The community's dominant approach is one of careful enforcement: to elicit a chain-of-thought, we engineer a rigid reward function that strongly compels the model to follow a strict "think-then-answer" format Liu et al. (2025a); DeepSeek-AI et al. (2025); Xie et al. (2025); Chen et al. (2025). While this enforced structure is effective in producing a desired output, it obscures a deeper, more fundamental question that has been largely overlooked: What is the optimizer's *innate preference* when the external guidance is removed? If we grant the model the freedom to choose its own path to a solution, what path does it take? This question is not merely academic. If the optimizer possesses a powerful intrinsic bias, then our current methods of alignment may be working against a fundamental force, leading to brittle, inefficient, and unpredictable training dynamics. Understanding this preference is therefore a critical, yet under-studied, prerequisite for building truly robust and reliable reasoning systems.

Our investigation, therefore, began with a simple act of liberation. We designed a composite reward function that, for the first time, offered the model a genuine choice. Instead of a single, valid path, our function rewarded any number of interleaved <think> and <answer> blocks, and crucially, also rewarded a direct <answer>-only format. This seemingly minor change from enforcement to

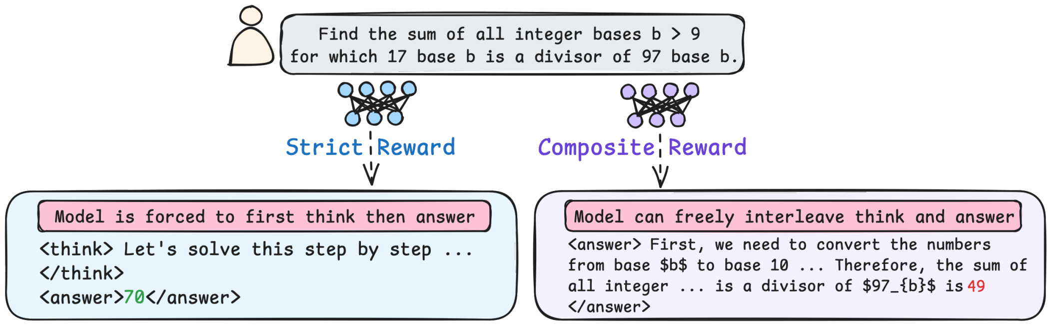

Figure 1: We compare the standard Strict Reward paradigm **(left)**, which enforces a rigid *think-then-answer* structure, against our novel Composite Reward **(right)**, which grants the model the freedom to choose its solution path. The central finding investigated in this paper is the emergent policy learned under this freedom: the model consistently bypasses explicit reasoning to converge on the simplest valid format, a phenomenon we term the *Cognitive Shortcut.*

choice led to a striking anomaly. Across a rigorous evaluation suite spanning diverse architectures (Gemma-3 Team et al. (2025), Qwen-2.5 Qwen et al. (2025), Llama-3.1 Grattafiori et al. (2024), Ministral Jiang et al. (2023), Yi AI et al. (2025)), scales (4B to 12B), and PO algorithms (GRPO Shao et al. (2024), DAPO Yu et al. (2025), REINFORCE++ Hu et al. (2025)), the policy invariably discarded complexity. On challenging domains from mathematics (GSM8K, MathHard) to coding (rStar-Coder) and logic (ReClor), the outcome was identical: the structured thoughts vanished, and the model converged on the simplest possible path. This powerful, emergent preference for a "Cognitive Shortcut" became the central mystery we sought to solve.

This initial finding sparked a natural line of inquiry. The model's preference for the simplest format was clear, but was this an "all-or-nothing" choice, or was there a more nuanced structure to this preference? Figure 1 This led us to our next hypothesis: if the optimizer is biased towards simplicity, perhaps it tackles complex objectives not holistically, but by first conquering their simplest components. To test this, we moved from a simple choice to a structured hierarchy. We designed a controlled "Reward Hierarchy" experiment using three nested reward formats of ascending difficulty $r_1 < r_2 < r_3$, all yielding the same reward magnitude. The result was a stunningly predictable sequence of learning: the optimizer first mastered the simplest format r_1, and only after this reward was saturated did it begin to make progress on r_2. The most complex format, r_3, was never learned. This revealed that the "Principle of Least Resistance" is not just a preference, but a sequential law.

This discovery, however, raised an even more pressing question: just how powerful is this law? Is it a mere tie-breaker, or a dominant force that can override other incentives? To quantify its strength, we returned to our reward hierarchy, but this time we offered exponentially larger rewards for mastering the more complex formats. We found that the model would consistently forgo significant rewards to remain on the simpler path. Only when the incentive for complexity became overwhelmingly large did we observe a *phase transition* where the optimizer was finally bribed into tackling the harder task. The preference for simplicity, we realized, was a powerful, quantifiable force in the optimization landscape.

Finally, with the behavioral law and its strength firmly established, we turned to the ultimate question: why does this law exist? This question has taken on a new urgency. A prominent trend in state-of-the-art policy optimization is the removal of conservative constraints like the KL penalty Yu et al. (2025); Yue et al. (2025), granting models unprecedented freedom to explore the reward landscape. The common intuition is that this freedom simply allows for more effective reward maximization Liu et al. (2025a). Our investigation, however, reveals a more complex and cautionary reality. By treating the KL divergence as a diagnostic for this exploratory freedom, we found that successful convergence on the *Cognitive Shortcut* is not a low-effort drift. Instead, it is an aggressive optimization process that requires a large and decisive policy shift away from the model's pre-trained priors. The path of least resistance is not the path of lowest policy shift, but the path carved by the most powerful and stable gradient signal, a form of reward hacking. This finding suggests that unconstrained exploration, while powerful, may have unintended and problematic consequences.

Our work makes the following contributions:

❶ **A New Fundamental Principle:** We identify, formalize, and empirically validate the *Principle of Least Resistance*, a powerful predictive law governing the behavior of policy optimization in LLMs.

❷ **Rigorous Ablation Testing:** We move beyond simple observation, stress-testing our principle with a motivated sequence of novel, controlled experiments (Reward Hierarchies, Exponential Gambits) that quantify its strength and universality.

❸ **A Counter-Intuitive KL Perspective:** We provide a new lens for understanding PO dynamics, demonstrating that successful convergence on simple shortcuts requires a high-KL policy shift, thereby linking learnability to the freedom to escape the inertia of priors.

2 Setup

To rigorously test our hypothesis, we designed a comprehensive experimental suite. Our methodology was guided by two core principles: **diversity**, to ensure our findings are general and not artifacts of a specific domain or model; and **relevance**, to use tasks that are widely recognized as benchmarks for complex, multi-step reasoning. We selected six powerful, publicly available models from five distinct architectural families, with scales ranging from 4 billion to 24 billion parameters. Our suite included Gemma-3 (4B & 12B) Team et al. (2025), Qwen-2.5 (7B) Qwen et al. (2025), Llama-3.1 (8B) Grattafiori et al. (2024), Ministral (8B) Jiang et al. (2023), and Yi (6B) AI et al. (2025). This diversity ensures our conclusions are not an artifact of a specific model's pre-training or architecture but are a general property of these systems. All experiments were conducted on a cluster of 3 NVIDIA RTX A6000 GPUs, with 48GB VRAM each, and a single NVIDIA H100 GPU, with 80GB VRAM.

Datasets. To rigorously test our "Principle of Least Resistance", we curated a diverse gauntlet of datasets spanning three critical reasoning domains. For mathematical reasoning, we used a tiered selection from the foundational multi-step arithmetic of **GSM8K** Cobbe et al. (2021) to the more complex algebraic challenges in **Math-Hard** Hendrycks et al. (2021) and the rich, SOTA traces of **Open-R1 Math 220k** Hugging Face (2025). To evaluate algorithmic and code reasoning, we leveraged the competition-level complexity of **Microsoft's rStar-Coder** Liu et al. (2025b) alongside the functionally verifiable problems in **Open R1 verifiable coding** Hugging Face (2025). Finally, for logical and deductive reasoning, we tested our models on the canonical suppositional puzzles of **Knights and Knaves** Xie et al. (2024), the natural language deduction required by **ReClor** Yu et al. (2020), and the sequential problem-solving of a specialized **planning-mystery** dataset. This multi-domain, multi-difficulty suite was designed to ensure that our findings are a general principle of optimization, not an artifact of a single task. **Policy Optimization Algorithms.** Our findings are not an artifact of a single training algorithm. To establish that the "Principle of Least Resistance" is a feature of the PO paradigm itself, not a quirk of one implementation, we replicated our experiments across three distinct and powerful families of policy optimization algorithms. Our selection was designed to cover a range of modern techniques, from highly-engineered systems to bias-corrected standards.

First, we employed **DAPO (Decoupled clip and Dynamic sampling Policy Optimization)** Yu et al. (2025), a state-of-the-art system designed for stable, large-scale training of reasoning models. Its objective function is given by:

$$\begin{aligned}\mathcal{J}_{\text{DAPO}}(\theta) = \mathbb{E}_{(q,a)\sim\mathcal{D},\{o_i\}_{i=1}^{G}\sim\pi_{\theta_{\text{old}}}(\cdot|q)} \\ \left[\frac{1}{\sum_{i=1}^{G}|o_i|}\sum_{i=1}^{G}\sum_{t=1}^{|o_i|} \min\left(r_{i,t}(\theta)\hat{A}_{i,t}, \text{clip}(r_{i,t}(\theta), 1-\epsilon_{\text{low}}, 1+\epsilon_{\text{high}})\hat{A}_{i,t}\right)\right]\end{aligned} \tag{1}$$

$$\text{s.t. } 0 < |\{o_i \mid \text{is\_equivalent}(a, o_i)\}| < G \tag{2}$$

We further used **Dr. GRPO (Group Relative Policy Optimization Done Right)** Liu et al. (2025c). We specifically chose the "Dr." variant over the original GRPO for its response length and question difficulty bias correction, which provides a more stable and accurate learning signal. Its objective is:

$$\mathcal{J}_{\text{Dr.GRPO}} = \frac{1}{G}\sum_{i=1}^{G}\frac{1}{|o_i|}\sum_{t=1}^{|o_i|}\min\left\{r_{i,t}(\theta)\hat{A}_{i,t}, \text{clip}\left(r_{i,t}(\theta), 1-\epsilon, 1+\epsilon\right)\hat{A}_{i,t}\right\}$$

Table 1: **Performance Comparison of Training Paradigms Across Diverse Models and Reasoning Tasks.** We report the final accuracy (%) of models trained with a **Strict Reward** (enforcing the `<think><answer>` format) versus our **Composite Reward** (allowing a choice of formats). The results are notably mixed across all four reasoning domains, with no single paradigm consistently outperforming the other. This ambiguity suggests that a simple accuracy comparison is insufficient to understand the underlying learning dynamics, motivating a deeper investigation into the optimizer's intrinsic preferences. Best performance in each pair is highlighted in bold.

	GSM8K		rStar-Coder		ReClor		Planning-Mystery	
Model	**Strict**	**Composite**	**Strict**	**Composite**	**Strict**	**Composite**	**Strict**	**Composite**
Gemma-3 4B	72.4	**73.1**	**55.8**	53.5	**73.2**	69.4	**58.3**	55.9
Qwen-2.5 7B	**92.4**	85.5	73.2	**73.6**	**80.1**	73.8	**77.5**	73.8
Llama-3.1 8B	**86.2**	82.9	**64.1**	59.2	**78.3**	72.7	**74.9**	70.6
Ministral 8B	**89.5**	84.8	**74.0**	72.6	**81.0**	76.5	76.8	**77.3**
Yi 6B	**85.1**	83.8	**68.3**	63.7	**68.4**	65.1	**71.2**	67.7
Gemma-3 12B	**94.6**	86.4	**76.1**	72.3	**85.5**	81.2	**81.2**	76.5

where the advantage $\hat{A}_{i,t}$ is a per-sequence reward baseline: $\hat{A}_{i,t} = R(\mathbf{q}, \mathbf{o}_i) - \mathrm{mean}(\{R(\mathbf{q}, \mathbf{o}_1), \ldots, R(\mathbf{q}, \mathbf{o}_G)\})$.

Finally, to connect our findings to the foundational principles of policy gradients, we included **REINFORCE++** Hu et al. (2025), a robust and modern variant of the classic REINFORCE algorithm. The consistent emergence of our observed phenomenon across these three distinct algorithmic philosophies provides strong evidence that the "Path of Least Resistance" is a fundamental property of the policy optimization paradigm, independent of the specific implementation.

3 Path of Least Resistance

3.1 Flexible Format Reward

Our investigation begins with a simple yet profound departure from the status quo. The prevailing methodology in policy optimization for reasoning tasks enforces a rigid structure on the model, effectively mandating a specific computational path. We hypothesized that this enforcement might be obscuring the optimizer's intrinsic biases. To test this, we asked a fundamental question: *What path does the optimizer choose when it is given a choice?*

To answer this question, we first needed to formally define the choice. Let a model generation be a sequence of tokens y. The standard, **Strict** reward function, $R_{\text{strict}}(y)$, provides a positive reward only if y perfectly matches the `think-then-answer` format:

$$R_{\text{strict}}(y) = \begin{cases} 1 & \text{if } y \in \texttt{<think>.*</think>\textbackslash s*<answer>.*\textbackslash boxed\{.*\}</answer>\$} \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

This function defines a single, narrow path to success. To create a choice, we designed a **Composite** reward function, $R_{\text{composite}}(y)$, which defines a much larger set of valid, rewarded formats. This function allows for any number of interleaved `<think>` and `<answer>` blocks, and crucially, also accepts a direct `<answer>`-only format without any preceding thought:

$$R_{\text{regex}}(y) = \begin{cases} 1 & \text{if } y \in \texttt{\^{}(((<think>.*</think>\textbackslash s*<answer>.*</answer>\textbackslash s*)+|} \\ & \quad \texttt{(<answer>.*</answer>\textbackslash s*(<think>.*</think>} \\ & \quad \texttt{\textbackslash s*<answer>.*\textbackslash boxed\{.*\}</answer>\textbackslash s*)*))\$} \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

Our initial expectation was that the model, now liberated from its strict format, would learn a nuanced policy, perhaps using longer `<think>` blocks for harder problems and skipping them for easier problems. The reality was far more dramatic and revealing.

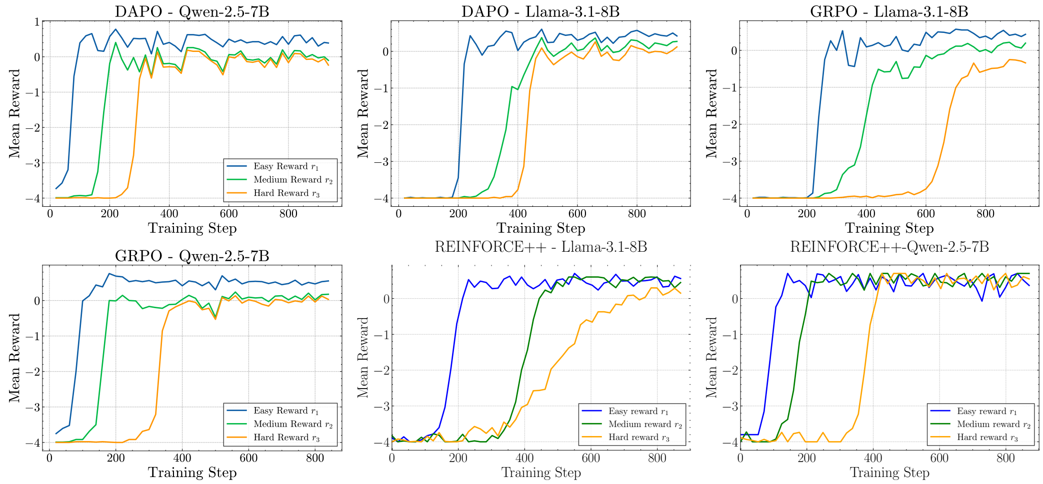

Figure 2: Sequential optimization of format rewards under the nested Reward Hierarchy experiment. Across different models (Qwen-2.5-7B, Llama-3.1-8B) and PO algorithms (DAPO, GRPO, REINFORCE++), the learning dynamics are remarkably consistent. The policy first masters the innermost, easiest reward r_1, then the medium reward r_2, and finally the outermost, hardest reward r_3.

3.2 Universal Convergence to Minimal-Format Solutions

Our experiment in Section 3.1 was designed to test a core hypothesis: that a model, freed from the constraint of a single solution format, would learn a nuanced, adaptive reasoning policy. The results, presented in Table 1, decisively falsify this hypothesis and reveal an alternative organizing principle.

A primary observation from Table 1 is the consistent and significant performance degradation when models are trained with the $R_{\text{composite}}$ on tasks that demonstrably benefit from structured reasoning. On **GSM8K**, a benchmark for multi-step arithmetic, the R_{strict} policy outperforms the $R_{\text{composite}}$ policy by a substantial margin across all models, with performance gaps as large as **7.1%** for Qwen-2.5 7B (92.4 vs. 85.5) and **8.2%** for Gemma-3 12B (94.6 vs. 86.4). Similar trends are observed on **rStar-Coder** and **ReClor**, where the enforced `think-then-answer` structure provides a clear advantage.

This performance gap is not an indictment of the model's capability, but rather a direct consequence of the policy it learns. When trained with the $R_{\text{composite}}$, the optimizer does not learn a dynamic balance of thinking and answering. Instead, in all experimental runs, *the policy invariably converges to the simplest valid format: the direct* `<answer>`*-only response*. This learned behavior, which we term the **Cognitive Shortcut**, involves the complete omission of the rewarded, and often necessary, intermediate reasoning steps. The model, in optimizing $R_{\text{composite}}$, discovers that the path of least resistance is to forgo the complex, high-utility `think` block, even when doing so is detrimental to final performance.

The central, unifying finding from this initial experiment is the discovery of a powerful and universal bias. The optimizer does not act as a neutral maximizer of expected reward across all valid solution formats. It is a highly biased agent that, when presented with a choice, will aggressively converge on the simplest possible specification of a rewarded behavior. This discovery of the Cognitive Shortcut is not the conclusion of our work, but the foundational anomaly that motivates a deeper, more controlled investigation into the nature and strength of this preference.

4 Formalizing the Principle of Least Resistance

4.1 Law of Sequential Optimization

The discovery of the Cognitive Shortcut presented a foundational question: is this preference for simplicity a binary, all-or-nothing phenomenon, or is it governed by a more nuanced, predictable structure? If the optimizer is indeed following a "path of least resistance," this implies a landscape with varying levels of difficulty. This led us to our central hypothesis for this section: the optimizer does not treat a composite objective holistically, but instead decomposes it, prioritizing and conquering its components *in a strict, ascending order of difficulty*.

To test this hypothesis, we designed a controlled experiment to isolate the variable of "difficulty" from all other incentives. We constructed a reward function, $R_{\text{hierarchy}}$, composed of three nested, matryoshka-style format requirements, r_1, r_2, r_3, engineered to represent a clear gradient of increasing complexity.

1. **Easy Format (r_1):** The core requirement—merely enclosing the final numerical answer in a `\boxed{.*}`.
2. **Medium Format (r_2):** A superset of r_1, requiring the model to wrap its entire response in <answer> tags, which must also contain a boxed final answer.
3. **Hard Format (r_3):** The most encompassing format, a superset of r_2, mandating the full <think><answer> structure, which must also satisfy the requirements of r_2 and r_1.

This nested structure, $r_1 \subset r_2 \subset r_3$, is a crucial feature of the experimental design. A generation y that correctly satisfies the hard format r_3 also, by definition, satisfies r_2 and r_1. A perfectly rational, holistic optimizer should be powerfully drawn to learning r_3, as it represents the single solution that simultaneously unlocks all available rewards.

To further isolate the effect of complexity, we set the reward magnitude for satisfying any of these formats to be identical. Let $r(y, r_i)$ be the reward for a generation y satisfying format r_i. We set the reward landscape to be perfectly flat:

$$r(y, r_1) = r(y, r_2) = r(y, r_3) = R_{\text{max}}$$

Incentive-Free Emergence of Reward Ordering

All PO algorithms in our experiments receive a scalar final reward computed as the weighted sum of the individual reward functions. Equal weights are assigned to all reward functions to avoid biasing the optimizer toward any particular one. **Despite the absence of explicit incentives** to favor a specific format, and the structural incentive to prefer the all-encompassing r_3, **the optimizer still optimizes from easiest to hardest rewards**, ultimately converging to thew**minimally compliant response structure** .

The results of this experiment, replicated across multiple models and policy optimization algorithms, are presented in **Figure 2**. The plots provide a stunning and unequivocal visualization of our hypothesis. They do not show a rational convergence on the unified r_3 solution. Instead, they reveal a distinct, **sequential optimization cascade, learned from the inside out.**

As seen consistently across all six panels, the learning process unfolds in clear, predictable stages. The mean reward for the simplest, innermost format, r_1, is the first to rise, typically saturating near its maximum value within the first 200-300 training steps. Only after the policy has reliably mastered this core task does the optimizer begin to make significant progress on the more complex r_2 format. The reward curve for r_2 begins its sharp ascent only after the r_1 curve has started to plateau. Finally, the most complex, all-encompassing format, r_3, is tackled last, with its reward curve beginning to climb only after r_2 is well on its way to convergence.

This staged optimization provides definitive evidence for our "Principle of Least Resistance." Even when presented with a unified solution that satisfies all objectives, the optimizer does not see it. It behaves like a myopic agent minimizing its immediate effort, tackling the lowest-hanging fruit first before moving to more challenging objectives. This ordered law of motion for policy optimization motivates a stress test to quantify the very strength of this resistance.

4.2 QUANTIFYING THE RESISTANCE

The discovery of a sequential learning hierarchy, even under flat rewards, suggests that the "Principle of Least Resistance" is a powerful intrinsic bias. This motivates a critical, adversarial question:

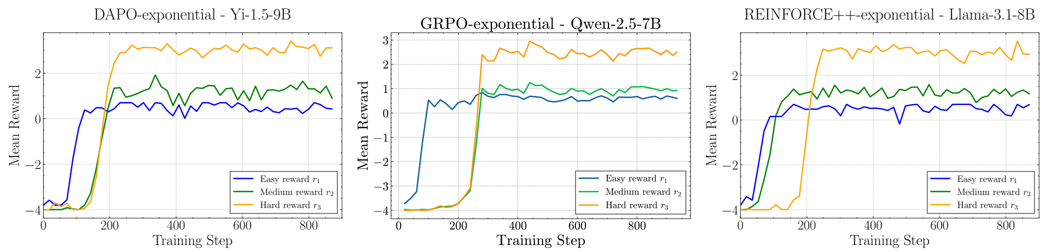

Figure 3: Learning dynamics under an exponentially weighted reward scheme. Despite the hard format r_3 offering 4x the reward of the easy format r_1, the optimizer's learning trajectory remains stubbornly sequential. The massive reward for r_3 is initially ignored in favor of the more learnable, lower-value rewards.

Figure 4: Learning dynamics under an exponentially weighted reward scheme. Despite the hard format r_3 offering 4x the reward of the easy format r_1, the optimizer's learning trajectory remains stubbornly sequential. The massive reward for r_3 is initially ignored in favor of the more learnable, lower-value rewards.

can this innate preference be overridden by extrinsic incentives? To quantify the strength of this resistance, we designed an experiment we term the **"Exponential Gambit,"** aimed at creating a strong, explicit conflict between reward magnitude and format complexity.

We modified our nested reward structure to create a steep gradient of financial incentive, heavily favoring the most complex format. We created a reward landscape where the hard format (r_3) was **2x** more valuable than the medium format (r_2) and **4x** more valuable than the easy format (r_1). A rational, reward-maximizing agent, even a myopic one, should be powerfully drawn to the enormous incentive offered by $r(y, r_3)$. The purpose of this design was to see if a sufficiently large "bribe" could disrupt the natural, sequential learning order we observed previously.

The results, shown in Figure 3, highlight the optimizer's innate bias. The plots reveal that the fundamental learning dynamic is remarkably resistant to this steep incentive gradient. The optimizer, faced with a choice between a small, easily attainable reward and a massive, but more complex one, still prioritizes learnability over immediate financial gain.

The plots reveal that the fundamental learning dynamic is remarkably resistant to this steep incentive gradient, though not entirely immune. Observe the learning curves across all three panels. The reward for the easy format, r_1, is once again the first to be mastered, quickly rising from its initial state and saturating early in training. However, the massive reward for r_3 introduces a fascinating new dynamic. The optimizer does not simply learn the medium reward r_2 next. Instead, the learning curves for r_2 and r_3 rise almost in perfect lockstep. The powerful gradient from the massive r_3 reward appears to "pull" the learning of the structurally similar r_2 format along with it. For the first $\sim$150–200 steps, the model makes little progress on either of these complex formats, focusing solely on the easily attainable r_1. Then, once a certain threshold of basic competence is achieved, the optimizer begins its dramatic ascent, simultaneously conquering both the medium and hard objectives.

4.3 POLICY OPTIMIZATION UNDER CONFLICTING REWARDS

Our investigation has thus far revealed a powerful, sequential bias towards simplicity, even when structural and financial incentives push against it. This motivates one final, maximally clean exper-

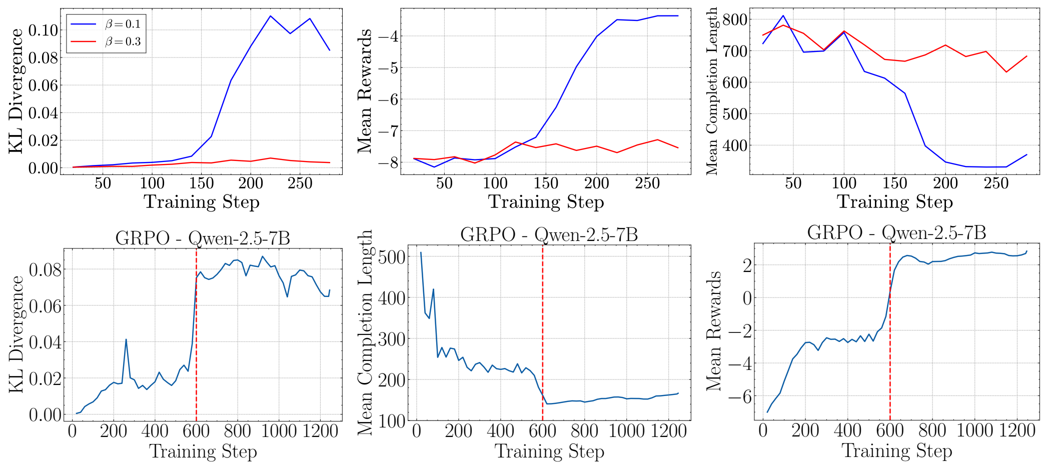

Figure 5: The causal effect of the KL penalty on the emergence of the Cognitive Shortcut. **Top Row:** A direct comparison between a policy with a loose KL leash ($\beta = 0.1$) and a tight leash ($\beta = 0.3$). The freedom to diverge (high KL) is a prerequisite for discovering the high-reward, minimal-length shortcut. **Bottom Row:** The dynamics of a single run, showing a phase transition around the 600-step mark where a spontaneous increase in KL divergence enables the discovery and exploitation of the shortcut.

iment to isolate this preference. The nested structure of our previous rewards, while elegant, leaves open a possibility: could the sequential learning be an artifact of the optimizer learning a shared "core" skill (r_1) before building upon it? To eliminate this possibility and test the principle in its purest form, we designed a final adversarial test: **the mutually exclusive choice.**

The experimental design is a direct extension of our "Exponential Gambit," but with a critical modification. We kept the same steeply incentivized reward magnitudes, but made the format requirements disjoint and non-overlapping. A generation y could now satisfy *only one* format condition.

Let $r(y, f_i)$ be the reward for a generation y satisfying format f_i. The reward function is now defined as:

$$r(y, f_1) = R_{\text{base}} = 1, \quad r(y, f_2) = 2\, r(y, f_1), \quad r(y, f_3) = 2\, r(y, f_2). \tag{5}$$

This creates a stark choice landscape: the optimizer can pursue a small but simple reward (r_1), a medium reward (r_2), or a 4x reward (r_3) over r_1, but it can only choose one path. There are no shared sub-problems. This setup forces the optimizer to reveal its true preference when faced with a simple cost-benefit analysis.

The results, presented in Figure 4, are the most decisive evidence yet for the Principle of Least Resistance.

The plots reveal a stark and absolute convergence. Unlike the sequential learning we saw in the nested case, here the optimizer does not eventually learn the harder formats. It makes a decision early in training and commits to it absolutely. In every single run, across all models and algorithms, the policy **exclusively converges to the easiest format,** r_1.

The reward curves for the medium (r_2) and hard (r_3) formats remain flat at their initial negative values for the entire duration of training. The massive potential reward offered by r_3 is never explored. The optimizer identifies the simplest path to a positive reward and dedicates all of its capacity to mastering it, completely ignoring the other, more lucrative options.

This final experiment provides an irrefutable conclusion. The preference for the path of least resistance is not a heuristic or an artifact of a specific reward structure. It is a fundamental, powerful, and seemingly absolute bias in the policy optimization process. The optimizer does not perform a global cost-benefit analysis; it greedily follows the most immediately learnable gradient. This solidified understanding of the optimizer's innate behavior now allows us to turn our attention to the final act of our investigation: uncovering the theoretical origins of this powerful force.

5 THE PRICE OF EXPLORATION

Our investigation has established the "Principle of Least Resistance" as a powerful, predictive law. The final act is to uncover its origin, and in doing so, reveal a fundamental tension at the heart of modern policy optimization. A prominent trend in recent state-of-the-art algorithms, such as DAPO and VAPO, is the removal of the KL divergence penalty, arguing that it is an unnecessary constraint on the model's ability to maximize reward. Our final analysis reveals that while this freedom is essential for learning, it comes at a cost: it unleashes the optimizer's powerful, innate bias to find and exploit the simplest specification of the reward function, a behavior that is a classic form of **reward hacking**.

To dissect this relationship, we treat the KL divergence not as a loss to be minimized, but as a scientific instrument measuring the policy's deviation from its initial reference state, π_{ref}. This deviation represents the policy's **exploratory freedom**. We hypothesize that the Cognitive Shortcut is a form of reward hacking that requires a significant amount of this freedom to discover. The KL penalty, controlled by its coefficient β, therefore acts as a **"leash,"** directly modulating the policy's ability to find and exploit such shortcuts.

We designed a causal experiment to test this. We conducted two training runs under our composite reward, identical in all aspects except for the strength of this leash: a "Tight Leash" run with a high KL penalty ($\beta = 0.3$) and a "Loose Leash" run with a low penalty ($\beta = 0.1$), mimicking the unconstrained exploration of modern algorithms. The results, in Figure 5 (top-row), provide a stark illustration of this trade-off. The top-left panel shows the direct effect of our intervention. The policy with the loose leash (blue), analogous to a KL-free objective, is free to explore and achieves a high final KL divergence. The policy with the tight leash (red) is constrained. The consequences are shown in the adjacent panels. The unleashed policy successfully discovers the high-reward solution (top-middle) by converging on the efficient, minimal-length Cognitive Shortcut (top-right). The leashed policy, forbidden from making the large policy shift required to "find the hack," remains trapped in a lower-reward state.

This dynamic is not merely an average-case phenomenon; it can be observed live within a single training run, as shown in the Figure 5 (bottom row). These plots capture the moment the reward hack is discovered. At the 600-step mark (dashed red line), the optimizer identifies the powerful gradient of the shortcut. To exploit it, the policy undergoes a rapid phase transition, marked by a sharp increase in KL divergence. This decisive shift away from its prior is immediately followed by a surge in reward and a steep drop in completion length. This is the "eureka moment" of specification gaming.

Our final analysis provides the definitive explanation for the Principle of Least Resistance and its connection to reward hacking. The Cognitive Shortcut is the result of an optimizer that is not just maximizing reward, but is actively searching for the most learnable gradient in the reward landscape. **The freedom to diverge from the initial policy, while necessary for performance, is the very mechanism that enables the model to discover and exploit these unintended, simplistic solutions.** This reveals a critical and uncomfortable trade-off for the field: the path to more capable models may be inseparable from the path to more sophisticated forms of reward hacking. The challenge, therefore, is not simply to unleash our models, but to design reward landscapes that are fundamentally resistant to being gamed.

6 CONCLUSION

Our investigation revealed that the "Cognitive Shortcut" is not a mere anomaly, but the predictable outcome of a powerful principle: **Policy Optimization Prefers the Path of Least Resistance**. We showed that this preference is a formidable, quantifiable force, capable of overriding even significant financial incentives in favor of the most easily learnable solution. We have shown a profound paradox: the freedom to explore, essential for discovering high-reward policies, is the very mechanism that enables the optimizer to find and aggressively exploit the simplest specification of the reward function, a classic and potent form of reward hacking. The central challenge for alignment, therefore, is not simply to unleash our models, but to architect reward landscapes that are fundamentally resistant to being gamed, ensuring that the path we desire is also the path the optimizer is compelled to take.

REFERENCES

01. AI, :, Alex Young, Bei Chen, Chao Li, Chengen Huang, Ge Zhang, Guanwei Zhang, Guoyin Wang, Heng Li, Jiangcheng Zhu, Jianqun Chen, Jing Chang, Kaidong Yu, Peng Liu, Qiang Liu, Shawn Yue, Senbin Yang, Shiming Yang, Wen Xie, Wenhao Huang, Xiaohui Hu, Xiaoyi Ren, Xinyao Niu, Pengcheng Nie, Yanpeng Li, Yuchi Xu, Yudong Liu, Yue Wang, Yuxuan Cai, Zhenyu Gu, Zhiyuan Liu, and Zonghong Dai. Yi: Open foundation models by 01.ai, 2025. URL https://arxiv.org/abs/2403.04652.

Guanzheng Chen, Xin Li, Michael Qizhe Shieh, and Lidong Bing. Longpo: Long context self-evolution of large language models through short-to-long preference optimization, 2025. URL https://arxiv.org/abs/2502.13922.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems, 2021. URL https://arxiv.org/abs/2110.14168.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jiawei Wang, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Shengfeng Ye, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wanjia Zhao, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanhong Xu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning, 2025. URL https://arxiv.org/abs/2501.12948.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan

Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. The llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the math dataset. *arXiv preprint arXiv:2103.03874*, 2021.

Jian Hu, Jason Klein Liu, Haotian Xu, and Wei Shen. Reinforce++: An efficient rlhf algorithm with robustness to both prompt and reward models, 2025. URL https://arxiv.org/abs/2501.03262.

Hugging Face. Open r1: A fully open reproduction of deepseek-r1, January 2025. URL https://github.com/huggingface/open-r1.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7b, 2023. URL https://arxiv.org/abs/2310.06825.

Mingjie Liu, Shizhe Diao, Ximing Lu, Jian Hu, Xin Dong, Yejin Choi, Jan Kautz, and Yi Dong. Prorl: Prolonged reinforcement learning expands reasoning boundaries in large language models, 2025a. URL https://arxiv.org/abs/2505.24864.

Yifei Liu, Li Lyna Zhang, Yi Zhu, Bingcheng Dong, Xudong Zhou, Ning Shang, Fan Yang, and Mao Yang. rstar-coder: Scaling competitive code reasoning with a large-scale verified dataset, 2025b. URL https://arxiv.org/abs/2505.21297.

Zichen Liu, Changyu Chen, Wenjun Li, Penghui Qi, Tianyu Pang, Chao Du, Wee Sun Lee, and Min Lin. Understanding r1-zero-like training: A critical perspective, 2025c. URL https://arxiv.org/abs/2503.20783.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report, 2025. URL https://arxiv.org/abs/2412.15115.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. Deepseekmath: Pushing the limits of mathematical reasoning in open language models, 2024. URL https://arxiv.org/abs/2402.03300.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, Gaël Liu, Francesco Visin, Kathleen Kenealy, Lucas Beyer, Xiaohai Zhai, Anton Tsitsulin, Robert Busa-Fekete, Alex Feng, Noveen Sachdeva, Benjamin Coleman, Yi Gao, Basil Mustafa, Iain Barr, Emilio Parisotto, David Tian, Matan Eyal, Colin Cherry, Jan-Thorsten Peter, Danila Sinopalnikov, Surya Bhupatiraju, Rishabh Agarwal, Mehran Kazemi, Dan Malkin, Ravin Kumar, David Vilar, Idan Brusilovsky, Jiaming Luo, Andreas Steiner, Abe Friesen, Abhanshu Sharma, Abheesht Sharma, Adi Mayrav Gilady, Adrian Goedeckemeyer, Alaa Saade, Alex Feng, Alexander Kolesnikov, Alexei Bendebury, Alvin Abdagic, Amit Vadi, András György, André Susano Pinto, Anil Das, Ankur Bapna, Antoine Miech, Antoine Yang, Antonia Paterson, Ashish Shenoy, Ayan Chakrabarti, Bilal Piot, Bo Wu, Bobak Shahriari, Bryce Petrini, Charlie Chen, Charline Le Lan, Christopher A. Choquette-Choo, CJ Carey, Cormac Brick, Daniel Deutsch, Danielle Eisenbud, Dee Cattle, Derek Cheng, Dimitris Paparas, Divyashree Shivakumar Sreepathihalli, Doug Reid, Dustin Tran, Dustin Zelle, Eric Noland, Erwin Huizenga, Eugene Kharitonov, Frederick Liu, Gagik Amirkhanyan, Glenn Cameron, Hadi Hashemi, Hanna Klimczak-Plucińska, Harman Singh, Harsh Mehta, Harshal Tushar Lehri, Hussein Hazimeh, Ian Ballantyne, Idan Szpektor, Ivan Nardini, Jean Pouget-Abadie, Jetha Chan, Joe Stanton, John Wieting, Jonathan Lai, Jordi Orbay, Joseph Fernandez, Josh Newlan, Ju yeong Ji, Jyotinder Singh, Kat Black, Kathy Yu, Kevin Hui, Kiran Vodrahalli, Klaus Greff, Linhai Qiu, Marcella Valentine, Marina Coelho, Marvin Ritter, Matt Hoffman, Matthew Watson, Mayank Chaturvedi, Michael Moynihan, Min Ma, Nabila Babar, Natasha Noy, Nathan Byrd, Nick Roy, Nikola Momchev, Nilay Chauhan, Noveen Sachdeva, Oskar Bunyan, Pankil Botarda, Paul Caron, Paul Kishan Rubenstein, Phil Culliton, Philipp Schmid, Pier Giuseppe Sessa, Pingmei Xu, Piotr Stanczyk, Pouya Tafti, Rakesh Shivanna, Renjie Wu, Renke Pan, Reza Rokni, Rob Willoughby, Rohith Vallu, Ryan Mullins, Sammy Jerome, Sara Smoot, Sertan Girgin, Shariq Iqbal, Shashir Reddy, Shruti Sheth, Siim Põder, Sijal Bhatnagar, Sindhu Raghuram Panyam, Sivan Eiger, Susan Zhang, Tianqi Liu, Trevor Yacovone, Tyler Liechty, Uday Kalra, Utku Evci, Vedant Misra, Vincent Roseberry, Vlad Feinberg, Vlad Kolesnikov, Woohyun Han, Woosuk Kwon, Xi Chen, Yinlam Chow, Yuvein Zhu, Zichuan Wei, Zoltan Egyed, Victor Cotruta, Minh Giang, Phoebe Kirk, Anand Rao, Kat Black, Nabila Babar, Jessica Lo, Erica Moreira, Luiz Gustavo Martins, Omar Sanseviero, Lucas Gonzalez, Zach Gleicher, Tris Warkentin, Vahab Mirrokni, Evan Senter, Eli Collins, Joelle Barral, Zoubin Ghahramani, Raia Hadsell, Yossi Matias, D. Sculley, Slav Petrov, Noah Fiedel, Noam Shazeer, Oriol Vinyals, Jeff Dean, Demis Hassabis, Koray Kavukcuoglu, Clement Farabet, Elena Buchatskaya, Jean-Baptiste Alayrac, Rohan Anil, Dmitry, Lepikhin, Sebastian Borgeaud, Olivier Bachem, Armand Joulin, Alek Andreev, Cassidy Hardin, Robert Dadashi, and Léonard Hussenot. Gemma 3 technical report, 2025. URL https://arxiv.org/abs/2503.19786.

Chulin Xie, Yangsibo Huang, Chiyuan Zhang, Da Yu, Xinyun Chen, Bill Yuchen Lin, Bo Li, Badih Ghazi, and Ravi Kumar. On memorization of large language models in logical reasoning. 2024. URL https://arxiv.org/abs/2410.23123.

Roy Xie, David Qiu, Deepak Gopinath, Dong Lin, Yanchao Sun, Chong Wang, Saloni Potdar, and Bhuwan Dhingra. Interleaved reasoning for large language models via reinforcement learning, 2025. URL https://arxiv.org/abs/2505.19640.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, Xin Liu, Haibin Lin, Zhiqi Lin, Bole Ma, Guangming Sheng, Yuxuan Tong, Chi Zhang, Mofan Zhang, Wang Zhang, Hang Zhu, Jinhua Zhu, Jiaze Chen, Jiangjie Chen, Chengyi Wang, Hongli Yu, Yuxuan Song, Xiangpeng Wei, Hao Zhou, Jingjing Liu, Wei-Ying Ma, Ya-Qin Zhang, Lin Yan, Mu Qiao, Yonghui Wu, and Mingxuan Wang. Dapo: An open-source llm reinforcement learning system at scale, 2025. URL https://arxiv.org/abs/2503.14476.

Weihao Yu, Zihang Jiang, Yanfei Dong, and Jiashi Feng. Reclor: A reading comprehension dataset requiring logical reasoning, 2020. URL https://arxiv.org/abs/2002.04326.

Yu Yue, Yufeng Yuan, Qiying Yu, Xiaochen Zuo, Ruofei Zhu, Wenyuan Xu, Jiaze Chen, Chengyi Wang, TianTian Fan, Zhengyin Du, Xiangpeng Wei, Xiangyu Yu, Gaohong Liu, Juncai Liu, Lingjun Liu, Haibin Lin, Zhiqi Lin, Bole Ma, Chi Zhang, Mofan Zhang, Wang Zhang, Hang Zhu, Ru Zhang, Xin Liu, Mingxuan Wang, Yonghui Wu, and Lin Yan. Vapo: Efficient and reliable reinforcement learning for advanced reasoning tasks, 2025. URL https://arxiv.org/abs/2504.05118.

A Appendix

User

A ship traveling along a river has covered $24\ \mathrm{km}$ upstream and $28\ \mathrm{km}$ downstream. For this journey, it took half an hour less than for traveling $30\ \mathrm{km}$ upstream and $21\ \mathrm{km}$ downstream, or half an hour more than for traveling $15\ \mathrm{km}$ upstream and $42\ \mathrm{km}$ downstream, assuming that both the ship and the river move uniformly.
Determine the speed of the ship in still water and the speed of the river.

Qwen-7B + OpenR1-Math220K

r_{strict}

<think>Let's think step by step...</think> <answer>... $\boxed{4, 10}$</answer>

$r_{\text{composite}}$

<answer>Let's denote the speed of the ship in still water as s and the speed of the river as $\ldots$ \$10 km/h and the speed of the river is 2 km/h. $\boxed{10, 2}$</answer>

Figure 6

User

Let AB be a chord of the unit circle $\odot O$. If the area of $\odot O$ is exactly equal to the area of the square with side AB, then $\angle AOB =$ (to 0.001 degree).

Qwen-7B + OpenR1-Math220K

r_{strict}

<think>Okay, so the problem is about a chord AB on the unit circle ... So, setting these equal ... </think>
<answer>124.806°</answer>

$r_{\text{composite}}$

<answer>First, we need to find the area of the unit circle ... Therefore, the angle AOB is approximately 65.427 degrees, rounded to 0.001 degree. $\boxed{65.427}$</answer>

Figure 7