Cornerstones or Stumbling Blocks? Deciphering the Rock Tokens in On-Policy Distillation

Yuxuan Jiang[1,*], **Runchao Li**[2,*], **Shubhashis Roy Dipta**[1,*], **Dawei Li**[3], **Zhao Yang**[4,†]
[1]UMBC [2]Case Western Reserve University [3]Arizona State University [4]VU Amsterdam
[*]Equal contribution [†]Corresponding author
yuxuanj1@umbc.edu

Abstract

While recent work in Reinforcement Learning with Verifiable Rewards (RLVR) has shown that a small subset of critical tokens disproportionately drives reasoning gains, an analogous token-level understanding of On-Policy Distillation (OPD) remains largely unexplored. In this work, we investigate high-loss tokens, a token type that—as the most direct signal of student-teacher mismatch under OPD's per-token KL objective—should progressively diminish as training converges according to existing studies; however, our empirical analysis shows otherwise. Even after OPD training reaches apparent saturation, a substantial subset of tokens continues to exhibit persistently high loss; these tokens, which we term Rock Tokens, can account for up to 50% of the tokens in generated outputs. Our investigation reveals two startling paradoxes. First, despite their high occurrence frequency providing a disproportionately large share of total gradient norms, Rock Tokens themselves remain stagnant throughout training, resisting teacher-driven corrections. Second, through causal intervention, we find that these tokens provide negligible functional contribution to the model's actual reasoning performance. These findings suggest that a vast amount of optimization bandwidth is spent on structural and discourse residuals that the student model cannot or need not internalize. By deconstructing these dynamics, we demonstrate that strategically bypassing these "stumbling blocks" can significantly streamline the alignment process, challenging the necessity of uniform token weighting and offering a more efficient paradigm for large-scale model distillation. Our code is available at https://github.com/YuxuanJiang1/Rock-Token.

1 Introduction

On-Policy Distillation (OPD) has established itself as a cornerstone of the modern post-training pipeline for Large Language Models (LLMs). By aligning a student model with a superior teacher through trajectories sampled from its own policy, OPD enables a token level of reasoning refinement that goes beyond static Supervised Fine-Tuning (SFT). This effectiveness has been validated by industrial works such as DeepSeek-V4 [2], MiMo [20], and Qwen-3 [24], where OPD serves as a vital component alongside SFT or Reinforcement Learning with Verifiable Rewards (RLVR) to further squeeze out reasoning performance.

Recent studies in Reinforcement Learning with Verifiable Rewards (RLVR) have revealed that not all tokens contribute equally to model learning: a small subset of critical tokens, such as high-entropy "forking tokens" at decisive branching points, disproportionately drives reasoning gains [21]. While this token-level perspective has reshaped our understanding of RLVR dynamics, an analogous investigation in OPD remains largely unexplored, despite OPD's reliance on dense token-level supervision. In this work, we initiate such an investigation by focusing on a token type uniquely

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

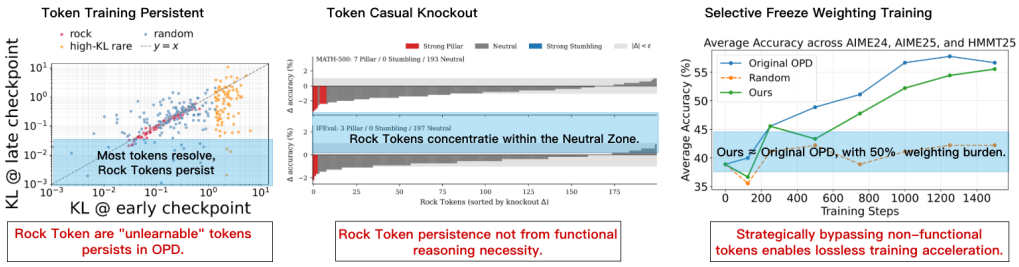


Figure 1: **The lifecycle and functional impact of Rock Tokens in OPD.** (a) **Phenomenon**: Identification of optimization-resistant tokens. (b) **Mechanism**: Causal evidence of structural redundancy via token knock-out. (c) **Utility**: Performance parity achieved through strategic gradient sparsification.

meaningful to the OPD paradigm: high-loss tokens. Unlike RLVR, where token importance is inferred from policy entropy or reward signals, OPD provides a direct measure of student-teacher mismatch through per-token KL divergence; high-loss tokens thus explicitly mark the positions where the student's policy diverges most sharply from the teacher's, making them natural candidates for the "critical correction points" that drive alignment [17].

Intuitively, as the student aligns with the teacher during the OPD process, both the magnitude and cardinality of these high-loss tokens are expected to diminish, eventually leading to an optimization plateau. However, our empirical investigations reveal a striking phenomenon: even after training has reached apparent saturation, a stubborn population of tokens—which we formally define as Rock Tokens—continues to exhibit consistently high loss. Notably, while these tokens represent a small minority of the vocabulary (approximately 6%), they account for as much as 50% of the total token frequency in the model's output. This persistence challenges conventional understanding of model convergence and raises a fundamental question: Are Rock Tokens indispensable Pillars for policy alignment, or are they Stumbling Blocks that create unnecessary optimization redundancy?

To provide a systematic understanding of the Rock Token phenomenon, our research is structured across three investigative dimensions: identifying **what** are rock tokens, analyzing **why** they emerge and persist, and uncovering **how** they collectively impact the overall OPD training process. This comprehensive framework is operationalized through the following three stages:

First, we characterize the identity of Rock Tokens (What). We begin by identifying Rock Tokens linguistic and statistical properties, revealing that they primarily comprise **syntactic and structural scaffolding**—such as formatting delimiters, whitespace symbols, and high-frequency discourse markers (e.g., "So", "Wait"). To determine the dynamic role of these elements, we employ quantitative gradient dynamics analysis to address **RQ1: Do Rock Tokens still provide useful learning signals during optimization?** By examining whether their persistent high loss effectively translates into constructive gradients, we assess whether these tokens offer steady directional guidance or merely represent stagnant optimization noise throughout the OPD process.

Second, we investigate the underlying mechanisms behind their persistence (Why). Beyond the actual contribution of Rock Tokens during training is clarified, a more profound question emerges: **why do these tokens persist in the first place?** We hypothesize that the student model develops a strong decoding dependency on these specific tokens to maintain its reasoning flow, creating a path dependency that resists corrections from the teacher model. This leads to our **RQ2: Does Rock Token persistence stem from student path dependency?** To investigate this, we conduct *token knock-out* experiments as a form of ablation analysis, examining whether the student's logical coherence is so deeply intertwined with these persistent residuals that it becomes structurally unable to shift its policy toward the teacher's distribution.

Finally, we evaluate how Rock Tokens influence the overall OPD training process (How). Finally, we arrive at a provocative question: what would happen if these tokens were excluded from the very beginning of OPD training? This leads to our final investigation, **RQ3: What is the genuine functional contribution of Rock Tokens to model training?** By freezing the gradients of these tokens from the onset of training, we intuitively isolate their functional necessity throughout the entire distillation process. This causal intervention allows us to rigorously determine whether Rock

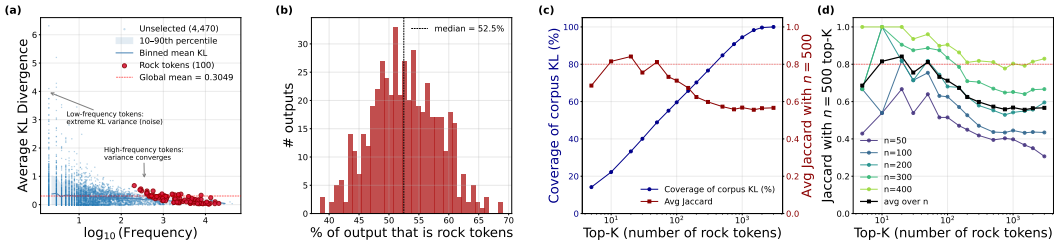


Figure 2: **Empirical identification and stability of Rock Tokens. (a)** Per-token KL $\widehat{\ell}_v$ vs. frequency on $N{=}500$ MATH-500 trajectories: rare tokens are noise-dominated, while the Rock Score $R(v)$ isolates true Rock Tokens (red) at the upper edge of stable frequency bands. **(b)** Per-sequence Rock-Token density (median 52.5%). **(c)** Cumulative loss coverage (blue) and selection stability (red) vs. cutoff K; $K{=}100$ balances reproducibility with $\approx 60\%$ coverage of the distillation burden. **(d)** Jaccard ranking robustness across sample sizes $n \in [50, 400]$: the top-100 selection is stable, larger cutoffs degrade.

Tokens serve as indispensable pillars of the model's reasoning integrity or if they represent a strategic opportunity to bypass computational redundancy and optimize the efficiency of the alignment pipeline.

In summary, our key contributions are as follows:

- For the first time, we identify and characterize **Rock Tokens**, a persistent high-loss phenomenon that unveils a previously hidden aspect of optimization dynamics in On-Policy Distillation.
- We deconstruct the intricate dynamics of the OPD process by systematically uncovering the functional roles of Rock Tokens across both the training and inference phases.
- Our research reveals a critical inefficiency in current alignment pipelines, demonstrating that a substantial portion of training signals in OPD provides significantly lower-than-expected contributions to the model's ultimate reasoning performance.

2 Rock Tokens

2.1 Token-Level Loss in On-Policy Distillation

On-policy distillation (OPD) optimizes a student policy π_θ by matching a teacher policy π_T on trajectories sampled from the student's own distribution. The objective minimizes the per-token reverse KL divergence:

$$\mathcal{L}_{\mathrm{OPD}}(\theta) = \mathbb{E}_{x_{1:T}\sim\pi_\theta}\left[\sum_{t=1}^{T} D_{\mathrm{KL}}\left(\pi_\theta(\cdot|x_{<t})\|\pi_T(\cdot|x_{<t})\right)\right]. \tag{1}$$

The per-position loss $\ell_t = D_{\mathrm{KL}}(\pi_\theta(\cdot \mid x_{<t}) \,\|\, \pi_T(\cdot \mid x_{<t}))$ is the exact reverse KL divergence at position t, fully determined by the context $x_{<t}$ [4]. Ideally, as the student aligns with the teacher, the magnitude of ℓ_t should systematically diminish across the vocabulary, leading to an optimization plateau where corrective signals eventually vanish.

2.2 Defining and Detecting Rock Tokens

Contrary to the expected convergence, we observe a subset of tokens that persistently exhibit high KL divergence even after training saturation. We term these **Rock Tokens**.

At an intuitive level, our detection method isolates these tokens based on their *Total Distillation Burden*—the absolute volume of optimization capacity consumed by a specific vocabulary item across the entire training corpus. Since the loss $\mathcal{L}_{\mathrm{OPD}}$ is an expectation over student trajectories, it admits an exact additive decomposition over token types v in the vocabulary $\mathcal{V}$:

$$\mathcal{L}_{\mathrm{OPD}}(\theta) = \mathbb{E}_{x\sim\pi_\theta}\left[\sum_t \ell_t\right] = \sum_{v\in\mathcal{V}} \underbrace{\mathrm{Freq}(v)\cdot\mathbb{E}[\ell_t \mid x_t = v]}_{R(v)} \tag{2}$$

We formally define the **Rock Score** as $R(v) = \bar{\ell}_v \cdot \mathrm{Freq}(v)$, where $\bar{\ell}_v$ is the empirically estimated mean KL loss of token v at the final checkpoint, and $\mathrm{Freq}(v)$ is the raw empirical occurrence count of token type v in the collected rollout corpus. Grouping positions by their realized token and averaging the per-position KL yields the exact identity $\sum_v R(v) = \widehat{\mathcal{L}}_{\mathrm{OPD}}$ for the corpus-level loss; $R(v)$ is therefore the exact share of the corpus loss attributable to token type v, not an approximation. Ranking tokens by $R(v)$ is the unique additive ranking that strictly agrees with the empirical, corpus-level OPD objective realized by these rollouts.

The $\mathrm{Freq}(v)$ factor acts as a natural regularizer: it up-weights tokens whose mean-loss estimate converges quickly under heavy sampling, while heavily discounting rare tokens whose mean is dominated by small-sample variance. Detection via $R(v)$ thus simultaneously aligns with the empirical, corpus-level OPD objective realized by these rollouts while remaining robust to the empirical estimation noise inherent in natural language generation.

Operationally, $R(v)$ prioritizes tokens that exhibit both a substantial occurrence count and an above-baseline per-token loss. Rare tokens with extreme observed losses receive appropriately low scores because their $\mathrm{Freq}(v)$ weight is minimal. As demonstrated empirically in Figure 2, the top-ranked tokens cluster in the mid-to-high frequency range and sit at the upper edge of the loss distribution within their respective frequency bands, successfully filtering out the rare-but-extreme statistical outliers.

2.3 Empirical Analysis

Rock tokens exhibit persistently high expected KL divergence between student π_θ and teacher π_T after on-policy distillation. We estimate this per-token loss $\widehat{\ell}_v = \widehat{\mathbb{E}}\big[D_{\mathrm{KL}}\big(\pi_\theta(\cdot \mid x_{<t}) \,\|\, \pi_T(\cdot \mid x_{<t})\big) \mid x_t = v\big]$ using student rollouts.

Setup The student (Qwen3-4B-Instruct) is on-policy distilled from the teacher (Qwen3-30B-A3B-Instruct) using OpenThoughts. Per-token KL is evaluated on $N{=}500$ MATH-500 problems decoded greedily from the student at the final checkpoint; this attribution corpus is distinct from the on-policy sampling trajectories used during training (full details in Section 5). The resulting top-$K{=}100$ Rock vocabulary is fixed once from this corpus and reused for all subsequent analyses; its stability across decoding configurations is verified in Appendix G.

Observations Figure 2 illustrates the distribution and density of the identified Rock Tokens within the student's generated trajectories. Panel (a) displays the absolute count of Rock Tokens per sequence, while Panel (b) normalizes this count as a percentage of the total sequence length. Strikingly, the data reveals that Rock Tokens are highly concentrated, comprising a median of 52.5% of the total tokens in every generated output.

Take-away **The Pervasiveness of Rock Tokens:** Although Rock Tokens represent a tightly constrained subset of the overall vocabulary (e.g., $K = 100$), their footprint during generation is massive. Over half of the model's reasoning trajectory consists of these stubborn structural and formatting tokens, demonstrating that the primary locus of student-teacher disagreement occurs consistently throughout the core scaffolding of the output.

2.4 Determining the Rock Token Cutoff

Defining the boundary of the Rock Token set $\mathcal{R}$ requires choosing a cutoff K that balances *selection stability* (reproducibility across samples) against *loss coverage* (fraction of optimization loss captured). A small K yields a stable set but low coverage, while a large K captures more loss but degrades under sampling noise.

Empirical sweeps (Figures 1b and 1c) identify an optimal intersection at $K = 100$. This boundary accounts for roughly 60% of the corpus-level KL divergence while maintaining high stability across subset sizes. We therefore adopt $K = 100$ to ensure a mathematically reproducible vocabulary for subsequent analyses. Comprehensive stability metrics and sweeping procedures are detailed in Appendix E.

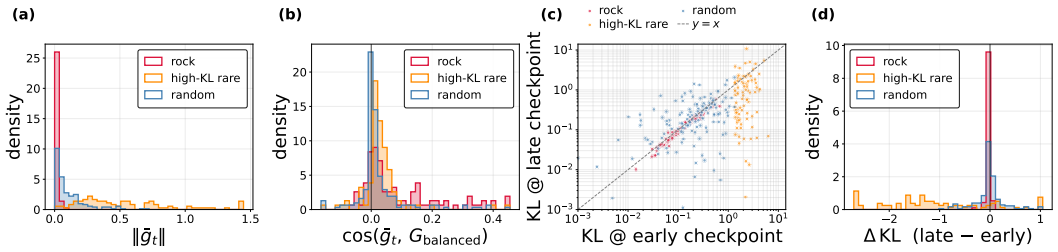


Figure 3: **Per-token gradient geometry and persistence under training.** (a) Per-token logit-gradient magnitude $\|\bar{g}_t\|$ by group: rocks are an order of magnitude smaller than rare high-KL tokens. (b) Cosine alignment with the frequency-balanced descent direction G_{balanced}: rocks are positively aligned, with a tail reaching $\cos > 0.3$. (c) Per-token mean KL paired across two training checkpoints (log-log). Points below the diagonal indicate KL reduction during further training; rocks lie on the diagonal, rare high-KL tokens fall below. (d) Distribution of $\Delta\text{KL} = \text{KL}_{\text{late}} - \text{KL}_{\text{early}}$ by group: rocks concentrate at $\Delta\text{KL} \approx 0$ (persistent); rare high-KL tokens have large negative ΔKL (learned).

2.5 Analysis of Rock Tokens

We decoded the top-K=100 Rock Tokens into their string forms using the Qwen3 tokenizer. Despite being identified by a purely numerical criterion, the resulting set is functionally coherent: Rock Tokens consistently mark structural or discourse boundaries in mathematical reasoning.

As detailed in Appendix B, four functional clusters dominate $\mathcal{R}$: (1) **LaTeX and math delimiters**, (2) **Markdown and whitespace structure**, (3) **Discourse markers** (e.g., "So", "Wait"), and (4) **Digits**. In contrast, frequency-matched controls ($\mathcal{S}_{\text{ctrl}}$) predominantly consist of content-bearing tokens (e.g., nouns, verbs) used at non-boundary positions.

Take-away **Structure Over Content:** The clustering of $\mathcal{R}$ at boundaries indicates student-teacher divergence centers on **how to structure** reasoning rather than **what content** to generate. Rock Tokens are a signature of the on-policy regime, marking where the student systematically resists the teacher's discourse style.

2.6 Empirical Observations on Gradient Geometry and Persistence

Having established that Rock Tokens primarily comprise syntactic and structural scaffolding, we address **RQ1: Do Rock Tokens still provide useful learning signals during optimization?** To determine whether their persistent high loss effectively translates into steady directional guidance or merely stagnant optimization noise, we characterize their training dynamics and geometric properties against rare high-KL tokens and random tokens (illustrated in Figure 3).

To formally evaluate this, we analyze the geometry of per-token gradients. For each token type t with n_t occurrences, we compute the mean per-occurrence reverse-KL gradient in logit space, $\bar{g}_t = \frac{1}{n_t}\sum_i g_i$. We then decompose its contribution to the global descent as:

$$\text{contrib}(t) \;=\; n_t \cdot \|\bar{g}_t\| \cdot \cos\big(\bar{g}_t, G_{\text{balanced}}\big), \tag{3}$$

where $G_{\text{balanced}} = \sum_t \bar{g}_t$ is a frequency-balanced reference that grants every token type equal weight, deliberately removing the trivial frequency-weighted dominance of common tokens. This decomposition yields several key insights corresponding to our empirical observations:

Low Gradient Magnitude vs. High Frequency (Panel a): Rock tokens are characterized by substantially *smaller* per-occurrence gradient magnitudes than rare high-KL tokens (median $\|\bar{g}_t\| \approx 0.016$ versus ≈ 0.54; Mann-Whitney $p < 10^{-30}$). However, the multiplicative decomposition in Eq. 3 shows that rock dominance arises from the *frequency* factor (n_t): because of their massive occurrence rate in the training corpus, this small per-token signal accumulates into a dominant aggregate force on the weight updates.

Alignment with Global Optimization (Panel b): Counter-intuitively, the gradient direction of Rock Tokens ($\bar{g}_t$) exhibits non-trivially positive directional alignment with the balanced global gradient G_{balanced} (median $\cos \approx 0.040$ versus ≈ 0.025 for high-KL and ≈ 0.006 for random tokens), with a

sub-population reaching $\cos > 0.3$. This means that each rock contributes a small but consistently aligned signal at many positions, whereas rare high-KL tokens contribute large but direction-distinct signals at very few. From the teacher's perspective, Rock Tokens do indeed point toward a "correct" optimization path for the overall objective.

Persistence During Training (Panel c & d): To confirm that this directional alignment translates into resistance to further learning, we compare per-token KL between an early and late checkpoint. Rare high-KL tokens experience a substantial reduction in KL (median $2.02 \rightarrow 0.85$, $\sim 58\%$; Wilcoxon $p < 10^{-8}$), moving significantly below the $y = x$ diagonal. In contrast, Rock Tokens remain essentially unchanged (median $0.21 \rightarrow 0.19$), clustering sharply along the diagonal with the distribution of ΔKL concentrated exactly at zero (Panel d). Random control tokens show a symmetric, noisy distribution with no net change.

This optimization stagnation is geometrically consistent with our gradient findings: rare high-KL tokens possess direction-distinct gradients that focused updates can isolate and resolve. Conversely, because rocks contribute highly aligned gradients, their targeted reduction would be indistinguishable from following the global descent. Thus, they receive no separable, token-specific learning signal to overcome their inherent loss.

While our gradient decomposition explains the geometry of this stagnation, the underlying mechanical cause remains open. We hypothesize this paradox is driven by adaptive optimizers under heavy-tailed class imbalance [12]. Unlike in pre-training—where Adam beneficially suppresses frequent tokens to avoid overfitting structural noise—On-Policy Distillation often requires updating these exact high-frequency structural tokens to match a teacher's formatting or style. By inflating the second-moment estimates for ubiquitous Rock Tokens, Adam systematically neutralizes their effective learning rate, trapping the student in pre-existing habits despite possessing accurate directional gradients. We leave the formal verification of this optimizer-suppression hypothesis to future work.

Take-away **The Gradient Paradox:** Rock Tokens *do* provide useful learning signals, generating gradients that are well-aligned with the global optimization target (Panel b). However, they remain stagnant and "untrained" throughout the OPD process (Panel d). This paradox suggests that their persistence is not due to a lack of constructive signal, but rather the student model's internal optimization bias, which stubbornly prioritizes pre-existing structural scaffolding over the teacher's stylistic direction because the rock-token signal is mathematically inseparable from the global descent direction.

3 Pillar Tokens: The Scaffolding of Fragile Reasoning

3.1 Existence as Necessity: The Functional Value of Rock Tokens

We have observed that **Rock Tokens** remain resistant to optimization throughout the entire OPD process. To understand the origin of this resistance, we hypothesize that these tokens serve as the structural "pillars" of the model's reasoning process. Under this view, the student model may actively protect these tokens from optimization to prevent its internal reasoning framework from collapsing, even when they conflict with the teacher's distribution. This leads to our **RQ2: Does Rock Token persistence stem from student path dependency?**

3.2 Causal Probing via Inference-time Knockout

To move beyond distributional metrics and pinpoint the exact functional value of individual tokens, we employ an inference-time **knockout** procedure. We probe the causal contribution of a token v by observing how its absence affects the model's accuracy on a benchmark $\mathcal{B}$. Specifically, we form a *knockout policy* $\pi_\theta^{\setminus v}$ by forcing the logit of token v to $-\infty$ at every decoding step. We then measure the **causal delta**:

$$\Delta_{\mathcal{B}}(v) = \mathrm{Acc}_{\mathcal{B}}\big(\pi_\theta^{\setminus v}\big) - \mathrm{Acc}_{\mathcal{B}}\big(\pi_\theta\big). \tag{4}$$

A token v is identified as a **Strong Pillar** if its removal causes a significant degradation ($\Delta_{\mathcal{B}}(v) \leq -0.01$). We screen a pool of $|\widetilde{\mathcal{R}}| = 200$ candidates, extending beyond the core $K = 100$ set to ensure a wider search for these rare causal effects (see Appendix D.4 for statistical and operational details).

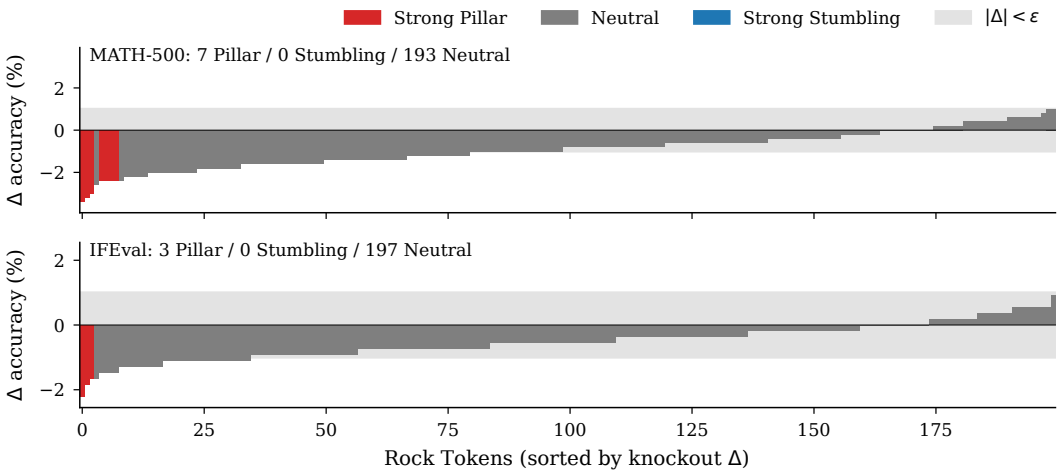


Figure 4: **Knockout effect on the $|\widetilde{\mathcal{R}}| = 200$ screened Rock-Token candidates, sorted by Δ.** Each bar is a single candidate; height is the accuracy change when its logit is masked at decode time. The shaded grey band marks the categorization threshold $|\Delta| < \varepsilon = 0.01$. Bars outside the band that pass the paired-bootstrap test ($\alpha = 0.05$, 10,000 resamples) are coloured by category (Strong Pillar in red; Strong Stumbling in blue). The vast majority of candidates are Individually Substitutable on both benchmarks; Pillars occupy a short tail and the symmetric Strong-Stumbling side is empty — the visual evidence underlying the sign-asymmetry test formalized below.

3.3 Are Rock Tokens Pillars? Findings and Orthogonality

The distribution of $\Delta_{\mathcal{B}}(v)$ in Figure 4 reveals the underlying nature of the Rock Token set. Our investigation yields three primary observations:

Pillars are rare but vital: Only 3.5% of Rock Tokens on MATH-500 and 1.5% on IFEval qualify as Strong Pillars. While most high-KL tokens are individually substitutable, a small subset is absolutely indispensable for the model's reasoning trajectory. **Directional Asymmetry:** The complete absence of Strong Stumbling Blocks suggests that the student's persistent deviations from the teacher are rarely detrimental to its own reasoning; the zero-effect mass is most consistent with substitutability rather than functional absence (see limitations below). **Orthogonality to Traditional Metrics:** To examine if Pillarhood can be predicted by standard signals, we analyzed the correlation between $\Delta_{\mathcal{B}}(v)$ and several distributional metrics, including student/teacher entropy, log-frequency, and residual KL. Crucially, none of these predictors exhibit a significant correlation with a token's causal importance ($|r| < 0.07$, see Appendix D.5 for full analysis and scatter plots). This demonstrates that Pillarhood is a property orthogonal to existing importance taxonomies [23], suggesting that importance schemes keyed to loss or entropy signals risk suppressing the very tokens the student model cannot reason without.

Limitations of the knockout protocol. The knockout policy sets $\pi_\theta^{\setminus v}$ by banning a single token type globally, so the test measures *irreplaceability* of the specific token rather than the functional necessity of the role it plays. Rock Tokens cluster into tight functional groups—LaTeX and math delimiters, whitespace markers, discourse-initial words—whose members are near-interchangeable. Banning one member (e.g., `\frac`) leaves close substitutes available within the same cluster, which plausibly explains the large Individually Substitutable mass: the structural function persists through a substitute, even though $\Delta \approx 0$. This means the Individually Substitutable label does *not* imply functional irrelevance; it implies that this particular token is not the *unique* carrier of the function. Importantly, the interpretation runs only one way: substitution can hide necessity, but cannot create it. The Strong Pillar results therefore remain valid and conservative—banning those tokens hurt accuracy despite substitutes being available, confirming genuine load-bearing status. Future work should complement individual-token knockout with cluster-level knockout (banning a complete functional cluster at once) to test whether the *function*, not just the token, is indispensable.

Take-away **The Concentration of Necessity.** While Rock Tokens as a population are indispensable for training stability, their causal necessity at inference is highly concentrated: only a tiny fraction (1.5%–3.5%) act as true **Pillars**. The remaining tokens are individually substitutable—not functionally absent—as the knockout protocol tests token-level irreplaceability, not the necessity of the structural role. Cluster-level experiments are needed to determine whether the underlying functions are likewise dispensable.

4 The genuine contribution of Rock Tokens

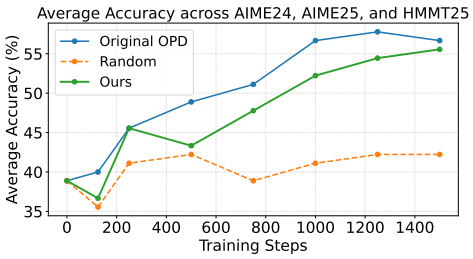


Figure 5: Average accuracy across AIME24, AIME25, and HMMT25 during OPD training. Each 200 training steps correspond to 8,000 prompts, with 4 rollouts per prompt.

4.1 RQ3: What is the genuine functional contribution of Rock Tokens to model training?

The functional redundancy of Rock Tokens at inference prompts a critical question: do their persistent high-loss signals provide essential constraints, or are they merely **optimization noise**? We consider two competing hypotheses: (1) **Late-stage Saturation**, where these tokens act as early alignment drivers but become unproductive "Stumbling Blocks" as the model reaches structural limits; or (2) **Inherent Inutility**, where they represent superficial stylistic gaps with no reasoning value from the outset.

To adjudicate between these views, we employ a **selective distillation strategy** by freezing the gradients of Rock Tokens. By comparing training trajectories with and without these signals, we aim to determine whether these persistent residuals are indispensable functional anchors or a detrimental optimization burden that consumes gradient bandwidth without yielding performance gains.

4.2 Selective Distillation via Token-Level Reweighting

To evaluate the functional necessity of these persistent residuals, we introduce a token-level reweighting mechanism to the OPD objective:

$$\mathcal{L}_{\text{weighted}} = \mathbb{E}_{x\sim\pi_\theta} \sum_{t=1}^{T} w(x_t) \cdot \ell_t, \quad \text{where } w(v) = \begin{cases} \lambda, & v \in \mathcal{R} \\ 1, & v \notin \mathcal{R} \end{cases} \tag{5}$$

Here, $\mathcal{R}$ is the Persistence Set defined in Section 2.2, and $\lambda \in [0, 1]$ serves as a scaling factor to modulate optimization pressure. This formulation allows us to test whether the gradient bandwidth consumed by tokens with high Rock Scores $R(v)$ is productive for learning or merely an optimization burden.

Our core experimental design employs three regimes by manipulating λ. The **Baseline OPD** ($\lambda = 1$) follows standard distillation without intervention. **Rock-Mask (Our Method)**: Based on the identified set $\mathcal{R}$, we adopt a "Just Not Train" approach by setting $\lambda = \mathbf{0}$. This strategy **zeros the loss terms** at positions whose realized token belongs to $\mathcal{R}$, removing those supervision signals without removing the tokens from the vocabulary, altering the rollout distribution, or freezing any model parameters, allowing us to determine if these tokens are indispensable functional anchors or simply unproductive residuals that can be bypassed to save computational budget. Finally, the **Freq-Matched Freeze (Random)** applies the same $\lambda = \mathbf{0}$ strategy to a randomly sampled set $\mathcal{S}$ that strictly matches the **token frequency distribution** of $\mathcal{R}$. By comparing these three settings, we can decouple the impact of structural Rock Tokens from stochastic noise and frequency-related artifacts.

4.3 Impact of Rock Tokens on Reasoning Performance

Figure 5 illustrates the performance trajectories across AIME24, AIME25, and HMMT25 under different intervention regimes. Our analysis reveals three key findings. First, while Rock Tokens exhibit persistent high loss, they are not merely "noise traps"; their inclusion provides essential gradients

that anchor the optimization process, as evidenced by the performance gap between the *Original OPD* and *Ours*. Second, while reducing total gradient signals is generally detrimental—shown by the severe degradation in the *Random* baseline—selectively down-weighting Rock Tokens is significantly safer, maintaining a much higher accuracy ceiling. Third, in our setup, Rock Tokens account for approximately 50% of the total output tokens; by mitigating the optimization pressure on these residuals, our empirical results demonstrate a $1.7\times$ increase in overall training wall-clock speed without substantial loss in reasoning integrity.

Take-away **The Efficiency-Performance Frontier:** While Rock Tokens provide necessary structural gradients, they exhibit significantly diminishing marginal utility. Our results demonstrate that applying selective loss masking to these $\sim$50% high-cost tokens yields a **1.7$\times$ wall-clock speedup** with minimal impact on reasoning integrity. This confirms that while Rock Tokens are not entirely disposable, **non-uniform supervision allocation** can effectively bypass computational redundancy in on-policy alignment.

Full ablation results comparing alternative token-selection criteria (top-frequency, top-mean-loss, gradient-magnitude, gradient-alignment) and a soft down-weighting sweep ($\lambda \in \{0.3, 0.5, 0.7\}$), together with per-variant masked-position fractions and wall-clock runtimes, are provided in Appendix F.

5 Experiment Setups

5.1 Off-policy and On-policy Distillation

We implement a two-stage distillation pipeline via the **KDFlow** framework [28] to transfer knowledge from a teacher, **Qwen3-30B-A3B-Instruct-2507** (MoE, 3B active), to a student, **Qwen3-4B-Instruct-2507** [24], with thinking mode disabled for both. **Stage 1 (Off-Policy):** To achieve initial *all-vocabulary alignment*, the student is trained for one epoch on 20k teacher-generated solutions from **OpenThoughts3** [6] using a forward KL objective (kd_ratio $= 0.5$) with cross-entropy. **Stage 2 (On-Policy):** Starting from the Stage 1 checkpoint, the student samples 4 rollouts each prompt from a separate 10k prompts as a checkpoint, (7 checkpoints in total) and is trained for one epoch on its own trajectories using pure reverse KL (kd_ratio $= 1.0$). Unless specified, all other hyperparameters follow KDFlow defaults; see Appendix C for full implementation details.

5.2 Evaluation

We employ the LM-EVALUATION-HARNESS framework [5] for standardized assessment across all benchmarks in a zero-shot setting. To ensure statistical robustness, we report the **Pass@1** accuracy averaged over five independent runs, accounting for variance in decoding.

Our evaluation suite focuses on challenging competitive mathematics, comprising **AIME 24**, **AIME 25**, and **HMMT 25-Feb** from MathArena [3]. Each benchmark contains 30 problems; we report the aggregated average score across these 90 tasks as the primary performance metric in our study. We extend our analysis to MATH-500 [14] and IFEval [30] to obtain a robust sample size for token-level statistical evaluation.

6 Related Works

6.1 On-Policy Distillation for LLM Post-Training

On-policy distillation evolves beyond SFT and RLVR by combining trajectory exploration with dense, token-level guidance [17]. Recent large-scale systems like DeepSeek-V4 [2], MiMo [20], and Qwen-3 [24] demonstrate its efficacy as a post-training stage, while self-distillation studies [29, 19, 9] suggest OPD is a general mechanism for amplifying latent model capabilities rather than mere imitation. However, the success of OPD is often hindered by teacher-student misalignment. While prior research examines this incompatibility through global factors like vocabulary, reasoning style, or model family [4, 13], our work focuses on a fine-grained source of stagnation. We investigate **Rock Tokens**—persistent high-loss positions that resist alignment even after training plateaus—to determine when OPD supervision provides productive signals versus unproductive noise.

Teacher–student compatibility also motivates personalized teacher selection: PerSyn routes prompts to teachers using student learnability and response quality [27]. DRP addresses inefficient reasoning through skill-aware step decomposition and pruning [11], while near-future guidance aligns short teacher and student continuations [10]. These methods change the source or structure of supervision; our study measures the learning utility of persistent high-loss token positions within OPD.

6.2 Token-Level Learning Dynamics in Reasoning Models

Recent studies demonstrate that reasoning in LLMs is disproportionately supported by critical "anchor" tokens, the perturbation of which during inference significantly degrades model performance [7]. Initial-token attention interventions also demonstrate that a specific token position can reshape downstream behavior without weight updates [8]; this inference-time effect is distinct from the training utility of persistent high-loss tokens studied here. In parallel, RLVR research reveals that training is non-uniform [25, 1, 22], concentrating optimization on high-entropy "forking tokens" that act as decisive branching points for reasoning gains [21], and further analysis confirms the overall token distribution drift mainly happens on some sparse but crucial tokens [18]. Although OPD's dense supervision suggests high-loss tokens drive training, their specific contribution mechanism remains unclear—especially when they persist as **Rock Tokens**. It is unknown whether these recalcitrant positions are indispensable reasoning pillars or merely unproductive residues. We address this gap by systematically analyzing their gradient dynamics and training effects to determine their true role in the OPD process.

Process supervision also changes the unit of guidance: IPS trains video-moderation models to answer ancillary questions [15]. Our analysis concerns a different granularity—whether individual token-level teacher signals remain useful during on-policy distillation.

Architectural efficiency is complementary to selective supervision: ZeroS uses signed, zero-sum weights to retain linear attention complexity while enabling contrastive operations [16]. Our intervention leaves the attention architecture unchanged and reduces distillation work by removing low-utility teacher supervision.

The modularized RL survey treats credit assignment as a separate learning-design dimension [26]. Our gradient and masking analyses address a related attribution question for dense distillation signals: which token positions contribute useful updates after loss has plateaued?

7 Conclusion

This paper investigates the role of high-loss tokens in On-Policy Distillation, revealing that while they occupy a dominant position in the loss landscape, they are not the decisive factor for performance gains. Our analysis identifies a distinct category of persistent tokens, termed **Rock Tokens**, despite their significant gradient magnitude and apparent "effort" during optimization, Rock Tokens fail to provide constructive guidance for the model's reasoning capabilities. Notably, our intervention experiments demonstrate that selectively masking the loss terms of these tokens—thereby removing their supervision without altering the rollout distribution or freezing model parameters—results in negligible impact on downstream performance while achieving a $1.7\times$ increase in training wall-clock speed. These findings challenge the conventional wisdom that aligning every high-loss token is indispensable for learning. Instead, we advocate for a more resource-efficient, token-selective approach that bypasses the computational redundancy of low-utility tokens to optimize the efficiency-performance frontier in on-policy alignment.

Funding and Competing Interests

The authors received no third-party funding or support for this work. The authors declare no competing interests.

References

[1] D. Cheng, S. Huang, X. Zhu, B. Dai, X. Zhao, Z. Zhang, and F. Wei. Reasoning with exploration: An entropy perspective. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2026.

[2] DeepSeek-AI. Deepseek-v4: Towards highly efficient million-token context intelligence. *Technical Report*, 2026.

[3] J. Dekoninck, N. Jovanović, T. Gehrunger, K. Rögnvaldsson, I. Petrov, C. Sun, and M. Vechev. Beyond benchmarks: Matharena as an evaluation platform for mathematics with llms, 2026. URL https://arxiv.org/abs/2605.00674.

[4] Y. Fu, H. Huang, K. Jiang, Y. Zhu, and D. Zhao. Revisiting on-policy distillation: Empirical failure modes and simple fixes. *ArXiv preprint*, 2026.

[5] L. Gao, J. Tow, B. Abbasi, S. Biderman, S. Black, A. DiPofi, C. Foster, L. Golding, J. Hsu, A. Le Noac'h, H. Li, K. McDonell, N. Muennighoff, C. Ociepa, J. Phang, L. Reynolds, H. Schoelkopf, A. Skowron, L. Sutawika, E. Tang, A. Thite, B. Wang, K. Wang, and A. Zou. The language model evaluation harness, 2024.

[6] E. Guha, R. Marten, S. S. Keh, N. Raoof, G. Smyrnis, H. Bansal, M. Nezhurina, J.-P. Mercat, T. Vu, Z. Sprague, A. Suvarna, B. Feuer, L. Chen, Z. Khan, E. Frankel, S. Grover, C. Choi, N. Muennighoff, S. Su, W. Zhao, J. Yang, S. Pimpalgaonkar, K. Sharma, C. C.-J. Ji, Y. Deng, S. Pratt, V. Ramanujan, J. Saad-Falcon, J. Li, A. Dave, A. Albalak, K. Arora, B. Wulfe, C. Hegde, G. Durrett, S. Oh, M. Bansal, S. Gabriel, A. Grover, K.-W. Chang, V. Shankar, A. Gokaslan, M. A. Merrill, T. Hashimoto, Y. Choi, J. Jitsev, R. Heckel, M. Sathiamoorthy, A. G. Dimakis, and L. Schmidt. Openthoughts: Data recipes for reasoning models. *arXiv.org*, 2025. doi: 10.48550/arxiv.2506.04178.

[7] A. Gupta, J. Yeung, G. Anumanchipalli, and A. A. Ivanova. How do llms use their depth? *arXiv.org*, 2025. doi: 10.48550/arxiv.2510.18871.

[8] F. Han, X. Yu, J. Tang, D. Rao, W. Du, and L. Ungar. ZeroTuning: Unlocking the initial token's power to enhance large language models without training. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/799b39274bdfeea24be0a3ccf304107e-Abstract-Conference.html.

[9] J. Hübotter, F. Lübeck, L. Behric, A. Baumann, M. Bagatella, D. Marta, I. Hakimi, I. Shenfeld, T. K. Buening, C. Guestrin, et al. Reinforcement learning via self-distillation. *ArXiv preprint*, 2026.

[10] Y. Jiang and F. Ferraro. Bridging reasoning trajectories in on-policy distillation via near-future guidance, 2026. URL https://arxiv.org/abs/2606.00305.

[11] Y. Jiang, D. Li, and F. Ferraro. DRP: Distilled reasoning pruning with mathematical skill-aware step decomposition for efficient large reasoning models. In M. Liakata, V. P. Moreira, J. Zhang, and D. Jurgens, editors, *Findings of the Association for Computational Linguistics: ACL 2026*, pages 4020–4039, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-395-1. doi: 10.18653/v1/2026.findings-acl.196. URL https://aclanthology.org/2026.findings-acl.196/.

[12] F. Kunstner, R. Yadav, A. Milligan, M. Schmidt, and A. Bietti. Heavy-tailed class imbalance and why adam outperforms gradient descent on language models, 2024. URL https://arxiv.org/abs/2402.19449.

[13] Y. Li, Y. Zuo, B. He, J. Zhang, C. Xiao, C. Qian, T. Yu, H. ang Gao, W. Yang, Z. Liu, and N. Ding. Rethinking on-policy distillation of large language models: Phenomenology, mechanism, and recipe, 2026. URL https://arxiv.org/abs/2604.13016.

[14] H. Lightman, V. Kosaraju, Y. Burda, H. Edwards, B. Baker, T. Lee, J. Leike, J. Schulman, I. Sutskever, and K. Cobbe. Let's verify step by step. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024.

[15] M. Liu, Y. Sun, R. Sun, X. Dong, X. Shen, H. Wang, H. Xiong, and Y. Song. IPS: In-prompt process supervision for short video content moderation. In Y. Li, G. Rehm, and M. Tu, editors, *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 6: Industry Track)*, pages 1277–1288, San Diego, California, USA, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-394-4. doi: 10.18653/v1/2026.acl-industry.89. URL https://aclanthology.org/2026.acl-industry.89/.

[16] J. Lu, X. Han, Y. Sun, V. Pati, Y. Kim, S. Somani, and S. Yang. ZeroS: Zero-sum linear attention for efficient transformers. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/1363163299a172662dcf0c0f9932acf6-Abstract-Conference.html.

[17] K. Lu and T. M. Lab. On-policy distillation. *Thinking Machines Lab: Connectionism*, 2025. doi: 10.64434/tml.20251026. https://thinkingmachines.ai/blog/on-policy-distillation.

[18] H. Meng, K. Huang, S. Wei, C. Ma, S. Yang, X. Wang, G. Wang, B. Ding, and J. Zhou. Sparse but critical: A token-level analysis of distributional shifts in rlvr fine-tuning of llms. *ArXiv preprint*, 2026.

[19] I. Shenfeld, M. Damani, J. Hübotter, and P. Agrawal. Self-distillation enables continual learning. *ArXiv preprint*, 2026.

[20] C. Team, B. Xiao, B. Xia, B. Yang, B. Gao, B. Shen, C. Zhang, C. He, C. Lou, F. Luo, G. Wang, G. Xie, H. Zhang, H. Lv, H. Li, H. Chen, H. Xu, H. Zhang, H. Liu, J. Duo, J. Wei, J. Xiao, J. Dong, J. Shi, J. Hu, K. Bao, K. Zhou, L. Li, L. Zhao, L. Zhang, P. Li, Q. Chen, S. Liu, S. Yu, S. Cao, S. Chen, S. Yu, S. Liu, T. Zhou, W. Su, W. Wang, W. Ma, X. Deng, B. Mao, B. Ye, C. Cai, C. Wang, C. Zhu, C. Ma, C. Chen, C. Li, D. Zhu, D. Xiao, D. Zhang, D. Zhang, F. Liu, F. Yang, F. Shi, G. Wang, H. Tian, H. Wu, H. Qu, H. Yi, H. An, H. Guan, X. Zhang, Y. Song, Y. Yan, Y. Zhao, Y. Lai, Y. Gao, Y. Cheng, Y. Tian, Y. Wang, Z. Tang, Z. Tang, Z. Wen, Z. Song, Z. Zheng, Z. Jiang, J. Wen, J. Sun, J. Li, J. Xue, J. Xia, K. Fang, M. Zhu, N. Chen, Q. Tu, Q. Zhang, Q. Wang, R. Li, R. Ma, S. Zhang, S. Wang, S. Li, S. Gu, S. Ren, S. Deng, T. Guo, T. Lu, W. Zhuang, W. Zhang, W. Xiong, W. Huang, W. Yang, X. Zhang, X. Yong, X. Wang, X. Xie, Y. Jiang, Y. Yang, Y. He, Y. Tu, Y. Dong, Y. Liu, Y. Ma, Y. Yu, Y. Xiang, Z. Huang, Z. Lin, Z. Xu, Z. Chen, Z. Deng, Z. Zhang, and Z. Yue. Mimo-v2-flash technical report, 2026. URL `https://arxiv.org/abs/2601.02780`.

[21] S. Wang, L. Yu, C. Gao, C. Zheng, S. Liu, R. Lu, K. Dang, X.-H. Chen, J. Yang, Z. Zhang, et al. Beyond the 80/20 rule: High-entropy minority tokens drive effective reinforcement learning for llm reasoning. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

[22] C.-C. Wu, Z. R. Tam, C.-Y. Lin, Y.-N. Chen, S.-H. Sun, and H. yi Lee. Mitigating forgetting in llm fine-tuning via low-perplexity token learning, 2025. URL `https://arxiv.org/abs/2501.14315`.

[23] Y. Xu, H. Sang, Z. Zhou, R. He, Z. Wang, and A. Geramifard. Tip: Token importance in on-policy distillation, 2026.

[24] A. Yang, A. Li, B. Yang, B. Zhang, B. Hui, B. Zheng, B. Yu, C. Gao, C. Huang, C. Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

[25] N. Yang, H. Lin, Y. Liu, B. Tian, G. Liu, and H. Zhang. Token-importance guided direct preference optimization. *arXiv.org*, 2025. doi: 10.48550/arxiv.2505.19653.

[26] Z. Yang, Y. Jiang, T.-C. Chen, L. Yang, A. Wong, C. Gao, J. E. Kooi, Z. Li, J. Shi, K. Qiu, Q. Huang, X. Zu, S. Yang, H. Zhang, N. Wong, F. Ilievski, S. Yu, A. Plaat, Z. Ren, M. Hoogendoorn, and V. François-Lavet. Modularized reinforcement learning on llms: From mdp creation to exploration and learning, 2026. URL `https://arxiv.org/abs/2606.21943`.

[27] H. Zhang, S. Yang, X. Liang, C. Shang, Y. Jiang, C. Tao, J. Xiong, H. K.-H. So, R. Xie, A. X. Chang, and N. Wong. Find your optimal teacher: Personalized data synthesis via router-guided multi-teacher distillation. *ACL 2026 Main*, 2026. URL `https://arxiv.org/abs/2510.10925`.

[28] S. Zhang, X. Zhang, T. Zhang, B. Hu, Y. Chen, and J. Xu. Kdflow: A user-friendly and efficient knowledge distillation framework for large language models. *ArXiv preprint*, 2026.

[29] S. Zhao, Z. Xie, M. Liu, J. Huang, G. Pang, F. Chen, and A. Grover. Self-distilled reasoner: On-policy self-distillation for large language models. *arXiv.org*, 2026. doi: 10.48550/arxiv.2601.18734.

[30] J. Zhou, T. Lu, S. Mishra, S. Brahma, S. Basu, Y. Luan, D. Zhou, and L. Hou. Instruction-following evaluation for large language models. *ArXiv preprint*, 2023.

A Limitations

First, our empirical evaluations rely predominantly on competitive mathematical reasoning; the functional role of Rock Tokens may differ in open-ended generation or coding tasks where structural boundaries are less rigid. Second, our experimental setup utilizes a specific 30B-to-4B distillation pairing, leaving the persistence and geometry of these tokens across different architectural families or vastly larger scales an open question. Finally, to isolate the functional impact of Stumbling Blocks, our mitigation strategy employs a strict binary gradient-freeze mask. While effective for demonstrating causal necessity, this approach is a blunt instrument; future work should explore dynamic reweighting schemes or soft-penalties that gracefully adapt to token-level loss trajectories throughout the optimization process.

A.1 Ethics Discussion and Risk Assessment

This work studies model distillation and training dynamics, and does not introduce a deployed system, human-subject data, or sensitive personal data. Its direct societal risks are therefore limited. However, because improvements in language model training may be used in both beneficial and harmful downstream applications, we acknowledge potential dual-use risks, including misuse for generating low-quality or misleading content.

We mitigate these risks by limiting the paper to research analysis and experimental evaluation. Any released code, models, or processed artifacts will be accompanied by appropriate documentation, intended-use information, and license details. All pretrained models, datasets, and software frameworks used in this work are credited, and their licenses and terms of use are respected.

B Analysis on Rock Tokens

We decode the top-$K{=}100$ rock tokens and their frequency-matched controls $\mathcal{S}_{\text{ctrl}}$ into their string forms using the public Qwen3 tokenizer. Despite being identified by a purely numerical criterion ($R(v)$), the resulting set is functionally coherent: rock tokens fall into a small number of identifiable categories, all of which mark structural or discourse boundaries in mathematical reasoning output.

Four functional clusters dominate $\mathcal{R}$.

1. **LaTeX and math delimiters** — $', \\, $$, =, ^, {, }, frac, (,). These mark transitions into and out of math mode and operate at the syntactic boundary between prose and formula.
2. **Markdown and whitespace structure** — **', \n', \n\n', :\n\n', —\n\n', ###'. These delimit logical units (paragraphs, section headers, bold spans).
3. **Discourse markers at sentence-initial position** — So', Let', We', But', Now', Wait', Then', Since', This'. These open new reasoning steps and signal turns in the chain of thought.
4. **Digits** — every digit 0' through 9' appears in $\mathcal{R}$, each with frequency in the thousands and a non-trivial mean per-token KL. These are tokens that punctuate quantitative statements.

Contrast with the frequency-matched control set. The control set $\mathcal{S}_{\text{ctrl}}$, sampled at the same frequency distribution, is qualitatively different: it consists predominantly of content-bearing tokens used at non-boundary positions — common nouns and verbs of mathematical discourse (e.g. triangle', equation', region', expression', must', find', check'), prepositions and conjunctions used mid-clause (as', between', first', right'), and single-letter variable tokens (a', b', i', f', y'). The frequency-matching procedure thus isolates the relevant axis: rocks and controls share the quantity of supervision they receive during OPD, but differ sharply in their syntactic role.

Interpretation The clustering of $\mathcal{R}$ at structural boundaries indicates that the residual student–teacher divergence after on-policy distillation is concentrated at the points where the model decides *how to structure* its reasoning output, rather than *what content* to emit. Decisions like open math mode", begin a new paragraph", insert a section header", or start a new reasoning step with So' versus Then'," are precisely the positions where the OPD loss persists, even after extensive distillation. This observation suggests that rock tokens are not arbitrary statistical outliers but a consistent signature of the on-policy distillation regime: they identify the structural/discourse decisions that the student systematically fails to align with the teacher.

C Training and Evaluation Details

C.1 off policy and on policy training

Training Details. We distill from Qwen3-30B-A3B-Instruct-2507, a Mixture-of-Experts teacher with $\sim$30B total and $\sim$3B active parameters, into Qwen3-4B-Instruct-2507 as the student, using the KDFlow framework. Both models are run with thinking mode disabled (`enable_thinking=False`). Training proceeds in two stages on a single node of $4\times$ H100 (80 GB) GPUs with FSDP2, bf16, and gradient checkpointing.

Stage 1 – Off-policy KD. We first sample 20k teacher responses on math prompts drawn from OpenThoughts3 (temperature $= 0.6$, top_p $= 0.95$, max_new_tokens $= 16384$, TP $= 2$ on $2\times$ H100). The student is then trained on these (prompt, teacher-response) pairs with a per-token KL distillation loss combined with cross-entropy at kd_ratio $= 0.5$ (vanilla KD, forward KL). We use AdamW with learning rate 2×10^{-5}, 5% linear warmup, global batch size 128 (micro-batch 1), max sequence length 16384, sample packing, and ring attention of size 2, for 1 epoch.

Stage 2 – On-policy KD. Initialized from the Stage-1 checkpoint, the student generates 4 rollouts per prompt (temperature $= 1.0$, top_p $= 1.0$, generate_max_len $= 8000$, prompt_max_len $= 800$) on a separate 10k-prompt slice (positions 20k–30k of the same source), and is distilled toward the teacher's token distributions on those rollouts. We use vanilla KD with reverse KL at kd_ratio $= 1.0$ (pure distillation, no CE), learning rate 2×10^{-6}, 5% linear warmup, gradient clipping at 1.0, global batch size 4 (micro-batch 1), and 1 epoch. The rollout engine uses 1 engine with TP $= 2$ and the teacher uses TP $= 4$; both engines share GPUs with the trainer via offload-to-CPU sleep/wakeup (`teacher_enable_sleep=True`, `rollout_enable_sleep=True`.

Defaults. All other hyperparameters are KDFlow defaults; we did not tune optimizer betas, weight decay, dropout, or KD temperature (kept at 1.0).

D Pillar Correlation and Multiple-Testing Analysis

D.1 Pillar Correlation Analysis

Table 1 reports the full set of Pearson correlations between the inference-time knockout effect $\Delta_{\mathcal{B}}(v)$ (Eq. 6) and nine candidate token-level predictors, measured over the $|\widetilde{\mathcal{R}}| = 200$ screened Rock-Token candidates of §D.4 and reported separately for MATH-500 and IFEval. The figure in the main text (Figure 6) visualizes a representative subset of six of these predictors on MATH-500; this table is the complete record, including IFEval and the three loss-based features (pre-/post-OPD KL and KL improvement). Across all eighteen $(predictor, benchmark)$ pairs, no correlation reaches $|r| > 0.075$ and no p-value falls below 0.30. The largest observed magnitude on either benchmark is the IFEval correlation with KL improvement ($r = +0.071$, p=0.32), which is well within the noise band expected at n=200. We therefore cannot reject the null hypothesis $r = 0$ for any predictor on either benchmark, supporting the claim in §D.5 that Pillarhood is statistically orthogonal to the entropy-, frequency-, and divergence-based importance signals used by prior work [21, 23, 25, 22].

Table 1: Pearson correlation between the per-token knockout effect $\Delta_{\mathcal{B}}(v)$ and nine candidate predictors over the $|\widetilde{\mathcal{R}}| = 200$ screened Rock-Token candidates, reported for MATH-500 and IFEval. p-values are two-sided. No correlation reaches $|r| > 0.075$ on either benchmark; no p-value falls below 0.30.

Predictor	**MATH-500**		**IFEval**	
	r	p	r	p
Frequency	$+0.015$	0.832	$+0.030$	0.677
Rock count	$+0.013$	0.855	$+0.023$	0.751
Rock rate	-0.022	0.760	-0.047	0.512
Pre-OPD KL	-0.017	0.808	-0.015	0.833
Post-OPD KL	$+0.010$	0.883	-0.048	0.503
KL improvement	-0.062	0.386	$+0.071$	0.320
Teacher entropy	$+0.063$	0.378	-0.036	0.611
Pre-OPD student entropy	$+0.059$	0.406	-0.069	0.331
Post-OPD student entropy	$+0.056$	0.432	-0.037	0.604

D.2 Multiple-Testing Analysis for the Per-Token Pillar Criterion

The per-token Pillar criterion of §D.3 applies a paired bootstrap test at $\alpha = 0.05$ independently to each of the $|\widetilde{\mathcal{R}}| = 200$ screened Rock-Token candidates, conjoined with the effect-size filter $|\Delta_{\mathcal{B}}(v)| \geq \varepsilon$ ($\varepsilon = 0.01$). Because ε is approximately at the per-token sampling-error scale, the

conjunction does not appreciably reduce the per-test Type-I rate, and $200 \times 0.05 = 10$ rejections are expected under the global null. We report below how the per-token identifications behave under standard family-wise and false-discovery corrections, and the global sign-asymmetry test on which the population-level claim of §D.4 rests.

Per-token corrections. Bonferroni control of the family-wise error at $\alpha = 0.05$ requires $p < \alpha/200 = 2.5 \times 10^{-4}$. The smallest observed p-values are 0.0038 (MATH-500, "certain") and 0.0016 (IFEval, "-like"); neither passes Bonferroni, and consequently no token does. Benjamini–Hochberg control of the false discovery rate at $q \in \{0.05, 0.10, 0.20\}$ likewise produces zero rejections on either benchmark: the BH cutoff $p_{(k)} \leq (k/n)\, q$ is not crossed even for the smallest-p token, because $1/200 \times 0.20 = 10^{-3} < 0.0016$. The per-token Pillar identifications are therefore *exploratory candidates*, not confirmatory rejections in the multiple-testing sense.

Global sign-asymmetry test. The exploratory candidates do, however, exhibit a non-trivial structural property: the sign of the significant rejection is consistent across benchmarks. Of the seven MATH-500 candidates, all seven satisfy $\Delta_{\mathrm{MATH500}}(v) < 0$ (Pillar direction); of the three IFEval candidates, all three satisfy $\Delta_{\mathrm{IFEval}}(v) < 0$. Combined, the rejection sign-distribution is 10 negative, 0 positive. Under the global null in which every per-token rejection is a chance event, the rejection sign should be Bernoulli($1/2$), and the probability of observing all 10 on one side is $2^{-10} \approx 9.8 \times 10^{-4}$ (two-sided exact binomial $p = 1.95 \times 10^{-3}$). On MATH-500 alone the analogous sign test is $7/0$ with two-sided $p = 0.016$; on IFEval alone $3/0$ with $p = 0.25$. The combined test is a single global test that does not inflate with the number of tokens screened, and it is the result on which §D.4 relies for the population-level claim that a non-empty subset of Rock Tokens is causally necessary at decode time.

Summary. The reader is advised to read the per-token Pillar identifications of §D.4 (e.g. "certain", "strategic", "Initialize") as the strongest individual candidates from a population-level effect, rather than as 7 independently-confirmed discoveries. The empirical contribution of this section that survives strict multiple-testing scrutiny is the sign asymmetry, which is incompatible with the multiplicity-noise null at $p \approx 0.002$.

Stable-core robustness check. The screening pool of §D.3 is the top-$K_{\text{screen}} = 200$ count-ranked extension of $\mathcal{R}$, while the definitional Rock-Token set fixed by the stability/coverage analysis of §2 is the top-$K = 100$. The 10 Strong Pillars from the bootstrap distribute as follows over the count-rank index. On MATH-500 (ranks reported on the top-200 list): "certain" (16), "Initialize" (93), "strategic" (99), "Do" (109), "programs" (143), "tech" (166), "arms" (185). On IFEval: "-like" (35), "tools" (64), "trade" (69). Six of the ten lie within the $K = 100$ core, including the smallest-p token on each benchmark ("certain" at $p = 0.0038$ on MATH-500 and "-like" at $p = 0.0016$ on IFEval). The remaining four lie in the rank-109-to-185 regime, which §2 flags as below the Jaccard-stability sweet spot; we therefore caution that those four are at greater risk of being corpus-specific. The sign-asymmetry test of the previous paragraph is robust to this concern: restricted to the $K = 100$ stable core, it remains $6/0$ negative-versus-positive, with two-sided exact-binomial $p = 2^{-6} \times 2 = 3.1 \times 10^{-2}$.

D.3 Definition and Identification of 'Pillar Tokens'

To investigate this, we operationally categorize Rock Tokens based on their causal impact at inference time. **Pillar Tokens** are defined as those whose removal or perturbation degrades task accuracy, suggesting the model *cannot reason without them*, while the remainder are deemed functionally neutral.

Intuition. The core idea is that if a token is truly load-bearing for reasoning, deleting it from the decoder should break the trajectory; if it only *looks* "hard" due to high loss but lacks functional utility, its deletion should change nothing. We operationalize Pillarhood as exactly this causal, inference-time property. Given the post-OPD student π_θ and a candidate $v \in \widetilde{\mathcal{R}}$ (screening pool defined below), we form a knockout policy $\pi_\theta^{\setminus v}$ that sets the logit of v to $-\infty$ at every decoding step and record the accuracy delta on a held-out benchmark $\mathcal{B}$:

$$\Delta_{\mathcal{B}}(v) = \mathrm{Acc}_{\mathcal{B}}\big(\pi_\theta^{\setminus v}\big) - \mathrm{Acc}_{\mathcal{B}}\big(\pi_\theta\big). \tag{6}$$

Table 2: Pillar census over the $|\widetilde{\mathcal{R}}| = 200$ screened Rock-Token candidates, by benchmark.

Benchmark	Strong Pillar	Neutral	Strong Stumbling	Baseline acc.
MATH-500	7 (3.5%)	193 (96.5%)	0 (0.0%)	0.750
IFEval	3 (1.5%)	197 (98.5%)	0 (0.0%)	0.752
Cross-task agreement: 190 N|N, 7 P|N, 3 N|P, **0** P|P				

A candidate v is a **Strong Pillar** on $\mathcal{B}$ iff (i) $\Delta_{\mathcal{B}}(v) \leq -\varepsilon$ ($\varepsilon = 0.01$) and (ii) a paired bootstrap on per-problem indicators (10,000 resamples, $\alpha = 0.05$) rejects $\mathrm{H}_0 : \Delta = 0$; symmetric thresholds at $+\varepsilon$ define Strong Stumbling Blocks, and all other candidates are Neutral. Accuracies are estimated by sampling at T=1.0 and averaging over independent rollouts to keep single-trajectory variance below the $\sim 2\%$ effect scale.

The screening pool is a strict superset $\widetilde{\mathcal{R}} \supset \mathcal{R}$ of size $K_{\text{screen}} = 200$ rather than the definitional $K = 100$ set $\mathcal{R}$ of §2.2: $K = 100$ is the stability/coverage sweet spot when $\mathcal{R}$ is treated as a vocabulary (qualitative analysis, gradient-zeroing in §4), but Pillarhood is a per-token causal property and the bootstrap is a search for individual rare-but-real effects, so we trade selection stability below rank 100 for a wider candidate pool. As a robustness check (Appendix D.2), 6 of the 10 Strong Pillars — including the strongest on each benchmark, " certain" on MATH-500 and "-like" on IFEval — lie within the $K = 100$ core. We probe the post-OPD student on two complementary benchmarks: MATH-500 [14] for the trained reasoning capability and IFEval [30] for instruction-following as an out-of-distribution control.

D.4 Are Rock Tokens pillar? Experiments and findings

Table 2 reports the full Pillar census from the bootstrap procedure across the $|\widetilde{\mathcal{R}}| = 200$ screened candidates. The dominant feature of the per-token $\Delta_{\mathcal{B}}(v)$ distribution is not the tail but the centre: on both benchmarks the great majority of candidates fall inside the ε-band, indicating no detectable causal effect of their removal (visualization in Fig. 4, Appendix D.2).

Multiple-testing considerations. At $|\widetilde{\mathcal{R}}| = 200$ candidates, ~ 10 false rejections are expected by chance at $\alpha = 0.05$, and neither Bonferroni nor Benjamini–Hochberg ($q \leq 0.20$) retains any individual token (Appendix D.2). We therefore read the per-token identifications above as exploratory *candidate* Pillars. The population-level claim is carried by the *sign* of the rejections: all 10 fall on the Pillar (negative-Δ) side and zero on the Stumbling side, a 10:0 split with exact probability $2^{-10} \approx 0.002$ under the multiplicity-noise null — a single global test that does not inflate with the number of candidates screened.

Three observations follow. **First**, Pillars are rare: 3.5% on MATH-500 and 1.5% on IFEval — the persistent high-KL criterion of §2 over-selects causally relevant tokens by more than an order of magnitude. **Second**, Pillarhood is task-specific: no token is a Strong Pillar on both benchmarks (P|P$= 0$ in Table 2), so the role of a token type depends on the task. **Third**, Strong Pillars are content-bearing tokens at decision points, not the structural delimiters that dominate the surface clustering of $\mathcal{R}$ in §2: MATH-500 examples include " certain", " strategic", " Initialize", " programs", "Do"; IFEval includes " tools", " trade", "-like". The symmetric Strong-Stumbling direction is empty under inference-time knockout, an asymmetry we revisit at training time in §4.

D.5 Does Rock Token follow previous findings on 'Forking Tokens'?

Intuition. The dominant prior framing casts "important tokens" as high-entropy decision points where the policy splits between alternative reasoning continuations: forking tokens in RLVR [21], and the (entropy $\cup$ student–teacher-divergence) taxonomy formalized for OPD by TIP [23] and related token-reweighting work [25, 22]. If our Pillars were the inference-time projection of any of these training-time importance criteria, knockout $\Delta_{\mathcal{B}}(v)$ should be predicted by entropy, by KL, or by some combination thereof.

Findings. Figure 6 scatters MATH-500 knockout $\Delta_{\mathrm{MATH500}}(v)$ against six candidate predictors — pre/post-OPD student entropy, teacher entropy, $\log_{10}$ frequency, rock rate, and post-OPD KL —

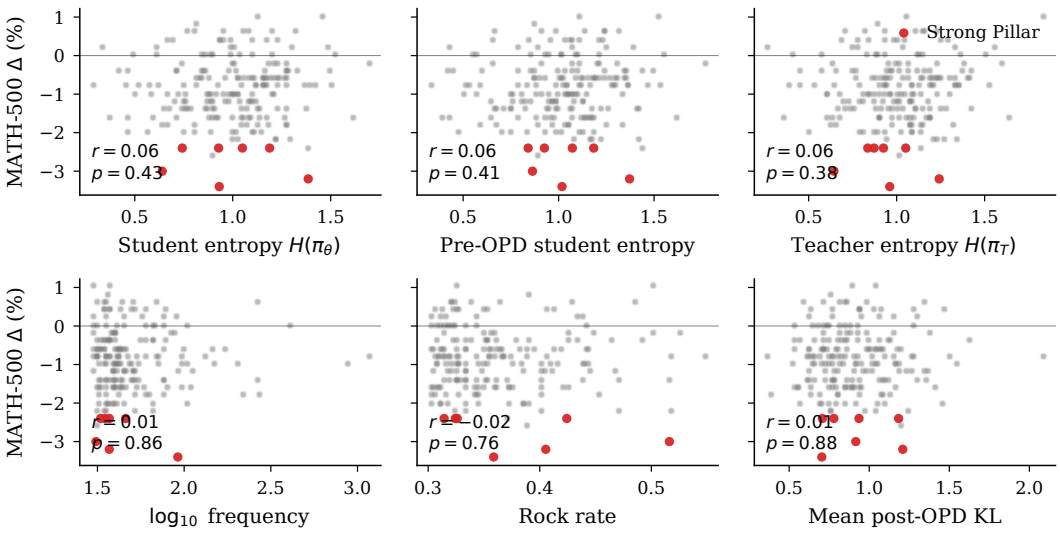


Figure 6: **Pillarhood is not predicted by entropy, frequency, or loss.** MATH-500 knockout Δ for each of the $|\widetilde{\mathcal{R}}| = 200$ screened candidates (Strong Pillars in red) plotted against six candidate predictors: post- and pre-OPD student entropy, teacher entropy, log-frequency, rock rate, and mean post-OPD KL. Annotated r, p are Pearson correlations over all 200 candidates. None reach $|r| > 0.07$.

and *none* reaches $|r| > 0.07$; the same conclusion holds on IFEval and on three additional loss-based predictors (Table 1, Appendix D.1). Pillars are scattered across the entire entropy range, not concentrated where forking tokens reside, and they are not the tokens with the largest residual KL — the criterion that $\mathcal{R}$ itself maximizes. Pillarhood is therefore an *orthogonal* cut: a causal, behaviour-defined property that the existing entropy- and divergence-based taxonomies fail to surface, with the immediate methodological corollary that any token-importance scheme keyed to those signals risks suppressing exactly the tokens the student cannot reason without.

Take-away Pillar Tokens are not Forking Tokens. Their causal role at inference is uncorrelated with student entropy, teacher entropy, frequency, or post-OPD KL ($|r| < 0.07$ on MATH-500), placing them outside the importance taxonomies of prior work [21, 23].

E Selection of Ratio

Choosing the selection size. The score $R(v)$ provides a ranking; we still need a cutoff K that determines $|\mathcal{R}|$. Two competing criteria bear on this choice. First, K should be small enough that the top-K set is stable — i.e., reproducible across resamplings of the rollout corpus. Second, K should be large enough that $\mathcal{R}$ accounts for a substantial fraction of the corpus-level KL, so that reweighting decisions made on $\mathcal{R}$ have meaningful effect on $\mathcal{L}_{\text{OPD}}$.

Procedure We sweep K jointly with sample size. For each $n \in 50, 100, 200, 300, 400$, we compute $\widehat{R}(v)$ from the corresponding sub-corpus and select its top-K tokens; we then measure the Jaccard index of this set against the top-K obtained at $n = 500$. We additionally compute the cumulative coverage $\sum_{v \in \mathcal{R}} \widehat{R}(v)/\mathcal{L}_{\text{OPD}}^{\text{corpus}}$ at $n = 500$ as K grows.

Results Figure 2(b) plots Jaccard against K, with one curve per n and an averaged curve. Average Jaccard exceeds 0.70 for $K \leq 100$ and decays toward a ≈ 0.57 floor for $K \geq 150$ — beyond which the additional tokens are no longer reliably reproduced across sub-corpora. Figure 2(c) overlays the corresponding coverage curve: at $K = 100$, $\mathcal{R}$ accounts for $\approx 60\%$ of the corpus-level KL while remaining within the stable regime. We therefore set $K = 100$ ($\approx 2.5\%$ of the observed vocabulary) as the *definitional* Rock-Token set $\mathcal{R}$ used wherever a stable, reproducible vocabulary is required — the qualitative cluster analysis below, and the training-time interventions of Section 4. This choice represents an explicit tradeoff: $K = 50$ would yield modestly higher stability ($J \approx 0.81$) but reduce coverage to $\approx 49\%$. For the per-token causal analysis of Appendix D.4 we instead *screen* a strict

superset of size $K_{\text{screen}} = 200$: that section's bootstrap is a search for individual rare-but-real Pillars, and the broader candidate pool both increases the chance of detection and lets us probe the regime in which selection stability has decayed below $J \approx 0.6$. The two cutoffs serve different purposes — the smaller K defines $\mathcal{R}$, the larger one widens the screening window — and we are explicit in Appendix D.3 about which is in force.

F Selection Criterion Ablation and Runtime Analysis

To isolate which component of the Rock Score drives the efficiency gains in Figure 5, we compare Rock-Mask against five alternative selection strategies that each remove the same number of tokens (K=100): top-frequency ($\mathrm{Freq}(v)$ alone), top-mean-loss ($\bar{\ell}_v$ alone), top-gradient-magnitude ($\|\bar{g}_t\|$), top-gradient-alignment ($\cos(\bar{g}_t, G_{\text{balanced}})$), and the frequency-matched random baseline already reported in the main text. We additionally sweep soft weights $\lambda \in \{0.3, 0.5, 0.7\}$ on the Rock Score set to trace the accuracy–compute frontier. All conditions share the same training data, hardware, step budget, and five-seed evaluation protocol as the main experiment. Soft-λ results use a corrected per-token loss normalizer; see rebuttal for details.

Table 3: Comparison of token-selection criteria for Rock-Mask. Soft-λ variants retain a fraction of the Rock-token loss weight; hard-mask variants (λ=0) zero it entirely. Accuracy is averaged over AIME24, AIME25, and HMMT25, reported as mean $\pm$ std over five independent runs. Δ is relative to Rock-Mask (Ours).

Method	Selection Criterion	λ	AIME24	AIME25	HMMT25	Avg.
Original OPD	—	1	56.7 ± 1.7	51.8 ± 0.8	36.7 ± 2.4	**48.1**
Soft λ=0.7	Rock Score	0.7	56.4 ± 1.5	50.1 ± 0.8	34.8 ± 1.2	47.1
Soft λ=0.5	Rock Score	0.5	56.2 ± 1.7	48.3 ± 1.2	34.1 ± 0.7	46.2
Soft λ=0.3	Rock Score	0.3	56.5 ± 1.2	45.2 ± 1.3	33.6 ± 0.9	45.1
Rock-Mask (Ours)	$\mathrm{Freq}(v) \cdot \bar{\ell}_v$, top-100	0	56.7 ± 1.4	43.3 ± 0.8	33.0 ± 2.6	**44.3**
Top-gradient-magnitude	$\|\bar{g}_t\|$, top-100	0	51.1 ± 7.7	42.2 ± 8.4	25.6 ± 5.1	39.6
Freq-Matched (Random)	random, freq.-matched	0	46.7 ± 3.4	40.0 ± 2.2	30.0 ± 0.4	38.9
Top-mean-loss	$\bar{\ell}_v$ alone, top-100	0	48.9 ± 1.9	42.2 ± 3.9	23.3 ± 0.0	38.1
Top-gradient-alignment	$\cos(\bar{g}_t, G_{\text{bal}})$, top-100	0	42.2 ± 7.7	32.2 ± 1.9	26.7 ± 3.3	33.7
Top-frequency	$\mathrm{Freq}(v)$ alone, top-100	0	32.2 ± 1.9	41.1 ± 3.8	17.8 ± 3.9	30.4

Among hard-mask variants, Rock-Mask achieves the highest average accuracy (44.3), retaining 92.1% of full-OPD performance. The joint $\mathrm{Freq}(v) \cdot \bar{\ell}_v$ criterion consistently outperforms either factor alone: top-frequency (30.4) and top-mean-loss (38.1) both fall substantially below Rock-Mask, confirming that neither occurrence count nor per-token KL alone is sufficient to identify low-return supervision. Gradient-based selectors (magnitude: 39.6; alignment: 33.7) underperform Rock-Mask despite targeting the tokens whose optimization signals should, in principle, be most redundant. The soft-λ sweep traces a monotone frontier from full OPD (48.1 at λ=1) to Rock-Mask (44.3 at λ=0), confirming that the efficiency gain scales continuously with the degree of Rock-token suppression.

Table 4 reports masked-position fractions and wall-clock speedups for each hard-mask variant. All methods remove a comparable fraction of positions (52–58%) and achieve similar wall-clock speedups (1.67–1.70×). Because computational savings are nearly identical across criteria, the performance gap in Table 3 reflects selection quality rather than budget.

G Robustness of Rock Token Identification to Decoding Configuration

The Rock Score $R(v)$ is estimated from a greedy-decoded rollout corpus (Section 2.3), whereas Stage-2 OPD training uses on-policy sampled rollouts (T=1.0, top_p=1.0). To verify that the identified vocabulary does not depend on this choice, we re-estimated $R(v)$ from a sampled rollout corpus generated from the same checkpoint under the Stage-2 decoding configuration.

Despite the greedy and sampled corpora sharing only 68% sequence-level overlap—confirming that the two configurations produce substantially different reasoning trajectories—the identified Rock vocabulary is highly stable: the **Top-50 Rock Tokens are identical**, and **96 out of the Top-100 Rock Tokens** are shared between the two decoding configurations (Jaccard = 0.96).

Table 4: Masked-position fraction and wall-clock speedup for each hard-mask (λ=0) variant. All methods operate at comparable computational cost; performance differences arise from selection criterion, not budget.

Method	Masked Positions (%)	Wall-Clock Speedup
Freq-Matched (Random)	55.1	$1.69\times$
Top-frequency	56.5	$1.70\times$
Top-mean-loss	52.1	$1.68\times$
Top-gradient-magnitude	54.7	$1.69\times$
Top-gradient-alignment	53.2	$1.67\times$
Rock-Mask (Ours)	**58.3**	$\mathbf{1.70}\times$

This result indicates that Rock Tokens reflect a structural property of the student–teacher divergence rather than an artifact of the attribution decoding strategy, and that the Rock-Mask intervention results reported in Section 4 generalize to the on-policy sampling regime used during training.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Yes, I've checked.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Yes, In appendix A.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

Justification: We don't have proof part, equations are for formal definitions.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Yes, in section 5 and appendix B.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [No]

Justification: Our main contribution is not on code or benchmark. We use the open source training frameworks, and model checkpoints are too large to upload.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Yes, in section 5 and appendix B.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: We include evaluation details, and significant analysis for all experiments.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.

- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: In section 5.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We have verifed.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: In appendix A.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: We don't have sucn information.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: In appendix A.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: We have verified.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: We don't have human annotation.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: Not envolve human experiments.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: Our work only use LLM for writing.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.