Critical Batch Size for LLM Policy Optimization

author names withheld

Under Review for the Workshop on High-dimensional Learning Dynamics, 2026

Abstract

Supervised learning's critical batch size defines the point at which increasing batch size leads to lost sample efficiency, which controls the extent to which data-parallel can be used to improve training efficiency. We study critical batch size for verifier-based reinforcement learning (RLVR) under a GRPO-style objective, where gradient noise depends on prompts B, rollouts per prompt K, and off-policy rollout reuse. We extend the noise-scale model from [9] to GRPO by decomposing on-policy noise into inter-prompt and intra-prompt terms, and modeling off-policy reuse as drift-inflated intra-prompt noise. We empirically measure the critical batch size in both on-policy and off-policy settings. In our experiments, we find that the gradient noise is dominated by the intra-prompt term and that the relevant batch dimension is approximately the total rollout count $N = BK$. We find that off-policy rollout reuse substantially increases the critical batch size relative to the on-policy setting, suggesting a practical parallelism advantage for RLVR post-training.

1. Introduction

The central design question in large-scale optimization is how much we can parallelize learning before it stops paying off. Classical results for supervised learning [9, 16] characterize this through the *critical batch size* (CBS): for batches below this scale, larger batches reduce the number of training steps nearly proportionally; above it, additional examples mostly waste compute. Equivalently, the CBS measures the maximum data-parallelism the objective can absorb per gradient step before per-sample returns diminish.

Unlike supervised training, reinforcement learning from verifiable rewards (RLVR) [8, 17] must generate trajectories from the current policy before each update, and this generation step often dominates wall-clock time. The resulting batch has three axes: B prompts, K rollouts per prompt, and multiple gradient updates from reused rollouts. Each axis affects gradient noise and therefore the CBS.

In this paper we explore the following question: *How does the critical batch size (CBS) of RLVR decompose across these three axes, and how does off-policy rollout reuse change the CBS relative to on-policy training?*

We answer by extending the McCandlish noise-scale model [9] to the two-level sampling structure of GRPO [17]. In the on-policy regime, we show that the prompt-axis CBS decomposes into an inter-prompt term and an intra-prompt rollout term, and we model off-policy rollout reuse as inflating the intra-prompt term according to the policy drift from the behavior policy. Empirically, on GRPO training of Qwen2.5-Math-Instruct-1.5B [21] on AIME1983-2024 [20], we find that standard on-policy regimes are dominated by intra-prompt rollout noise, so the total rollout batch $N = BK$ is the primary statistical batch dimension. Finally, we show that off-policy reuse substantially in-

creases the fitted critical rollout batch size, extending the apparent linear-scaling regime relative to on-policy training. These results characterize when GRPO benefits from more prompts, more rollouts, or more rollout reuse, and provide a practical noise-scale framework for allocating parallelism in RLVR post-training.

2. Background

We provide a brief overview of the primary works which we use to formalize CBS for RLVR; additional related work is given in Appendix A.

GRPO GRPO is a policy-optimization algorithm designed for LLM training. Optimization alternates between two stages: (1) response generation, and (2) policy updates. In the first phase, B prompts from the dataset are selected at random $\{q_i\}_{i=1}^{B}$, and K responses per prompt are generated from the policy. Response j for prompt i, denoted $o_{i,j}$, receives reward $r_{i,j}$ and group-relative advantage $A_{i,j} = \frac{r_{i,j}-\mu(q_i)}{\sigma(q_i)}$ where $\mu(q_i)$ and $\sigma(q_i)$ are the mean and standard deviation of the rewards for prompt q_i. The BK trajectories are split into minibatches $\mathcal{B}_m$, each containing $b := B/M$ prompts (we assume mini-batches split by prompt), and cycle for E epochs, giving $T := ME$ inner updates.

Critical batch size McCandlish et al. [9] show that stochastic gradient noise sets a ceiling on useful SGD batch size. For loss $L(\theta) = \mathbb{E}_x[L_x(\theta)]$, gradient $G = \nabla L$, Hessian $H = \nabla^2 L$, and unbiased batch estimator $\mathrm{G}_{\text{est}} = \frac{1}{B}\sum_i \nabla L_{x_i}$ with $\mathrm{Cov}(\mathrm{G}_{\text{est}}) = \Sigma/B$, a second-order expansion of $L(\theta - \epsilon\,\mathrm{G}_{\text{est}})$ gives an optimal step size and per-step loss decrement that both scale as $1/(1 + \mathcal{B}_{\text{noise}}/B)$, where

$$\mathcal{B}_{\text{noise}} = \frac{\mathrm{tr}(H\Sigma)}{G^\top H G}, \qquad S(B) = S_\infty\left(1 + \frac{\mathcal{B}_{\text{critical}}}{B}\right), \qquad \mathcal{B}_{\text{critical}} \equiv \mathcal{B}_{\text{noise}}\,.$$

Thus, for $B \ll \mathcal{B}_{\text{critical}}$, doubling B roughly halves steps-to-target; for $B \gg \mathcal{B}_{\text{critical}}$, larger batches mostly waste samples. Since the RL objective shifts during training, we follow standard RL for LLMs practice and fit $\mathcal{B}_{\text{critical}}$ using stable downstream error as a proxy for loss.

3. Theoretical CBS for Policy Optimization

We extend the McCandlish noise-scale framework to GRPO training. The key structural difference from supervised learning is that GRPO batches have two axes of variation, prompts and rollouts per prompt, which leads to a noise scale that decomposes into inter-prompt and intra-prompt components. We state results first for the on-policy regime in Theorem 1, and then for the off-policy regime in Theorem 2.

Theorem 1 (On-Policy CBS) *Under on-policy evaluation ($T = 1$), let $\{q_i\}_{i=1}^{B}$ be prompts drawn i.i.d., and let $\{o_{i,k}\}_{k=1}^{K}$ be trajectories drawn i.i.d. from $\pi_\theta(\cdot \mid q_i)$ for each prompt i. The per-prompt mean gradient is*

$$\hat{g}(q;\, o_{1:K}) \;=\; \frac{1}{K}\sum_{k=1}^{K} \hat{A}_k(o_{1:K})\,\nabla_\theta \log \pi_\theta(o_k \mid q).$$

Define the inter-prompt and intra-prompt variance as

$$\Sigma_q \;=\; \mathrm{Var}_q\big(\, \mathop{\mathbb{E}}_{o|q}[\hat{g}(q;o_{1:K})]\big), \qquad \Sigma_{o|q} \;=\; \mathbb{E}_q\Big[\mathit{Var}_{o|q}\big(\hat{g}(q;\,o_{1:K})\big)\Big].$$

Let $\sigma^2_{\mathit{inter}} = \mathit{tr}(H\,\Sigma_q)/(G^\top H G)$ *and* $\sigma^2_{\mathit{intra}} = \mathit{tr}(H\,\Sigma_{o|q})/(G^\top H G)$*, where* H *is the surrogate objective Hessian. The critical batch size, as a function of the number of trajectories per prompt, is:*

$$\mathcal{B}_{\mathit{critical}}(K) \;=\; \sigma^2_{\mathit{inter}} \;+\; \frac{\sigma^2_{\mathit{intra}}}{K}. \tag{1}$$

We defer the proof to Appendix C. $\mathcal{B}_{\text{critical}}$ decomposes additively into a prompt-diversity term σ^2_{inter}, and a within-prompt term $\sigma^2_{\text{intra}}/K$, which decays at the standard $1/K$ rate with added rollouts. As $K \to \infty$, $\mathcal{B}_{\text{critical}} \to \sigma^2_{\text{inter}}$; the gap between $\mathcal{B}_{\text{critical}}$ at small and large K is therefore a direct estimator of the variance ratio $\sigma^2_{\text{intra}} \,/\, \sigma^2_{\text{inter}}$.

Next, we turn to the critical batch size for the off-policy regime. We make the following assumptions, discussed in further detail in Appendix D.2: **(A1)** $\mathrm{KL}(\pi_{\theta_t} \| \pi_{\theta_0}) \ll 1$ for all $t < T$; **(A2)** $G(\theta)$, $H(\theta)$, $\Sigma_q(\theta)$, $\Sigma_{o|q}(\theta)$ are approximately constant in θ over the inner loop; **(A3)** the objective is used without clipping.

Theorem 2 (Off-Policy CBS) *Under assumptions (A1)–(A3), and with* Σ_q, $\Sigma_{o|q}$, $\sigma^2_{\mathit{inter}}$, $\sigma^2_{\mathit{intra}}$ *as in Theorem 1 (which the off-policy quantities reduce to under (A2)), the critical batch size at inner step* t *is:*

$$\mathcal{B}_{\mathit{critical}}{}^{(t)}(K) \;=\; \sigma^2_{\mathit{inter}} \;+\; \big[1 + (t\kappa)^2\big]\, \frac{\sigma^2_{\mathit{intra}}}{K}, \tag{2}$$

where $\kappa^2 := \epsilon^2\, G^\top \bar{F}\, G \;\approx\; 2\,\mathrm{KL}_{\textit{per inner step}}$*, with* $\bar{F} = \mathbb{E}_q[F_q(\theta_0)]$ *the prompt-averaged Fisher information and* ϵ *the learning rate.*

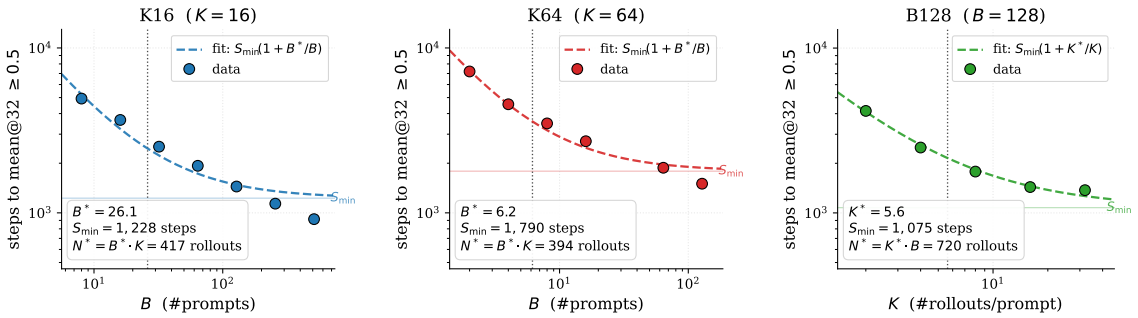
We defer the proof to Appendx D.2. Only the intra-prompt term inflates with off-policy reuse. The inter-prompt contribution σ^2_{inter} is unchanged because prompts are drawn exogenously from the data distribution, independently of any policy. The intra-prompt term $\sigma^2_{\text{intra}}/K$ is multiplied by $1 + (t\kappa)^2$, quadratic in the cumulative parameter drift $t\kappa$. Setting $t = 0$ recovers Theorem 1. The drift parameter κ is fully determined by the per-inner-step KL: a smaller learning rate reduces κ and extends the off-policy reuse horizon.

4. Empirical Analysis of CBS for GRPO

We use the Qwen2.5-Math-Instruct-1.5B [21] model and train on the DAPO [22] dataset. For validation, we use the AIME1983- 2024 [20]. We report all accuracies with `pass@1/mean@32`, representing the average correctness of the model when prompted with a question 32 times. We report hyperparameters for all experiments in Appendix E.2, F.2. Every experiment runs on 4 NVIDIA H100 GPUs, and we use the Verl library [18].

4.1. On-Policy Results

We first evaluate the on-policy noise decomposition $B_{\text{crit}} = \sigma^2_{\text{inter}} + \sigma^2_{\text{intra}}/K$ (equivalently, the rollout-batch CBS $N_{\text{crit}} = K\sigma^2_{\text{inter}} + \sigma^2_{\text{intra}}$ where $N = BK$). Performance is tracked via mean@32



(a) K16: $K = 16$; varying B. (b) K64: $K = 64$; varying B. (c) B128: Varying K; $B = 128$.

Figure 1: McCandlish-style fits of steps-to-target S versus the swept axis for the three on-policy experiments. Each panel fits the hyperbolic form $S(x) = S_{\min}(1 + x^*/x)$ to the data and extracts the critical value $x^* \in \{B^*, K^*\}$. Translated to rollout units, the three independent experiments yield $N^* \in \{417, 394, 720\}$ with geometric mean ≈ 484.

on the evaluation set. For each run with prompt batch B and rollouts-per-prompt K, we record the number of outer iterations S required to first reach mean@32 ≥ 0.5; this defines the empirical steps-to-target curve $S(B, K)$ from which CBS is fitted.

We conduct three experiment sweeps, each holding one of $\{B, G\}$ fixed and varying the other. The inner gradient batch in each case is $B \times G$ rollouts. Across the three settings, we measure 24 (B, G, S) measurement triples (further details given in Appendix E.1).

For each experiment we fit the hyperbolic form $S(x) = a + b/x$ to the swept axis $x \in \{B, K\}$ and extract a critical value $x^* = b/a$ (as shown in Figure 1). Table 2 summarizes the extracted critical values translated to the rollout-batch CBS $N^* = B^*K$ (or $N^* = BK^*$).

To extract $(\sigma^2_{\text{inter}}, \sigma^2_{\text{intra}})$ jointly, we fit all data points to

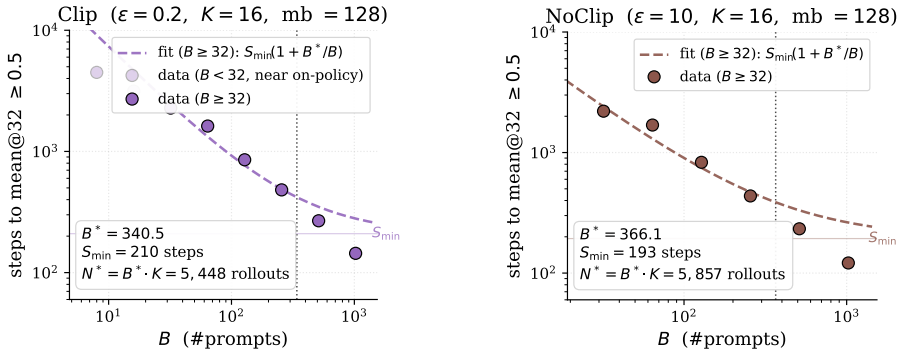
$$S(B, K) = S_{\min}\left(1 + \frac{\sigma^2_{\text{inter}}}{B} + \frac{\sigma^2_{\text{intra}}}{BK}\right),$$

yielding $\sigma^2_{\text{inter}} \approx 3.2$, $\sigma^2_{\text{intra}} \approx 311$, and $S_{\min} \approx 1465$, with mean relative residual of 15%. The residual is dominated by the B128 sweep, where the joint model under-predicts K^* by roughly $2\times$ relative to the per-experiment fit; the K16 and K64 sweeps alone are reproduced to within ± 4. Thus the intra-prompt term dominates inter-prompt.

4.2. Off-Policy Results

In the off-policy setting, each parameter update involves collecting $B \times K$ rollouts at the behavior policy and processing them in mini-batches of b prompts for $T = B/b$ gradient steps. We test the off-policy drift-inflated noise scale $N^{(t)}_{\text{crit}} = K\sigma^2_{\text{inter}} + \rho(t)\sigma^2_{\text{intra}}$ with $\rho(t) \approx 1 + (t\kappa)^2$, and the consequent saturation prediction $N^* \gtrsim b/\kappa$ in rollout units, against runs that take multiple gradient steps per data collection.

We conduct two experiment sweeps, both with $K = 16$ rollouts per prompt and 128 samples per mini-batch (so $T = B/8$ inner steps per outer iteration), differing only in the PPO clipping range ϵ: (1) **Clip**, in which we use the standard PPO clipping value $\epsilon = 0.2$, and (2) **NoClip**, in which we set a high clipping value $\epsilon = 10$ that results in effectively no clipping.



(a) Clip: $\epsilon = 0.2$.

(b) NoClip: $\epsilon = 10$.

Figure 2: McCandlish-style fits of steps-to-target S versus prompt batch B for the two off-policy sweeps, target $\text{mean@32} \geq 0.5$. Hyperbolic fit $S(B) = S_{\min}(1 + B^*/B)$ is restricted to $B \geq 32$. Translated to rollout units, the two settings give $N^* \approx 5\,448$ and $N^* \approx 5\,857$.

We fit the hyperbolic form on each experiment (Figure 2). The measured per-inner-step drift is similar for Clip and NoClip, with $\kappa = 0.080$ and $\kappa = 0.088$, respectively; the corresponding fitted rollout knees are $N^* \approx 5{,}448$ and $N^* \approx 5{,}857$ (see Table 4). At matched B, Clip and NoClip have indistinguishable $S \cdot N$ at the target metric, with the difference between them appearing only in the asymptotic ceiling and not in the linear-regime CBS.

4.3. Discussion of results

Our empirical findings suggest the following main takeaways. Further discussion of results is given in Appendix E regarding the on-policy setting and Appendix F regarding off-policy.

Intra-prompt rollout noise dominates inter-prompt noise. In our empirical estimates $\sigma^2_{\text{inter}} \approx 3.2$ and $\sigma^2_{\text{intra}} \approx 311$, most of the noise comes from variation among rollouts for the same prompt rather than variation across prompts.

The relevant batch axis is total rollouts, not prompt count alone. For all $K \lesssim 100$ (which covers standard practice), $N^* \approx \sigma^2_{\text{intra}} \approx 311$ is the only number that matters for gradient noise. The split of N between B and K is a hardware-utilization choice, not a statistical-efficiency one.

Off-policy offers a substantial parallelism advantage relative to on-policy. The empirical off-policy CBS is about $N^* \approx 5,400$–$5,900$ rollouts, compared to the on-policy CBS of roughly 311 rollouts. This suggests a compute-parallelism advantage of off-policy GRPO at fixed problem scale.

5. Conclusion

We study critical batch size for LLM policy optimization using GRPO, decomposing on-policy noise into inter-prompt and intra-prompt components and modeling off-policy reuse as drift-inflated noise. Empirically, the total generated rollouts are the relevant unit for on-policy batch scaling, while off-policy reuse substantially increases the fitted CBS by amortizing data collection across multiple updates. We leave to future work broader validation across model scales, task distributions, training stages, and different PPO-style training objectives.

References

[1] Aili Chen, Aonian Li, Bangwei Gong, Binyang Jiang, Bo Fei, Bo Yang, Boji Shan, Changqing Yu, Chao Wang, Cheng Zhu, et al. Minimax-m1: Scaling test-time compute efficiently with lightning attention. *arXiv preprint arXiv:2506.13585*, 2025.

[2] Corinna Cortes, Yishay Mansour, and Mehryar Mohri. Learning bounds for importance weighting. In J. Lafferty, C. Williams, J. Shawe-Taylor, R. Zemel, and A. Culotta, editors, *Advances in Neural Information Processing Systems*, volume 23. Curran Associates, Inc., 2010. URL https://proceedings.neurips.cc/paper_files/paper/2010/file/59c33016884a62116be975a9bb8257e3-Paper.pdf.

[3] Josef Dai, Xuehai Pan, Ruiyang Sun, Jiaming Ji, Xinbo Xu, Mickel Liu, Yizhou Wang, and Yaodong Yang. Safe rlhf: Safe reinforcement learning from human feedback. *arXiv preprint arXiv:2310.12773*, 2023.

[4] Jacob Hilton et al. Scaling data-constrained language models. *Advances in Neural Information Processing Systems*, 2023.

[5] Prateek Jain, Sham M. Kakade, Rahul Kidambi, Praneeth Netrapalli, and Aaron Sidford. Parallelizing stochastic gradient descent for least squares regression: Mini-batching, averaging, and model misspecification. *Journal of Machine Learning Research*, 18(223):1–42, 2018. URL http://jmlr.org/papers/v18/16-595.html.

[6] Sham M Kakade. A natural policy gradient. *Advances in neural information processing systems*, 14, 2001.

[7] Devvrit Khatri, Lovish Madaan, Rishabh Tiwari, Rachit Bansal, Sai Surya Duvvuri, Manzil Zaheer, Inderjit S. Dhillon, David Brandfonbrener, and Rishabh Agarwal. The art of scaling reinforcement learning compute for llms. *arXiv preprint arXiv:2510.13786*, 2025.

[8] Nathan Lambert, Jacob Morrison, Valentina Pyatkin, Shengyi Huang, Hamish Ivison, Faeze Brahman, Lester James V. Miranda, Alisa Liu, Nouha Dziri, Shane Lyu, Yuling Gu, Saumya Malik, Victoria Graf, Jena D. Hwang, Jiangjiang Yang, Ronan Le Bras, Oyvind Tafjord, Chris Wilhelm, Luca Soldaini, Noah A. Smith, Yizhong Wang, Pradeep Dasigi, and Hannaneh Hajishirzi. Tülu 3: Pushing frontiers in open language model post-training. *arXiv preprint arXiv:2411.15124*, 2024.

[9] Sam McCandlish, Jared Kaplan, Dario Amodei, and Dario Amodei OpenAI. An empirical model of large-batch training. *arXiv preprint arXiv:1812.06162*, 2018.

[10] Alberto Maria Metelli, Matteo Papini, Francesco Faccio, and Marcello Restelli. Policy optimization via importance sampling. *CoRR*, abs/1809.06098, 2018. URL http://arxiv.org/abs/1809.06098.

[11] Clara Mohri, Amir Globerson, Haim Kaplan, Tomer Koren, and Yishay Mansour. Cost-aware learning. *arXiv preprint arXiv:2604.28020*, 2026.

[12] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744, 2022.

[13] Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741, 2023.

[14] John Schulman, Sergey Levine, Pieter Abbeel, Michael Jordan, and Philipp Moritz. Trust region policy optimization. In *International conference on machine learning*, pages 1889–1897. PMLR, 2015.

[15] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

[16] Christopher J Shallue, Jaehoon Lee, Joseph Antognini, Jascha Sohl-Dickstein, Roy Frostig, and George E Dahl. Measuring the effects of data parallelism on neural network training. *Journal of Machine Learning Research*, 20(112):1–49, 2019.

[17] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Mingchuan Zhang, Y.K. Li, Y. Wu, and Daya Guo. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

[18] Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng, Haibin Lin, and Chuan Wu. Hybridflow: A flexible and efficient rlhf framework. *arXiv preprint arXiv: 2409.19256*, 2024.

[19] Richard S Sutton, David McAllester, Satinder Singh, and Yishay Mansour. Policy gradient methods for reinforcement learning with function approximation. In S. Solla, T. Leen, and K. Müller, editors, *Advances in Neural Information Processing Systems*, volume 12. MIT Press, 1999. URL https://proceedings.neurips.cc/paper_files/paper/1999/file/464d828b85b0bed98e80ade0a5c43b0f-Paper.pdf.

[20] Hemish Veeraboina. Aime problem set 1983-2024, 2023. URL https://www.kaggle.com/datasets/hemishveeraboina/aime-problem-set-1983-2024.

[21] An Yang, Beichen Zhang, Binyuan Hui, Bofei Gao, Bowen Yu, Chengpeng Li, Dayiheng Liu, Jianhong Tu, Jingren Zhou, Junyang Lin, et al. Qwen2. 5-math technical report: Toward mathematical expert model via self-improvement. *arXiv preprint arXiv:2409.12122*, 2024.

[22] Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, et al. Dapo: An open-source llm reinforcement learning system at scale. *arXiv preprint arXiv:2503.14476*, 2025.

[23] Hanlin Zhang, Depen Morwani, Nikhil Vyas, Jingfeng Wu, Difan Zou, Udaya Ghai, Dean Foster, and Sham Kakade. How does critical batch size scale in pre-training?, 2025. URL https://arxiv.org/abs/2410.21676.

[24] Rui Zheng, Shihan Dou, Songyang Gao, Yuan Hua, Wei Shen, Binghai Wang, Yan Liu, Senjie Jin, Qin Liu, Yuhao Zhou, et al. Secrets of rlhf in large language models part i: Ppo. *arXiv preprint arXiv:2307.04964*, 2023.

Contents

Appendix A. Related work

RL for Language Models. While policy gradient methods [6, 14, 19] have been used in many settings, the works of Dai et al. [3], Ouyang et al. [12] have garnered significant attention for their application to LLM post-training. Various methods such as DPO [13], PPO [15, 24], and GRPO [17] have been shown to be extremely effective in improving accuracy on standard reasoning benchmarks. Further recent work [1, 7, 11, 22] explore the best way to scale RL compute.

Critical batch size and scaling laws Several works have characterized *critical batch size* and related limits of mini-batch parallelism [5, 9, 16]. Shallue et al. [16] empirically characterize diminishing returns from large mini-batches across neural-network workloads, while Jain et al. [5] theoretically analyze analogous limits in least-squares regression. [9] develop the theory on which we base our analysis and include empirical findings for RL tasks (Atari and Dota), but do not specifically focus on the RLVR setting. [23] also performs an empirical study of CBS for pretraining. ScaleRL [7] studies how to scale RL compute for LLM post-training and provides empirical scaling laws for validation performance under increased rollout/training compute, whereas our focus is on explaining the critical batch size and gradient-noise decomposition underlying when additional rollout parallelism remains useful.

Appendix B. Rederivation of McCandlish result

Consider a model parameterized by $\theta \in \mathbb{R}^D$, with per-example loss $L_x(\theta)$ for data point $x \sim \rho$, and full loss $L(\theta) = \mathbb{E}_{x\sim\rho}[L_x(\theta)]$. Let $G(\theta) \equiv \nabla L(\theta)$ be the true gradient and $H(\theta) \equiv \nabla^2 L(\theta)$ the true Hessian at θ.

The gradient is estimated by averaging over a batch of B i.i.d. samples $x_1, \ldots, x_B \sim \rho$:

$$G_{\text{est}}(\theta) \;=\; \frac{1}{B}\sum_{i=1}^{B} \nabla_\theta L_{x_i}(\theta), \qquad x_i \sim \rho. \tag{3}$$

Because each $\nabla_\theta L_{x_i}(\theta)$ is an i.i.d. unbiased estimator of $G(\theta)$, linearity of expectation and the standard variance-of-the-mean identity give

$$\mathop{\mathbb{E}}_{x_{1:B}\sim\rho}[G_{\text{est}}(\theta)] \;=\; G(\theta), \qquad \mathrm{Cov}_{x_{1:B}\sim\rho}(G_{\text{est}}(\theta)) \;=\; \frac{1}{B}\,\Sigma(\theta), \tag{4}$$

where $\Sigma(\theta)$ is the per-example gradient covariance:

$$\Sigma(\theta) \;\equiv\; \mathrm{Cov}_{x\sim\rho}(\nabla_\theta L_x(\theta)) \;=\; \mathop{\mathbb{E}}_{x\sim\rho}\Big[(\nabla_\theta L_x(\theta))\,(\nabla_\theta L_x(\theta))^\top\Big] - G(\theta)G(\theta)^\top. \tag{5}$$

Perturb the parameters along a direction V with step size ϵ: $\theta \mapsto \theta - \epsilon V$. Taylor-expanding L to second order gives

$$L(\theta - \epsilon V) \;\approx\; L(\theta) \;-\; \epsilon\, G^\top V \;+\; \tfrac{1}{2}\,\epsilon^2\, V^\top H V. \tag{6}$$

If we had access to the true gradient and chose $V = G$, then $L(\theta - \epsilon G) \approx L(\theta) - \epsilon|G|^2 + \frac{1}{2}\epsilon^2\, G^\top H G$. Differentiating w.r.t. ϵ and setting to zero,

$$-|G|^2 + \epsilon\, G^\top H G = 0 \quad\Longrightarrow\quad \boxed{\epsilon_{\max} \;=\; \frac{|G|^2}{G^\top H G}.}$$

Now use the noisy estimator $V = \mathrm{G}_{\text{est}}$. The expected updated loss is

$$\mathbb{E}[L(\theta - \epsilon\,\mathrm{G}_{\text{est}})] \;=\; L(\theta) \;-\; \epsilon\,|G|^2 \;+\; \tfrac{1}{2}\,\epsilon^2\!\left(G^\top H G \;+\; \frac{\mathrm{tr}(H\Sigma)}{B}\right). \tag{7}$$

Differentiating with respect to ϵ and setting to zero:

$$-|G|^2 + \epsilon\left(G^\top H G + \frac{\mathrm{tr}(H\Sigma)}{B}\right) = 0,$$

yielding

$$\epsilon_{\text{opt}}(B) \;=\; \frac{|G|^2}{G^\top H G + \mathrm{tr}(H\Sigma)/B} \;=\; \frac{|G|^2/(G^\top H G)}{1 + \mathrm{tr}(H\Sigma)/(B\,G^\top H G)}.$$

Defining the *gradient noise scale*

$$\boxed{\mathcal{B}_{\text{noise}} \;=\; \frac{\mathrm{tr}(H\Sigma)}{G^\top H G},} \tag{8}$$

this becomes

$$\epsilon_{\text{opt}}(B) \;=\; \arg\min_{\epsilon}\,\mathbb{E}[L(\theta - \epsilon\,\mathrm{G}_{\text{est}})] \;=\; \frac{\epsilon_{\max}}{1 + \mathcal{B}_{\text{noise}}/B}. \tag{9}$$

Let $\mathcal{L}_{\text{opt}}(B) \equiv L(\theta) - \mathbb{E}[L(\theta - \epsilon_{\text{opt}}\,\mathrm{G}_{\text{est}})]$, that is,

$$\mathcal{L}_{\text{opt}}(B) \;=\; \epsilon_{\text{opt}}\,|G|^2 \;-\; \tfrac{1}{2}\,{\epsilon_{\text{opt}}}^2\!\left(G^\top H G + \tfrac{\mathrm{tr}(H\Sigma)}{B}\right).$$

At the optimum, $\epsilon_{\text{opt}}\big(G^\top H G + \mathrm{tr}(H\Sigma)/B\big) = |G|^2$, so the second term equals $\frac{1}{2}\,\epsilon_{\text{opt}}\,|G|^2$ and

$$\mathcal{L}_{\text{opt}}(B) \;=\; \tfrac{1}{2}\,\epsilon_{\text{opt}}\,|G|^2 \;=\; \frac{1}{2}\,\frac{|G|^4}{G^\top H G + \mathrm{tr}(H\Sigma)/B}.$$

Factoring out $G^\top H G$ in the denominator,

$$\mathcal{L}_{\text{opt}}(B) \;=\; \frac{\mathcal{L}_{\max}}{1 + \mathcal{B}_{\text{noise}}/B}, \qquad \mathcal{L}_{\max} \;=\; \frac{1}{2}\,\frac{|G|^4}{G^\top H G}. \tag{10}$$

- **Small-batch regime** ($B \ll \mathcal{B}_{\text{noise}}$): $\mathcal{L}_{\text{opt}}(B) \approx \mathcal{L}_{\max}\; B/\,\mathcal{B}_{\text{noise}}$, so progress per step grows linearly with B. Equivalently, total examples $E = BS$ to reach a target loss is roughly constant in B: doubling B halves the steps needed.
- **Large-batch regime** ($B \gg \mathcal{B}_{\text{noise}}$): $\mathcal{L}_{\text{opt}}(B) \approx \mathcal{L}_{\max}$, so increasing B has negligible effect on the loss decrement; extra compute is wasted.

The transition happens at $B \approx \mathcal{B}_{\text{noise}}$, and hence $\mathcal{B}_{\text{critical}} := \mathcal{B}_{\text{noise}} = \frac{\mathrm{tr}(H\Sigma)/B}{G^\top H G}$.

Appendix C. Proof of Theorem 1

For minibatch $\mathcal{B}_m$, GRPO maximizes

$$J_m(\theta) = \frac{1}{b} \sum_{(i,k)\in\mathcal{B}_m} \frac{1}{|o_{i,k}|} \sum_{\tau=1}^{|o_{i,k}|} \min\Big(\rho_{i,k,\tau}(\theta)\hat{A}_{i,k}, \mathrm{clip}(\rho_{i,k,\tau}(\theta), 1-\varepsilon, 1+\varepsilon)\hat{A}_{i,k}\Big), \tag{11}$$

where $\rho_{i,k,\tau}(\theta) = \pi_\theta(o_{i,k,\tau} \mid q_i, o_{i,k,<\tau}) / \pi_{\theta_{\text{old}}}(o_{i,k,\tau} \mid q_i, o_{i,k,<\tau})$. We note that $J_m(\theta)$ is is a surrogate objective.

We re-state the theorem and provide the proof.

Theorem 1 (On-Policy CBS) *Under on-policy evaluation ($T = 1$), let $\{q_i\}_{i=1}^B$ be prompts drawn i.i.d., and let $\{o_{i,k}\}_{k=1}^K$ be trajectories drawn i.i.d. from $\pi_\theta(\cdot \mid q_i)$ for each prompt i. The per-prompt mean gradient is*

$$\hat{g}(q;\, o_{1:K}) \;=\; \frac{1}{K}\sum_{k=1}^{K} \hat{A}_k(o_{1:K})\, \nabla_\theta \log \pi_\theta(o_k \mid q).$$

Define the inter-prompt and intra-prompt variance as

$$\Sigma_q \;=\; \mathrm{Var}_q\big(\mathop{\mathbb{E}}_{o|q}[\hat{g}(q; o_{1:K})]\big), \qquad \Sigma_{o|q} \;=\; \mathbb{E}_q\Big[\mathit{Var}_{o|q}\big(\hat{g}(q;\, o_{1:K})\big)\Big].$$

Let $\sigma^2_{\mathit{inter}} = \mathit{tr}(H\,\Sigma_q)/(G^\top H G)$ and $\sigma^2_{\mathit{intra}} = \mathit{tr}(H\,\Sigma_{o|q})/(G^\top H G)$, where H is the surrogate objective Hessian. The critical batch size, as a function of the number of trajectories per prompt, is:

$$\mathcal{B}_{\mathit{critical}}(K) \;=\; \sigma^2_{\mathit{inter}} + \frac{\sigma^2_{\mathit{intra}}}{K}. \tag{1}$$

Proof In the on-policy regime ($T = 1$), the importance ratios in $J_m(\theta)$ (Eq. 11) are identically 1 at $\theta = \theta_{\text{old}}$, and clipping is inactive. The surrogate objective reduces to

$$J(\theta;\, \theta_{\text{old}},\, \mathcal{D}) \;=\; \frac{1}{B}\sum_{i=1}^{B}\frac{1}{K}\sum_{k=1}^{K} \hat{A}_k(o_{i,1:K})\, \frac{\pi_\theta(o_{i,k} \mid q_i)}{\pi_{\theta_{\text{old}}}(o_{i,k} \mid q_i)}.$$

For a given batch $\mathcal{D} = \{(q_i,\, o_{i,1:K})\}$, this is a fixed function of θ.

The stochastic gradient at $\theta = \theta_{\text{old}}$ is

$$\mathrm{G}_{\text{est}} \;=\; \nabla_\theta J\big|_{\theta=\theta_{\text{old}}} \;=\; \frac{1}{B}\sum_{i=1}^{B} \hat{g}(q_i;\, o_{i,1:K}),$$

where $\hat{g}(q;\, o_{1:K}) = \frac{1}{K}\sum_{k=1}^{K} \hat{A}_k(o_{1:K})\, \nabla_\theta \log \pi_\theta(o_k \mid q)$.

Moments of the gradient estimator. Since prompts are drawn i.i.d. and rollouts are drawn i.i.d. from $\pi_{\theta_{\text{old}}}(\cdot \mid q)$, and under our advantage formulation the estimator is unbiased:

$$\mathbb{E}[\mathrm{G}_{\text{est}}] \;=\; G(\theta_{\text{old}}).$$

Applying the law of total variance across prompts and rollouts:

$$\mathrm{Cov}(\mathrm{G}_{\text{est}}) \;=\; \frac{1}{B}\,\Sigma_q \;+\; \frac{1}{BK}\,\Sigma_{o|q},$$

where Σ_q and $\Sigma_{o|q}$ are as defined in the theorem statement.

Quadratic approximation. We Taylor expand the surrogate to second order around θ_{old}. Let $H = \nabla^2_\theta J\big|_{\theta=\theta_{\text{old}}}$ denote the Hessian of the surrogate. Taking expectations over the batch randomness and using $\mathbb{E}[\mathrm{G_{est}}\,\mathrm{G_{est}}^\top] = GG^\top + \mathrm{Cov}(\mathrm{G_{est}})$:

$$\mathbb{E}\big[J(\theta_{\text{old}} - \epsilon\ \mathrm{G_{est}})\big] \;\approx\; J(\theta_{\text{old}}) \;-\; \epsilon\,|G|^2 \;+\; \frac{\epsilon^2}{2}\left(G^\top H G \;+\; \frac{\mathrm{tr}(H\,\Sigma_q)}{B} \;+\; \frac{\mathrm{tr}(H\,\Sigma_{o|q})}{BK}\right).$$

Optimal step size. Differentiating with respect to ϵ and setting to zero:

$$\epsilon_{\text{opt}}(B) \;=\; \frac{|G|^2}{G^\top H G + \frac{\mathrm{tr}(H\,\Sigma_q)}{B} + \frac{\mathrm{tr}(H\,\Sigma_{o|q})}{BK}} \;=\; \frac{\epsilon_{\max}}{1 + \mathcal{B}_{\text{noise}}\,/B},$$

where $\epsilon_{\max} = |G|^2/(G^\top H G)$ and

$$\mathcal{B}_{\text{noise}} \;=\; \frac{\mathrm{tr}(H\,\Sigma_q)}{G^\top H G} \;+\; \frac{\mathrm{tr}(H\,\Sigma_{o|q})}{K\,G^\top H G}$$

is obtained by factoring $1/B$ from both noise terms in the denominator.

Critical batch size. Defining $\sigma^2_{\text{inter}} = \mathrm{tr}(H\,\Sigma_q)/(G^\top H G)$ and $\sigma^2_{\text{intra}} = \mathrm{tr}(H\,\Sigma_{o|q})/(G^\top H G)$:

$$\mathcal{B}_{\text{critical}}(K) \;=\; \sigma^2_{\text{inter}} \;+\; \frac{\sigma^2_{\text{intra}}}{K}.$$

The critical batch size decomposes into an inter-prompt component (irreducible by adding rollouts) and an intra-prompt component (reduced by increasing K). ■

Appendix D. Proof of Theorem 2

We extend the McCandlish noise model to the off-policy setting by combining it with the off-policy importance-sampling framework of Metelli et al. [10], which characterizes how the variance of an importance-sampling estimator scales with the dissimilarity between the target and behavior distributions.

D.1. Background: importance sampling and Rényi-bounded variance

Since the off-policy GRPO gradient is built from importance-weighted samples, we first recall the relevant tools from [10] for bounding the variance of such estimators.

Setting. Let P, Q be probability measures on a measurable space $(\mathcal{X}, \mathcal{F})$ with P absolutely continuous with respect to Q, and let $f : \mathcal{X} \to \mathbb{R}$ be bounded ($\|f\|_\infty < \infty$). Given i.i.d. samples $x_1, \ldots, x_N \sim Q$, the *importance-sampling (IS) estimator* of $\mathbb{E}_{x\sim P}[f(x)]$ is

$$\hat{\mu}_{P/Q} \;=\; \frac{1}{N}\sum_{i=1}^{N} \frac{p(x_i)}{q(x_i)}\, f(x_i) \;=\; \frac{1}{N}\sum_{i=1}^{N} w_{P/Q}(x_i)\, f(x_i),$$

where $w_{P/Q}(x) := p(x)/q(x)$ is the importance weight. This estimator is unbiased, $\mathbb{E}_{x\sim Q}[\hat{\mu}_{P/Q}] = \mathbb{E}_{x\sim P}[f]$, but its variance can blow up when P and Q are dissimilar, because the weights $w_{P/Q}$ become heavy-tailed.

Rényi divergence. The dissimilarity between P and Q is measured by the family of α-Rényi divergences,

$$D_\alpha(P\|Q) \;=\; \frac{1}{\alpha-1}\log\int q(x)\left(\frac{p(x)}{q(x)}\right)^\alpha dx, \qquad \alpha\in(0,1)\cup(1,\infty),$$

with the special cases $D_1(P\|Q)=\mathrm{KL}(P\|Q)$ as $\alpha\to 1$ and $D_\infty(P\|Q)=\log \operatorname{ess\,sup}_x w_{P/Q}(x)$. We write $d_\alpha(P\|Q) := \exp D_\alpha(P\|Q)$ for the exponentiated form. The Rényi divergence directly controls the moments of the importance weights: in particular, $\mathbb{E}_{x\sim Q}[w_{P/Q}(x)^2]=d_2(P\|Q)$.

Variance bound (Lemma 4.1 of Metelli et al. [10]). The variance of the IS estimator is bounded by

$$\mathrm{Var}_{x\sim Q}\big[\hat\mu_{P/Q}\big] \;\le\; \frac{1}{N}\,\|f\|_\infty^2\, d_2(P\|Q). \tag{12}$$

Compared to a Monte Carlo estimator using N samples drawn directly from P, the IS variance is inflated by the factor $d_2(P\|Q)$. As P drifts away from Q, this factor grows and the estimator effectively uses fewer samples. Similar results for supervised learning have been shown in [2].

Fisher expansion (Theorem 4.2 of Metelli et al. [10]). For parametric families $\{p_\omega : \omega\in\Omega\subseteq\mathbb{R}^p\}$ with differentiable density and Fisher information matrix $F(\omega)=\mathbb{E}_{x\sim p_\omega}[\nabla\log p_\omega\,\nabla\log p_\omega^\top]$,

$$D_\alpha(p_{\omega'}\|p_\omega) \;=\; \frac{\alpha}{2}\,(\omega'-\omega)^\top F(\omega)\,(\omega'-\omega) \;+\; o(\|\omega'-\omega\|_2^2), \tag{13}$$

and consequently

$$d_2(p_{\omega'}\|p_\omega) \;\approx\; 1+(\omega'-\omega)^\top F(\omega)\,(\omega'-\omega).$$

For our setting this says that when the trained policy π_{θ_t} stays close to the behavior policy π_{θ_0} in parameter space, the variance amplification of the IS-corrected estimator is, to leading order, quadratic in $\theta_t-\theta_0$ in the Fisher metric.

D.2. Theorem proof

We begin by stating the assumptions. Next we restate the theorem and provide the proof.

Small-drift assumptions. We assume a small-drift regime:

- **(A1)** $\mathrm{KL}(\pi_{\theta_t}\|\pi_{\theta_0})\ll 1$ for all $t<T$, ensuring that the parameter displacement remains in the regime where the second-order Fisher expansion of the Rényi divergence as in Eq. 13 is accurate.
- **(A2)** $G(\theta)$, $H(\theta)$, $\Sigma_q(\theta)$, $\Sigma_{o|q}(\theta)$ are approximately constant in θ over the inner loop. This is motivated by A1: small policy drift suggests these quantities change slowly, but this is assumed independently, particularly for the second-order terms H and $\Sigma_{o|q}$.
- **(A3)** The objective is used without clipping. Following [4], when clipping is present, we treat the proximal-policy role of θ_0 (controlling the update size) as orthogonal to its behavior-policy role (determining w_t); the noise analysis here concerns only the latter.

Theorem 2 (Off-Policy CBS) *Under assumptions (A1)–(A3), and with Σ_q, $\Sigma_{o|q}$, σ^2_{inter}, σ^2_{intra} as in Theorem 1 (which the off-policy quantities reduce to under (A2)), the critical batch size at inner step t is:*

$$\mathcal{B}_{critical}{}^{(t)}(K) \;=\; \sigma^2_{inter} \;+\; \big[1 + (t\kappa)^2\big]\,\frac{\sigma^2_{intra}}{K}, \tag{2}$$

where $\kappa^2 := \epsilon^2\, G^\top \bar{F}\, G \;\approx\; 2\,\mathrm{KL}_{\text{per inner step}}$, with $\bar{F} = \mathbb{E}_q[F_q(\theta_0)]$ the prompt-averaged Fisher information and ϵ the learning rate.

Proof One outer iteration of off-policy GRPO consists of: (i) collecting a dataset of B prompts $q_{1:B} \sim \rho$ with K rollouts each, $o_{i,1:K} \sim \pi_{\theta_0}(\cdot \mid q_i)$, where θ_0 is the behavior policy fixed for the iteration; (ii) splitting the dataset into M minibatches of $b := B/M$ prompts, where each minibatch contains all K rollouts for its prompts; and (iii) sweeping E epochs over them, for a total of $T := ME$ inner gradient steps. We index the inner steps by $t \in \{0, 1, \ldots, T-1\}$, with parameter θ_t at step t and θ_0 shared with the behavior policy.

Unlike the on-policy case, the importance ratios in $J_m(\theta)$ (Eq. 11) are no longer identically 1 at inner step $t > 0$, since $\theta_t \neq \theta_0$. Under (A3), clipping is removed, and the surrogate objective for a minibatch at inner step t is

$$J_t(\theta;\, \theta_0,\, \mathcal{D}_t) \;=\; \frac{1}{b}\sum_{i \in \mathrm{mb}_t} \frac{1}{K}\sum_{k=1}^{K} w_t(q_i, o_{i,k})\, \hat{A}_k(o_{i,1:K})\, \frac{\pi_\theta(o_{i,k} \mid q_i)}{\pi_{\theta_0}(o_{i,k} \mid q_i)},$$

with importance weight $w_t(q, o) := \pi_{\theta_t}(o \mid q)/\pi_{\theta_0}(o \mid q)$. For a given minibatch, this is a fixed function of θ. The gradient at $\theta = \theta_t$ is

$$\hat{G}_t \;=\; \frac{1}{b}\sum_{i \in \mathrm{mb}_t} \hat{g}_t(q_i; o_{i,1:K}), \qquad \hat{g}_t(q; o_{1:K}) \;=\; \frac{1}{K}\sum_{k=1}^{K} w_t(q, o_k)\, \hat{A}_k(o_{1:K})\, \nabla_\theta \log \pi_{\theta_t}(o_k \mid q).$$

Mean and covariance of $\hat{G}_t$. The IS identity gives $\mathbb{E}_{o|q,\pi_{\theta_0}}[w_t f] = \mathbb{E}_{o|q,\pi_{\theta_t}}[f]$, hence $\mathbb{E}_{\theta_0}[\hat{G}_t] = G(\theta_t)$. Applying the law of total variance over prompts and rollouts as in the on-policy case,

$$\mathrm{Cov}_{\theta_0}(\hat{G}_t) \;=\; \frac{1}{b}\,\Sigma_q^{(t,0)} \;+\; \frac{1}{bK}\,\Sigma_{o|q}^{(t,0)},$$

with

$$\Sigma_q^{(t,0)} \;:=\; \mathrm{Var}_q\!\left(\mathop{\mathbb{E}}_{o|q,\pi_{\theta_0}}[\hat{g}_t(q;\, o_{1:K})]\right), \quad \Sigma_{o|q}^{(t,0)} \;:=\; \mathop{\mathbb{E}}_{q}\Big[\mathrm{Var}_{o|q,\pi_{\theta_0}}\big(\hat{g}_t(q;\, o_{1:K})\big)\Big].$$

The inner expectation in $\Sigma_q^{(t,0)}$ collapses, by the IS identity, to the on-policy per-prompt gradient at θ_t, giving $\Sigma_q^{(t,0)} = \Sigma_q(\theta_t) \approx \Sigma_q(\theta_0)$ under (A2). The inter-prompt term is therefore not drift-inflated. The intra-prompt term, in contrast, picks up the Rényi inflation: applying (12) prompt-by-prompt and substituting (13),

$$\mathrm{tr}(H\,\Sigma_{o|q}^{(t,0)}) \;\approx\; \rho(t)\;\mathrm{tr}(H\,\Sigma_{o|q}(\theta_0)), \qquad \rho(t) \;:=\; \mathop{\mathbb{E}}_{q}[d_2(\pi_{\theta_t}(\cdot|q) \,\|\, \pi_{\theta_0}(\cdot|q))]\,.$$

Noise scale. Plugging this into the McCandlish second-order Taylor expansion of $\mathbb{E}[J_t(\theta_t - \epsilon\hat{G}_t;\ \theta_0,\ \mathcal{D}_t)]$ and minimizing over ϵ yields the same functional form as on-policy, with the intra-prompt noise term replaced by its drift-inflated counterpart:

$$\mathcal{B}_{\text{noise}}{}^{(t)} \;=\; \sigma^2_{\text{inter}} \;+\; \rho(t)\,\frac{\sigma^2_{\text{intra}}}{K}, \tag{14}$$

giving

$$\epsilon_{\text{opt}}{}^{(t)} \;=\; \frac{\epsilon_{\max}}{1 + \mathcal{B}_{\text{noise}}{}^{(t)}/b}, \qquad \mathcal{L}_{\text{opt}}{}^{(t)} \;=\; \frac{\mathcal{L}_{\max}}{1 + \mathcal{B}_{\text{noise}}{}^{(t)}/b}.$$

The drift inflation $\rho(t)$ and the scale κ. The factor $\rho(t)$ admits a closed form under a mean-trajectory approximation. To leading order, the small-drift dynamics give $\mathbb{E}[\theta_t - \theta_0] \approx -t\,\epsilon\,G(\theta_0)$. We approximate $\rho(t)$ by evaluating the Fisher quadratic form at this mean trajectory rather than taking its expectation over the stochastic path; this is accurate when minibatches are large enough that θ_t concentrates around its mean. By the Fisher expansion (13),

$$\rho(t) \;\approx\; 1 + t^2\epsilon^2\, G^\top \bar{F}\, G \;=\; 1 + (t\kappa)^2, \qquad \kappa^2 \;:=\; \epsilon^2\, G^\top \bar{F}\, G, \tag{15}$$

where $\bar{F} := \mathbb{E}_q[F_q(\theta_0)]$ is the prompt-averaged Fisher information. The quantity κ has a directly measurable interpretation. The KL divergence between adjacent inner-step policies admits the same Fisher expansion:

$$\mathrm{KL}(\pi_{\theta_1} \| \pi_{\theta_0}) \;\approx\; \tfrac{1}{2}\,(\Delta\theta)^\top F\,\Delta\theta \;=\; \tfrac{1}{2}\,\kappa^2,$$

hence $\kappa^2 \approx 2\,\mathrm{KL}_{\text{per inner step}}$.

Off-policy critical batch size. Substituting (15) into (14),

$$\mathcal{B}_{\text{critical}}{}^{(t)} \;:=\; \mathcal{B}_{\text{noise}}{}^{(t)} \;=\; \sigma^2_{\text{inter}} + \big[1 + (t\kappa)^2\big]\,\frac{\sigma^2_{\text{intra}}}{K}. \tag{16}$$

At $t = 0$ this recovers the on-policy CBS exactly. For $t > 0$, only the intra-prompt term is inflated, and the inflation is quadratic in the cumulative drift $t\kappa$. ■

Appendix E. Full empirical analysis and details for On-Policy GRPO

We evaluate the on-policy noise decomposition $B_{\text{crit}} = \sigma^2_{\text{inter}} + \sigma^2_{\text{intra}}/K$ (equivalently, the rollout-batch CBS $N_{\text{crit}} = K\sigma^2_{\text{inter}} + \sigma^2_{\text{intra}}$ where $N = BK$) against runs of GRPO on AIME-1983-2024 with the Qwen2.5 base model. Performance is tracked via mean@32 on the evaluation set. For each run with prompt batch B and rollouts-per-prompt K, we record the number of outer iterations S required to first reach mean@32 ≥ 0.5; this defines the empirical steps-to-target curve $S(B, K)$ from which CBS is fitted.

E.1. Experiments

We conduct three experiment sweeps, each holding one of $\{B, K\}$ fixed and varying the other. The inner gradient batch in each case is $B \times K$ rollouts (single, fully on-policy gradient step per outer iteration).

- **K16**: $K = 16$ fixed; $B \in \{8, 16, 32, 64, 128, 256, 512, 1024, 2048\}$.
- **K64**: $K = 64$ fixed; $B \in \{2, 4, 8, 16, 32, 64, 128, 256, 512, 1024\}$.
- **B128**: $B = 128$ fixed; $K \in \{2, 4, 8, 16, 32\}$.

Together these provide 24 (B, K, S) measurement triples. The first two probe the B-axis of the noise decomposition at two different values of K; the third probes the K-axis at fixed B and provides an independent identification of σ^2_{intra}.

E.2. Hyperparameters

Hyperparameter	$K = 16 \mid 64$	$B = 128$
Train Prompt Batch Size (`train_prompt_bsz`)	*	128
Responses per Prompt (`n_resp_per_prompt`)	(16 | 64)	*
PPO Mini Batch Size (`train_prompt_mini_bsz`)	* × (16 | 64)	128 × *
Advantage Estimator (`adv_estimator`)	`grpo`	`grpo`
Use KL in Reward (`use_kl_in_reward`)	False	False
KL Coefficient (`kl_coef`)	0.0	0.0
Use KL Loss (`use_kl_loss`)	True	True
KL Loss Coefficient (`kl_loss_coef`)	0.001	0.001
Clip Ratio Low (`clip_ratio_low`)	0.2	0.2
Clip Ratio High (`clip_ratio_high`)	0.2	0.2
Clip Ratio C (`clip_ratio_c`)	3.0	3.0
Loss Aggregation Mode (`loss_agg_mode`)	`token-mean`	`token-mean`
Learning Rate (`optim.lr`)	1e-6	1e-6
Entropy Coefficient (`entropy_coeff`)	0	0
Gradient Clip (`grad_clip`)	1.0	1.0
Validation Temperature (`val_kwargs.temperature`)	1.0	0.0

Table 1: Key hyperparameters for on-policy experiments

E.3. Results

Figure 6 shows the raw mean@32 training trajectories for the three on-policy sweeps with the ScaleRL sigmoid [7] fitted per batch and overlaid as solid curves; the fitted asymptotic pass rate A from Eq. 1 of Khatri et al. [7] is reported in each legend. The fits faithfully track the data across the full $\sim 10^4$-step range, with the smaller batches reaching higher A once the ScaleRL extrapolation is taken into account. Trajectories that exhibit late-training collapse (sustained drop >0.15 from running max) are truncated before fitting.

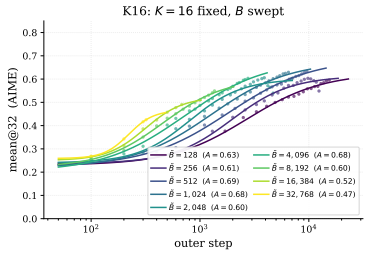


Figure 3: K16

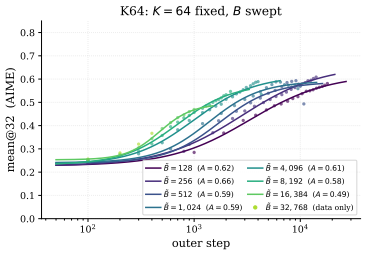


Figure 4: K64

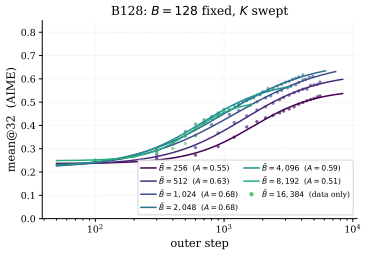


Figure 5: B128

Figure 6: Training trajectories (mean@32 vs. outer step) for the three on-policy sweeps. Markers are raw evaluations; solid curves are per-batch ScaleRL sigmoid fits $R_C - R_0 = (A - R_0)/(1 + (C_{\text{mid}}/C)^B)$ from Khatri et al. [7], extrapolated $1.5\times$ beyond the observed range. Asymptotic pass rate A is in the legend. Larger batches converge in fewer outer steps but, under matched compute, hit a lower A ceiling.

For each experiment we fit the hyperbolic form $S(x) = a + b/x$ to the swept axis $x \in \{B, K\}$ and extract a critical value $x^* = b/a$. Translation to the rollout-batch CBS $N^* = B^*K$ (or $N^* = BK^*$) yields three independent estimates that should agree if the decomposition is correct. Figure 10 shows the per-experiment hyperbolic fits, and Table 2 summarizes the extracted critical values.

Experiment	**Swept axis**	$S_{\min}$	B^* **(prompts)**	K^* **(rollouts/prompt)**	$N^* = B^*K$ **or** BK^* **(rollouts)**
K16	B at $K = 16$	1,228	26.1	—	417
K64	B at $K = 64$	1,790	6.2	—	394
B128	K at $B = 128$	1,075	—	5.6	720

Table 2: Per-experiment CBS fits at the target mean@32 ≥ 0.5. Translated to rollout units, the three experiments give consistent values clustered around $N^* \approx 484$ (geometric mean), supporting the prediction that N^* is approximately K-invariant when $K\sigma^2_{\text{inter}} \ll \sigma^2_{\text{intra}}$.

To extract $(\sigma^2_{\text{inter}}, \sigma^2_{\text{intra}})$ jointly, we fit all eighteen data points to the three-parameter form

$$S(B, K) \;=\; S_{min} \cdot \left(1 \;+\; \frac{\sigma^2_{\text{inter}}}{B} \;+\; \frac{\sigma^2_{\text{intra}}}{BK}\right), \tag{17}$$

yielding

$$\sigma^2_{\text{inter}} \approx 3.2, \qquad \sigma^2_{\text{intra}} \approx 311, \qquad S_{min} \approx 1465, \tag{18}$$

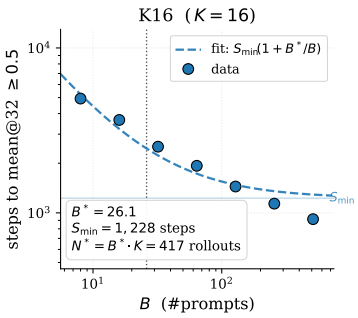


Figure 7: K16: $K = 16$ fixed, B swept.

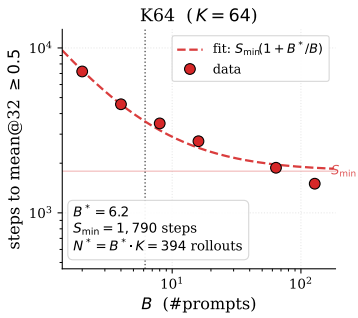


Figure 8: K64: $K = 64$ fixed, B swept.

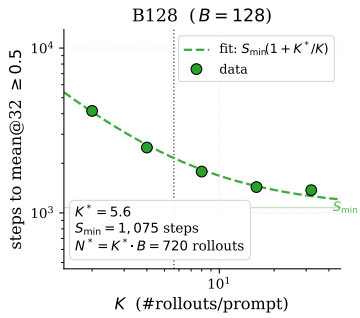


Figure 9: B128: $B = 128$ fixed, K swept.

Figure 10: McCandlish-style fits of steps-to-target S versus the swept axis for the three on-policy experiments. Each panel fits the hyperbolic form $S(x) = S_{\min}(1 + x^*/x)$ to the data and extracts the critical value $x^* \in \{B^*, K^*\}$ at which the noise term equals the signal term. Translated to rollout units, the three independent experiments yield $N^* \in \{417, 394, 720\}$ with geometric mean ≈ 484, confirming that N^* is approximately K-invariant within the regime where $K\sigma^2_{\text{inter}} \ll \sigma^2_{\text{intra}}$. The hyperbolic form fits the data tightly across both axes.

with mean relative residual of 15% across the data, dominated by the B128 sweep, where the joint model under-predicts K^* by roughly $2\times$ relative to the per-experiment B128 hyperbola fit; the K16 and K64 sweeps alone are reproduced to within ± 4. The intra-prompt term dominates inter-prompt.

E.4. Interpretation

Inter-prompt variance is small but nonzero. The group-relative advantage normalization in GRPO subtracts the prompt-conditional reward mean, eliminating across-prompt variance in reward scale. The residual $\sigma^2_{\text{inter}} \approx 3$ measures variation in the prompt-conditional gradient *direction*, which the per-prompt baseline does not address. That this residual is two orders of magnitude smaller than σ^2_{intra} confirms the design intent of GRPO's normalization works as expected.

Intra-prompt variance is the dominant source of gradient noise. The estimate $\sigma^2_{\text{intra}} \approx 311$ has a direct physical reading: it is the average noise-to-signal ratio of a single rollout's gradient contribution. Roughly σ^2_{intra} rollouts are required for the policy-gradient direction to emerge from per-rollout noise. This number is set by the product of (i) reward variance per prompt $\mathbb{E}_o[A^2 \mid q]$, which is maximized near the success-rate boundary; (ii) score-function magnitude $\mathbb{E}_o[\|\nabla \log \pi\|^2]$, which scales with trajectory length and policy entropy; and (iii) the inverse of the squared population gradient $|G|^{-2}$.

The decomposition implies a regime-balance threshold. Define $K_{\text{balance}} = \sigma^2_{\text{intra}}/\sigma^2_{\text{inter}} \approx 100$, the value of K at which the inter and intra contributions to N_{crit} are equal. For $K \ll K_{\text{balance}}$, the noise is overwhelmingly intra-dominated and the rollout-batch CBS $N^* \approx \sigma^2_{\text{intra}}$ is approximately

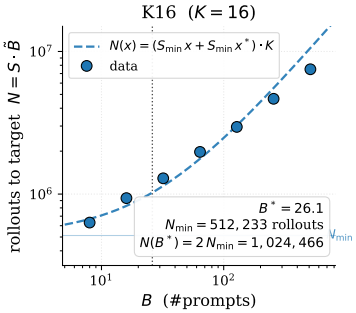


Figure 11: K16

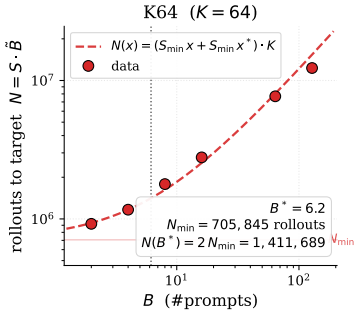


Figure 12: K64

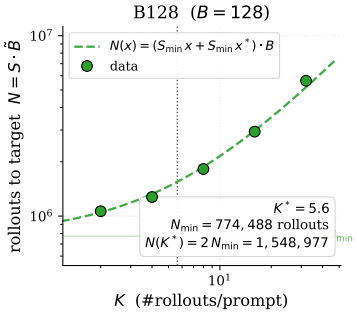


Figure 13: B128

Figure 14: Pareto frontiers in rollout units: total rollouts $N = S \cdot \tilde{B}$ to reach mean@32 ≥ 0.5 vs. the swept axis. The McCandlish hyperbola $S(x) = S_{\min}(1 + x^*/x)$ implies $N(x) = S_{\min} \cdot \texttt{fixed} \cdot (x + x^*)$, linear in x with floor $N_{\min} = S_{\min}\, x^* \cdot \texttt{fixed}$ at $x \to 0$ and slope $S_{\min}\, \texttt{fixed}$ at $x \to \infty$. The knee at $x = x^*$ doubles the floor: every batch beyond the critical value buys parallelism at strictly increasing rollout cost. The three experiments yield $N_{\min} \in \{0.51, 0.71, 0.77\} \times 10^6$ rollouts, again consistent within a factor of 1.5.

K-invariant: only the total rollout count $N = BK$ matters for gradient noise, regardless of how it is partitioned. For $K \gtrsim K_{\text{balance}}$, the $\sigma^2_{\text{inter}}/B$ term becomes binding and additional K gives diminishing returns.

There is an irreducible floor on B. As $K \to \infty$, the per-step variance approaches $\sigma^2_{\text{inter}}/B$, requiring $B \gtrsim \sigma^2_{\text{inter}} \approx 3$ to keep total noise below unity regardless of K. Practically, $B \gtrsim 2\sigma^2_{\text{inter}} \approx 10$ is needed to keep the inter-prompt term well below the intra term at typical K.

E.5. Insights

The on-policy analysis yields several actionable takeaways.

1. **The relevant batch dimension is $N = BK$, not B.** For all $K \lesssim 100$ (which covers standard practice), $N^* \approx \sigma^2_{\text{intra}} \approx 311$ is the only number that matters for gradient noise. The split of N between B and K is a hardware-utilization choice, not a statistical-efficiency one.

2. **There is a hard lower bound on B.** Even with infinite K, $B \gtrsim \sigma^2_{\text{inter}} \approx 3$ is required for a trustworthy gradient direction. Practically $B \gtrsim 10$ is safe.

3. **$K_{\text{balance}} \approx 100$ marks where K-scaling stops being free.** Standard practice ($K \in \{16, 64\}$) sits comfortably below this; pushing K to ≥ 256 requires growing B proportionally to avoid an inter-prompt penalty.

4. **$\sigma^2_{\text{intra}} \approx 311$ is a problem-property, not a hyperparameter.** It depends on reward variance, trajectory length, and policy entropy, all of which evolve during training. CBS should therefore drift over training, and is expected to be larger for longer-trajectory or higher-entropy tasks.

5. **The group-relative baseline performs as designed.** The empirical $\sigma^2_{\text{inter}} \approx 3$ is small enough that, in this regime, GRPO's per-prompt normalization is not the bottleneck. Improvements to gradient estimation should target intra-prompt noise (e.g., via importance-weighted advantages, control variates within rollouts) rather than the inter-prompt baseline.

Appendix F. Full empirical analysis for Off-policy GRPO

We test the off-policy drift-inflated noise scale $N_{\text{crit}}^{(t)} = K\sigma_{\text{inter}}^2 + \rho(t)\sigma_{\text{intra}}^2$ with $\rho(t) \approx 1 + (t\kappa)^2$, and the consequent saturation prediction $N^* \gtrsim b/\kappa$ in rollout units, against runs that take multiple gradient steps per data collection. Setup matches the on-policy experiments (Qwen2.5 base, AIME-1983-2024, mean@32). Each outer iteration collects $B \times K$ rollouts at the behavior policy and processes them in mini-batches of b samples for $T = BK/b$ inner gradient steps.

F.1. Experiments

We conduct two experiment sweeps, both with $K = 16$ rollouts per prompt and $b = 128$ samples per mini-batch (so $T = B/8$ inner steps per outer iteration), differing only in the PPO clipping range ϵ:

- **Clip**: standard PPO clipping, $\epsilon = 0.2$. $B \in \{8, 16, 32, 64, 128, 256, 512, 1024\}$.
- **NoClip**: effectively no clipping, $\epsilon = 10$. $B \in \{32, 64, 128, 256, 512, 1024\}$.

For $B = 8$, $T = 1$ and the run is identical to on-policy GRPO; for $B \geq 16$, multiple inner steps are taken on each batch and drift accumulates. We extract the per-inner-step drift κ directly from the running KL-to-behavior diagnostic via $\kappa^2/2 \approx$ `actor/kl_loss`$/T$, an independent measurement that does not rely on CBS fitting.

F.2. Hyperparameters

Hyperparameter	Clip Ratio = 0.2	Clip Ratio = 10.0
Train Prompt Batch Size (`train_prompt_bsz`)	*	*
Responses per Prompt (`n_resp_per_prompt`)	16	16
PPO Mini Batch Size (`train_prompt_mini_bsz`)	64	64
Advantage Estimator (`adv_estimator`)	`grpo`	`grpo`
Use KL in Reward (`use_kl_in_reward`)	False	False
KL Coefficient (`kl_coef`)	0.0	0.0
Use KL Loss (`use_kl_loss`)	True	True
KL Loss Coefficient (`kl_loss_coef`)	0.001	0.001
Clip Ratio Low (`clip_ratio_low`)	0.2	10.0
Clip Ratio High (`clip_ratio_high`)	0.2	10.0
Clip Ratio C (`clip_ratio_c`)	3.0	3.0
Loss Aggregation Mode (`loss_agg_mode`)	`token-mean`	`token-mean`
Learning Rate (`optim.lr`)	1e-6	1e-6
Entropy Coefficient (`entropy_coeff`)	0	0
Gradient Clip (`grad_clip`)	1.0	1.0
Validation Temperature (`val_kwargs.temperature`)	1.0	1.0

Table 3: Key hyperparameters for off-policy experiments

F.3. Results

Figure 17 shows the mean@32 training trajectories for the two off-policy sweeps, with per-batch ScaleRL sigmoid fits [7] overlaid as solid curves. Two qualitative differences from on-policy are immediately visible: (i) at matched $\tilde{B}$, the curves are vertically shifted up across the full step range — i.e., off-policy reuse of rollouts amortizes data collection across $T = \tilde{B}/128$ inner steps and the curves reach any given mean@32 at roughly $T\times$ fewer outer steps; (ii) the NoClip ($\epsilon = 10$) asymptotes are systematically higher than Clip ($\epsilon = 0.2$), with A reaching 0.86 for NoClip vs. $\leq$0.70 for Clip — independent of CBS, this corroborates the clipping-bias interpretation of the final-accuracy gap.

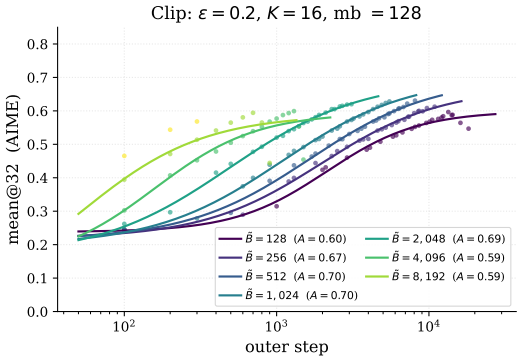


Figure 15: Clip ($\epsilon = 0.2$)

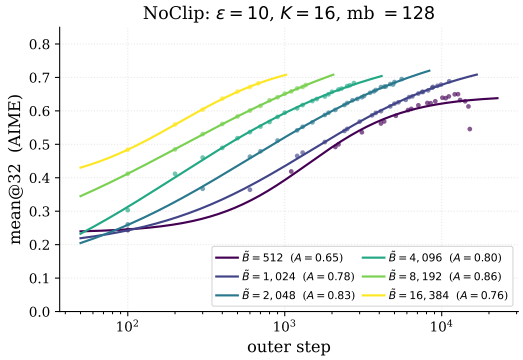


Figure 16: NoClip ($\epsilon = 10$)

Figure 17: Training trajectories for the off-policy sweeps. Markers are raw evaluations; solid curves are per-batch ScaleRL sigmoid fits from Khatri et al. [7], extrapolated 1.5$\times$ beyond observed step counts. Asymptotic pass rate A is reported in each legend. The NoClip recipe reaches strictly higher A at every $\tilde{B}$, separate from any CBS-related effect.

Figure 20 shows the McCandlish steps-to-target hyperbola for each off-policy sweep. Both settings yield essentially the same fit, with $N^* \approx 5\,400$–$5\,900$ rollouts when restricted to the $B \geq 32$ regime; at matched B, Clip and NoClip have indistinguishable $S \cdot N$ at the target metric, with the difference between them appearing only in the asymptotic ceiling and not in the linear-regime CBS. Table 4 summarizes per-experiment drift measurements and CBS-related quantities.

The key empirical observation is that κ measured per-inner-step is approximately B-independent within each setting, varying within a factor of ~ 1.5 across a 64-fold range of B. The two settings yield nearly identical per-step drift, indicating that PPO clipping at $\epsilon = 0.2$ does not meaningfully reduce drift in this regime; the empirical clipping fraction `actor/pg_clipfrac` remains near zero throughout, so the clipped and unclipped objectives produce nearly identical gradients.

For comparison with the on-policy CBS of ≈ 311 rollouts, both off-policy settings extend the linear-scaling regime by approximately 18$\times$ in rollout units, indicating a substantial off-policy advantage in data-parallel scalability.

F.4. Interpretation

The experiments are consistent with three structural predictions of the off-policy model.

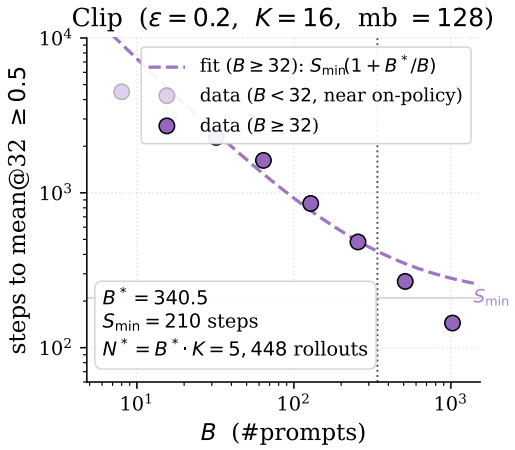


Figure 18: Clip: $\epsilon = 0.2$.

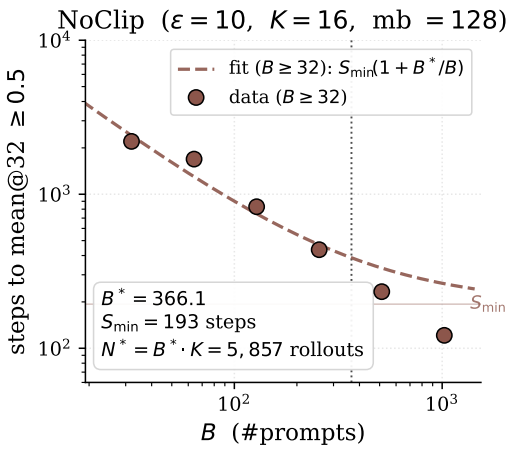


Figure 19: NoClip: $\epsilon = 10$.

Figure 20: McCandlish-style fits of steps-to-target S versus prompt batch B for the two off-policy sweeps, target mean@32 ≥ 0.5. Hyperbolic fit $S(B) = S_{\min}(1 + B^*/B)$ is restricted to $B \geq 32$ (the regime where $T = B/8 \geq 4$ inner steps per outer iteration are taken; the $B = 8, 16$ Clip points sit closer to the on-policy regime and are shown in grey but not used in the fit). Translated to rollout units, the two settings give $N^* \approx 5\,448$ (Clip) and $N^* \approx 5\,857$ (NoClip) — within 8% of each other and in both cases $\sim 18\times$ the on-policy CBS (≈ 311), a genuine off-policy parallelism advantage.

Experiment	κ **(median,** $B \geq 16$**)**	b/κ **(projected** N^***)**	**Empirical** N^* **(B≥32 fit)**	**Final mean@32**
Clip	0.080	$\approx 1\,600$ rollouts	$\approx 5\,448$ rollouts	≈ 0.63
NoClip	0.088	$\approx 1\,455$ rollouts	$\approx 5\,857$ rollouts	≈ 0.70

Table 4: Off-policy CBS estimates. "Projected N^*" is the conservative lower bound b/κ from the Rényi-bounded variance argument. "Empirical N^*" is from a hyperbolic fit on $S(B)$ at target mean@32 ≥ 0.5, restricted to $B \geq 32$ (the regime where $T \geq 4$ inner steps are taken). The two settings yield nearly identical empirical N^* — within 8% of each other — confirming that PPO clipping at $\epsilon = 0.2$ does not shift the linear-regime CBS in this drift regime. The final-accuracy ceiling differs meaningfully and independently of CBS.

κ is approximately constant in B. The theory predicts $\kappa^2 \approx \epsilon_{\text{lr}}^2 \cdot G^\top \bar{F} G$ depends on the learning rate and Fisher curvature but not on B. Empirically, κ is stable within a factor of ~ 1.5 across the full sweep in both settings, with a mild upward drift at very large T in Clip ($\kappa = 0.074$ at $T = 2$ rising to $\kappa = 0.112$ at $T = 128$) consistent with second-order compounding in the Fisher expansion. This validates the use of κ as a hyperparameter-level constant for predicting safe parallelism.

κ is insensitive to ϵ when clipping is rarely active. The clipping fraction is negligible throughout training in both Clip and NoClip. The theoretical drift-suppression effect of clipping requires the

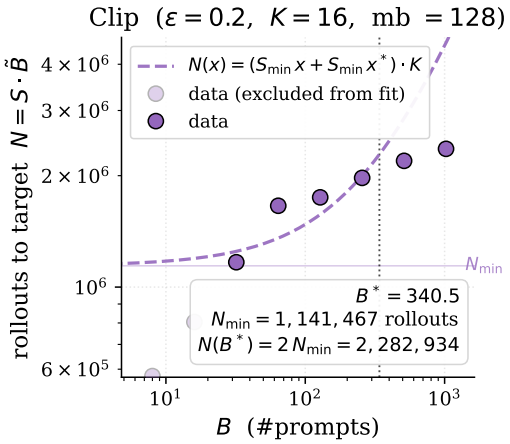


Figure 21: Clip ($\epsilon = 0.2$).

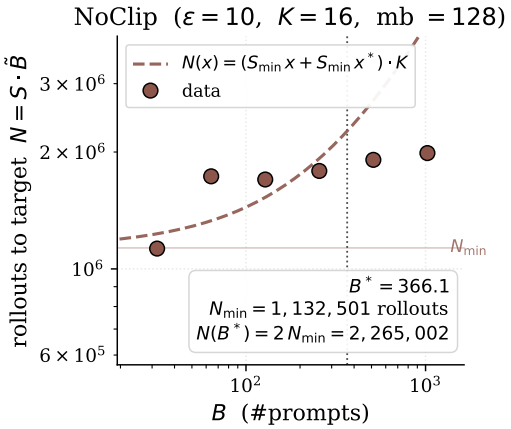


Figure 22: NoClip ($\epsilon = 10$).

Figure 23: Pareto frontiers in rollout units for the off-policy sweeps: total rollouts $N = S \cdot \tilde{B}$ to reach mean@32 ≥ 0.5 vs. prompt batch B. Hyperbolic fit restricted to $B \geq 32$ as in Table 4; the $B = 8, 16$ Clip points (greyed) lie in the near-on-policy regime ($T \leq 2$). Both settings yield $N_{\min} \approx 1.14 \times 10^6$ rollouts — a roughly $2\times$ floor relative to on-policy K16 — and reach $2N_{\min}$ at the knee $B^* \approx 350$.

clip to actually bind, which it does not in this regime, so the two objectives behave identically at the gradient level. This explains why κ matches between the two settings to within 10% (0.080 vs. 0.088).

Off-policy CBS in rollout units is strictly larger than on-policy. On-policy CBS sits at $N^* \approx \sigma^2_{\text{intra}} \approx 311$ rollouts; off-policy CBS sits at $N^* \approx 5\,400$–$5\,900$ rollouts. This is the genuine parallelism advantage: each batch of rollouts can support T inner gradient steps before drift saturates, amortizing the data collection cost across multiple updates and extending the McCandlish linear regime accordingly.

The theoretical lower bound $N^* \geq b/\kappa \approx 1\,500$ rollouts is satisfied with margin, but empirical CBS exceeds it by a factor of ≈ 3.5–$4\times$. This is consistent with the Rényi d_2 bound being conservative in practice: importance ratios stay close to 1 (clipfrac ≈ 0 throughout), and the actual variance inflation $\rho(t)$ is much smaller than the worst-case $1 + (t\kappa)^2$ bound. The b/κ prediction should be read as a sufficient condition for being in the off-policy linear regime, not a tight estimate of saturation.

F.5. Insights

The off-policy analysis yields the following operational and conceptual conclusions.

1. κ **is directly measurable during training.** The diagnostic $\kappa \approx \sqrt{2 \cdot \texttt{actor/kl\_loss}/T}$ is computable at every step from quantities already logged by standard PPO/GRPO implementations. It provides a direct estimator of per-inner-step drift in Fisher units, and serves as an early-warning signal when scaling B beyond the safe regime.

2. $N^* \geq b/\kappa$ **is the practical lower bound on usable parallelism.** Empirically the safe regime extends several times further than this bound, so b/κ is conservative. In conjunction with monitoring `pg_clipfrac`, this gives a practitioner-friendly recipe for setting batch sizes in off-policy GRPO without requiring CBS to be fit empirically.

3. **The off-policy parallelism gain is** $\sim 18\times$ **in rollout units.** On-policy CBS of ~ 311 rollouts grows to $\sim$ 5 400–5 900 rollouts in the off-policy regime, set by the b/κ ratio with a multiplicative slack of ≈ 3.5–$4\times$ reflecting Rényi bound looseness. This is the genuine compute-parallelism advantage of off-policy GRPO at fixed problem scale.

4. **PPO clipping does not protect against drift in this regime.** The clipping fraction is negligible throughout training, so the clipped objective behaves identically to the unclipped one at the gradient level. The variance-reduction story behind PPO clipping is irrelevant for the noise-scale analysis here.

5. **Removing PPO clipping (**$\epsilon = 10$**) strictly improves final accuracy in this regime.** Both CBS and final accuracy are at least as good as with clipping, suggesting that the bias introduced by clipping — not the variance protection it provides — is the binding constraint on accuracy ceiling.