Baseline Estimation Bounds For GRPO

Yoav Kor

July 2026

1 Introduction

GRPO is a widely used algorithm to post train LLMs. while its proven to work, a notable problem is the credit assigment problem. Specifically for dataset X and a prompt $p \sim X$ GRPO sample G completion per round and estimate the Advantage baseline as $\frac{1}{G} \cdot 1[\text{rollout}_g = \text{correct}]$ for $g \in \{1, \ldots, G\}$ While it has enormous success and is generally consider pretty stable, GRPO assign the same advantage for every token (i.e. reward-baseline). As already known from control variate theory, the optimal baseline for variance reduction is usually estimated with $V^{\pi}(s)$ for every state s in a trajectory. GRPO estimates this at every position in a trajectory with the base level of the question, i.e. the expected success of the policy of solving the question. This of course, does not change between different states of the trajectory (as the baseline is projected to every token). This causes might cause high gradient estimator variance (with respect to the sampled trajectory). Then, the updates during RL session becomes unstable. The question is if we could do better, i.e. find a better baseline estimation than GRPO's, such that each state recieve a fine grained signal and the gradient estimator becomes less noisy.

2 Deriving upper bound on the improvment

First, we need to ask ourselves whether we can improve over the GRPO-based-gradient-estimator variance. In this section we will focus on reducing variance through a better baseline estimation, while leaving the token level reward constant, as in the regular GRPO framework, which for a question $x \sim Q$ and a policy trajectory y assign a token y_t, the final trajectory reward $r(x, y)$.

2.1 A2C Policy Gradient Variance

Let $X \sim D$ be a sampled prompt and let $Y = (Y_1, \ldots, Y_T) \sim \pi_\theta(\cdot \mid X)$ be a complete autoregressively sampled response. At token position t, define the state as the prompt and generated prefix,

$$S_t = (X, Y_{<t}),$$

and define

$$\tau_{t:} = (Y_t, Y_{t+1}, \ldots, Y_T)$$

as the trajectory suffix from position t onward. Thus, $\tau_{t:} \sim \pi_\theta(\cdot \mid S_t)$ means sampling the current token and all subsequent tokens autoregressively, conditional on the prefix S_t. The policy-gradient estimator is

$$\mathbb{E}_{X\sim D,\, Y\sim\pi_\theta(\cdot|X)}\left[\frac{1}{T}\sum_{t=1}^{T}\nabla_\theta \log \pi_\theta(Y_t \mid S_t)\big(G_t - b(S_t)\big)\right], \qquad |Y| = T.$$

In practice, we estimate this expectation using a finite Monte Carlo sample. This introduces variance because only finitely many questions and trajectories per question can be sampled at each gradient step.

The notation $S_t \sim d_t^\pi$ denotes the state distribution induced by first sampling $X \sim D$ and then sampling the prefix $Y_{<t}$ from π_θ. The return G_t is determined by the sampled suffix $\tau_{t:}$ and any additional reward or environment randomness.

To identify which part of this variance a baseline can reduce, define the per-token gradient contribution

$$Z_t = \nabla_\theta \log \pi_\theta(Y_t \mid S_t)\big(G_t - b(S_t)\big).$$

Provided that $b(S_t)$ does not depend on the sampled action Y_t, subtracting it does not change the expected policy gradient. By the law of total variance,

$$\begin{aligned}&\operatorname{Var}_{\substack{S_t\sim d_t^\pi,\\ \tau_{t:}\sim\pi_\theta(\cdot|S_t)}}(Z_t)\\ &\quad= \mathbb{E}_{S_t\sim d_t^\pi}\left[\operatorname{Var}_{\tau_{t:}\sim\pi_\theta(\cdot|S_t)}(Z_t \mid S_t)\right] + \operatorname{Var}_{S_t\sim d_t^\pi}\left(\mathbb{E}_{\tau_{t:}\sim\pi_\theta(\cdot|S_t)}[Z_t \mid S_t]\right).\end{aligned}$$

The baseline can reduce the first term: variance due to the sampled action and the stochastic future return, conditional on the current state. It cannot reduce the second term, which captures variation of the true expected gradient across sampled questions and states. Consequently, a better baseline cannot eliminate variance arising from prompt sampling, state visitation, or variation of the expected policy gradient across states. It also cannot generally eliminate all conditional variance, because the return need not be perfectly predictable from the state.

2.2 Optimal baseline

The value function

$$V^\pi(s) = \mathbb{E}_{\tau_{t:}\sim\pi_\theta(\cdot|s)}[G_t \mid S_t = s]$$

minimizes the unweighted return prediction error

$$\mathbb{E}_{\tau_{t:}\sim\pi_\theta(\cdot|s)}\left[(G_t - b(s))^2 \mid S_t = s\right].$$

However, the variance-reduction-optimal scalar baseline for the policy-gradient estimator is, more precisely,

$$b^*(s) = \frac{\mathbb{E}_{\tau_{t:}\sim\pi_\theta(\cdot|s)}\left[\left\|\nabla_\theta \log \pi_\theta(Y_t \mid s)\right\|^2 G_t \mid S_t = s\right]}{\mathbb{E}_{Y_t\sim\pi_\theta(\cdot|s)}\left[\left\|\nabla_\theta \log \pi_\theta(Y_t \mid s)\right\|^2 \mid S_t = s\right]}.$$

Thus, $V^\pi(s)$ is optimal only when the squared score norm is constant or conditionally uncorrelated with the return.[1] Nevertheless, it is a natural practical target.
To improve over the GRPO group-mean baseline, we therefore seek a state-dependent estimator that approaches $V^\pi(s)$ (while remembering that return MSE is only a proxy for policy-gradient variance).

2.3 Comparing baseline estimators

Define $p^\pi(x)$ as the true success rate of a policy π for a given question $x \sim D$. We can say GRPO estimates $V^\pi(x, s) = p^\pi(x)$ for every s. We assume that the group success rate approximate exactly the real success rate of the policy for a given question (i.e. if we sample G=8 per group then the sume of rewards averaged of the group size G equals the true policy's success rate on this question). Recall that in order to minimize the gradient estimator variance we want to minimize:

$$\mathbb{E}_{S_t \sim d_t^\pi(\cdot|x),\, \tau_{t:} \sim \pi_\theta(\cdot|S_t)} \left[\|\nabla_\theta \log \pi_\theta(Y_t \mid S_t)\|^2 \left(G_t - b(S_t)\right)^2 \right].$$

Also recall that we drop the magnitude of the gradient term in order to have a practial target. That is, we minimize

$$\mathbb{E}_{\tau_{t:} \sim \pi_\theta(\cdot|s)} \left[(G_t - b(s))^2 \mid S_t = s \right].$$

Therefore, we evaluate a baseline using this loss: the mean squared error between the sampled return and the baseline. We will use this loss term to rank how good a given baseline is in terms of variance reduction. The lower this value becomes, the better our baseline is. This will be termed fron now on as a given baseline's *MSE*

2.4 Optimal Baseline MSE

To provides a practical ceiling on the attainable improvement and helps determine whether pursuing a more accurate baseline is worthwhile, we need to have the optimal baseline's loss. By definition, we wont be able to do better than that with any other baseline. For the value-function baseline (which as we seen is the optimal one), and for a given state S, the MSE is

$$\text{MSE}(V(s_t)) = \mathbb{E}_{\tau_{t:} \sim \pi_\theta(\cdot|s_t)} \left[(r(\tau) - V(s_t))^2 | s_t \right] = V(s_t)(1 - V(s_t)).$$

The last term is due to the RV $r(\tau)$ being a Bernouli RV with paramter $(V(S))$. This is the best we could get for a given state. Intuition: That term is what

[1] Here, "conditionally uncorrelated" means that, after fixing the state s, the squared score norm does not systematically increase or decrease with G_t. Formally, $\text{Cov}_{\tau_{t:} \sim \pi_\theta(\cdot|s)}(\|\nabla_\theta \log \pi_\theta(Y_t \mid s)\|^2, G_t \mid S_t = s) = 0$. Under this condition, the numerator of $b^*(s)$ factorizes into $\mathbb{E}_{Y_t \sim \pi_\theta(\cdot|s)}[\|\nabla_\theta \log \pi_\theta(Y_t \mid s)\|^2]\, \mathbb{E}_{\tau_{t:} \sim \pi_\theta(\cdot|s)}[G_t]$, and hence $b^*(s) = \mathbb{E}_{\tau_{t:} \sim \pi_\theta(\cdot|s)}[G_t] = V^\pi(s)$.

left to reduce from the stochastic process of sampling from the policy. The variance (over contiuations from a give state s_t) is basically 0 at when $V(s_t)$ is approaching 0 or 1, i.e. when there is not a lot of room for the sampling process to change the probability that this question will be answered wrong/correct given the current context we have at the given time. For the entire states distribution this means the best MSE is given by:

$$\begin{aligned}
\mathrm{MSE}(V) &= \mathbb{E}_{S_t \sim d_t^\pi(\cdot|x)}\left[V(S_t)(1-V(S_t))\right] \\
&= \mathbb{E}_{S_t \sim d_t^\pi(\cdot|x)}[V(S_t)] - \mathbb{E}_{S_t \sim d_t^\pi(\cdot|x)}[V(S_t)^2] \\
&= p - \left(\mathrm{Var}_{S_t \sim d_t^\pi(\cdot|x)}[V(S_t)] + p^2\right) \\
&= p(1-p) - \mathrm{Var}_{S_t \sim d_t^\pi(\cdot|x)}[V(S_t)],
\end{aligned}$$

2.5 GRPO Baseline MSE

We now evaluate the GRPO baseline using the same loss. For a fixed question x, let

$$p = p^\pi(x) = \mathbb{P}_{\tau \sim \pi_\theta(\cdot|x)}(r(\tau) = 1)$$

and assume that the group-mean baseline recovers this probability exactly, so that $b_{\mathrm{GRPO}} = p$. Its MSE over states and continuations is

$$\begin{aligned}
\mathrm{MSE}(b_{\mathrm{GRPO}}) &= \mathbb{E}_{\substack{S_t \sim d_t^\pi(\cdot|x), \\ \tau_{t:} \sim \pi_\theta(\cdot|S_t)}}\left[(r(\tau) - p)^2\right] \\
&= \mathbb{E}_{\tau \sim \pi_\theta(\cdot|x)}\left[(r(\tau) - p)^2\right] \\
&= \mathrm{Var}_{\tau \sim \pi_\theta(\cdot|x)}[r(\tau)] \\
&= p(1-p),
\end{aligned}$$

Here, the second equality follows because sampling a prefix state $S_t \sim d_t^\pi(\cdot \mid x)$ and then sampling its continuation $\tau_{t:} \sim \pi_\theta(\cdot \mid S_t)$ induces the same distribution over complete trajectories as sampling $\tau \sim \pi_\theta(\cdot \mid x)$ directly. The last equality follows because $r(\tau)$ is Bernoulli with success probability p.

To compare this result with the value-function baseline, first note that the law of total expectation gives

$$\mathbb{E}_{S_t \sim d_t^\pi(\cdot|x)}[V(S_t)] = p.$$

We can therefore decompose the GRPO MSE algebraically as

$$\begin{aligned}
p(1-p) &= p - p^2 \\
&= \mathbb{E}_{S_t \sim d_t^\pi(\cdot|x)}[V(S_t)] - \mathbb{E}_{S_t \sim d_t^\pi(\cdot|x)}[V(S_t)]^2 \\
&= \mathbb{E}_{S_t \sim d_t^\pi(\cdot|x)}[V(S_t)] - \mathbb{E}_{S_t \sim d_t^\pi(\cdot|x)}[V(S_t)^2] \\
&\quad + \left(\mathbb{E}_{S_t \sim d_t^\pi(\cdot|x)}[V(S_t)^2] - \mathbb{E}_{S_t \sim d_t^\pi(\cdot|x)}[V(S_t)]^2\right) \\
&= \mathbb{E}_{S_t \sim d_t^\pi(\cdot|x)}[V(S_t)(1-V(S_t))] + \mathrm{Var}_{S_t \sim d_t^\pi(\cdot|x)}(V(S_t)),
\end{aligned}$$

which is the law of total variance written for the binary reward. The first term is precisely the value-baseline MSE:

$$\mathrm{MSE}(V) = \mathbb{E}_{S_t \sim d_t^\pi(\cdot|x)}\left[V(S_t)(1-V(S_t))\right].$$

It measures the uncertainty that remains after the generated prefix is known. The second term, $\mathrm{Var}_{S_t \sim d_t^\pi(\cdot|x)}[V(S_t)]$, measures how much the probability of eventual success varies across prefixes. The GRPO baseline uses only the question-level probability p and therefore cannot account for this between-prefix variation.

It follows that the maximum absolute MSE reduction obtainable by replacing the question-level GRPO baseline with the exact state-value baseline is

$$\boxed{\mathrm{MSE}(b_{\mathrm{GRPO}}) - \mathrm{MSE}(V) = \mathrm{Var}_{S_t \sim d_t^\pi(\cdot|x)}[V(S_t)]}.$$

When $0 < p < 1$, the corresponding relative reduction is

$$\boxed{\frac{\mathrm{MSE}(b_{\mathrm{GRPO}}) - \mathrm{MSE}(V)}{\mathrm{MSE}(b_{\mathrm{GRPO}})} = \frac{\mathrm{Var}_{S_t \sim d_t^\pi(\cdot|x)}[V(S_t)]}{p(1-p)}}.$$

This ratio lies in $[0, 1]$. It equals zero when the success probability is constant across prefixes, in which case the GRPO and value baselines have the same MSE. It becomes larger when the generated prefix substantially changes the probability of eventual success. Thus, in this simplified analysis, the variance of $V(S_t)$ across visited states quantifies the maximum potential benefit of token-level credit assignment by only controlling the baseline function.

3 Experiments

The experiment estimates how much the probability of eventual success varies across prefixes generated for the same question. Under the MSE analysis above, this between-state variation determines the maximum improvement available to a state-dependent baseline over the question-level GRPO baseline.

3.1 Questions, model, and reward

We use Qwen3-4B and the Hugging Face copy of DAPO-Math-17k. Because this copy repeats each underlying question approximately 103 times for the `verl` data loader, we first deduplicate examples using `extra_info.index`. We then apply a deterministic hash-ordered shuffle with seed 0 and select $N_0 = 100$ distinct questions.

All responses are sampled from Qwen3-4B in non-thinking mode with temperature 1.0, top-p = 1.0, and top-k disabled. The maximum generated-response length is 3,072 tokens. These settings define the policy π whose state values are being estimated.

We use a binary correctness reward. The predicted answer is extracted from the trailing `Answer:` line, with a `\boxed{}` expression used as a fallback, and is compared with the reference answer using `math_verify`. Responses that terminate because their token budget is exhausted receive reward zero.

3.2 Initial rollouts and filtering

For every selected question, we generate $G = 16$ independent source responses, giving 1,600 responses in total, and evaluate their binary rewards. Across all source responses the mean accuracy is 0.331, the truncation rate is 0.089, and the mean generated length is 1,485 tokens.

We retain only questions whose source-rollout rewards are mixed: at least one rollout must be correct and at least one must be incorrect. Of the 100 questions, 26 have reward zero for all rollouts, 3 have reward one for all rollouts, and 71 have mixed rewards. Removing both constant-reward groups therefore leaves $N = 71$ questions. Note that filtering only the all-zero groups would leave 74; both constant-reward cases must be removed, since a group with no reward variation has $p \in \{0, 1\}$ and an undefined relative ceiling.

This filtering focuses the experiment on questions for which the source policy exhibits nonzero empirical outcome variation. Consequently, the results are conditional on the retained questions and do not estimate the corresponding quantities over the original DAPO-Math-17k distribution.

3.3 Prefix construction and continuation sampling

For each retained question x_q, every one of its $G = 16$ source responses is used as a source trajectory. We select prefixes at the seven relative response positions

$$\mathcal{A} = \{0.10, 0.25, 0.40, 0.55, 0.70, 0.85, 0.95\}.$$

If $A_{q,g,\ell}$ is the token index associated with relative position ℓ, the corresponding state is

$$S_{q,g,\ell} = (x_q, Y_{<A_{q,g,\ell}}), \qquad g \in \{1, \ldots, G\}, \quad \ell \in \{1, \ldots, L\},$$

where $L = 7$. The grid deliberately excludes both endpoints, and the reported average over states is therefore specific to this position distribution.

For every state $S_{q,g,\ell}$, we generate $M = 16$ independent continuations using the same sampling policy. If the prefix contains k generated tokens, each continuation receives at most $3{,}072 - k$ additional tokens rather than a fresh 3,072-token budget. This preserves the total response-length constraint faced by the source trajectory and prevents late forks from escaping truncation-related failure modes merely because they were restarted from a prefix.

The trajectory-derived portion of the experiment contains

$$NGL = 71 \times 16 \times 7 = 7{,}952$$

states, or $K = GL = 112$ states per question. We additionally fork once from the bare prompt of every retained question, adding 71 validation states and producing 8,023 estimated states in total. The bare-prompt states are used only for validation and are not included in the between-prefix variance estimate. At $M = 16$ continuations per state this amounts to 128,368 sampled continuations.

3.4 Estimating state values

For each trajectory-derived or bare-prompt state S, let

$$R_{S,m} \in \{0,1\}, \qquad m \in \{1,\dots,M\},$$

be the rewards of its sampled continuations. We estimate its value by

$$\widehat{V}(S) = \frac{1}{M}\sum_{m=1}^{M} R_{S,m}.$$

Conditioned on the state, this estimator satisfies

$$\mathbb{E}[\widehat{V}(S) \mid S] = V(S), \qquad \mathrm{Var}(\widehat{V}(S) \mid S) = \frac{V(S)(1-V(S))}{M}.$$

3.5 Estimating variation across states

The $K = 112$ states of a question are not an i.i.d. sample. They are produced by a two-stage design: $G = 16$ source trajectories are drawn from the policy, and $L = 7$ positions are taken along each one. States sharing a trajectory are strongly correlated, because they share a prefix and share whatever made that trajectory succeed or fail. We therefore estimate the variance in the form the design actually has, using the law of total variance,

$$\mathrm{Var}_{S\sim d^{\pi}(\cdot|x_q)}[V(S)] = \underbrace{\mathrm{Var}_g(\mathbb{E}_\ell[V \mid g])}_{\text{between trajectories}} + \underbrace{\mathbb{E}_g(\mathrm{Var}_\ell[V \mid g])}_{\text{within trajectory}},$$

rather than pooling all K values into a single sample variance.

Write $\widehat{V}_{q,g,\ell} = \widehat{V}(S_{q,g,\ell})$, let

$$m_{q,g} = \frac{1}{L}\sum_{\ell=1}^{L} \widehat{V}_{q,g,\ell}, \qquad \overline{V}_q = \frac{1}{G}\sum_{g=1}^{G} m_{q,g},$$

and let

$$\widehat{\nu}(S) = \frac{\widehat{V}(S)\left(1-\widehat{V}(S)\right)}{M-1}, \qquad \bar{\nu}_q = \frac{1}{K}\sum_{S\in\mathcal{S}_q} \widehat{\nu}(S)$$

be the unbiased estimator of the per-state Monte Carlo noise variance and its question-level mean. Monte Carlo noise enters the two levels differently: a single $\widehat{V}(S)$ carries the full $\bar{\nu}_q$, whereas a trajectory mean $m_{q,g}$ averages L independent noise terms and therefore carries only $\bar{\nu}_q/L$. The two corrected components are

$$\widehat{\sigma}^2_{W,q} = \frac{1}{G}\sum_{g=1}^{G}\left[\frac{1}{L-1}\sum_{\ell=1}^{L}\left(\widehat{V}_{q,g,\ell} - m_{q,g}\right)^2\right] - \bar{\nu}_q,$$

$$\widehat{\sigma}^2_{B,q} = \frac{1}{G-1}\sum_{g=1}^{G}\left(m_{q,g} - \overline{V}_q\right)^2 - \frac{\bar{\nu}_q}{L}, \qquad \widehat{\sigma}^2_{V,q} = \widehat{\sigma}^2_{W,q} + \widehat{\sigma}^2_{B,q}.$$

Pooling all K values into one sample variance is unbiased only under i.i.d. sampling and is biased downward here, since the between-trajectory spread has 16 effective replicates rather than 112. None of the 71 questions yields a negative corrected component, so the optional nonnegative truncation $\max\{0, \cdot\}$ is never applied.

For each question with $0 < \overline{V}_q < 1$ the estimated ceiling is $\widehat{\rho}_q = \widehat{\sigma}^2_{V,q}/[\overline{V}_q(1-\overline{V}_q)]$, and the headline statistic is the equally weighted mean $\widehat{\rho} = \frac{1}{N}\sum_q \widehat{\rho}_q$.

3.6 Validation checks and uncertainty

The primary validation is the martingale identity

$$\mathbb{E}[V(S_t) \mid X = x] = p^{\pi}(x),$$

which requires the mean of $\widehat{V}(S)$ to be constant across all seven relative positions. Because it must hold at every position simultaneously, it is a sensitive joint check of prefix construction, continuation budgets, and reward evaluation: an off-by-one in the prefix index, a mis-specified token budget, or an inconsistent answer parser would all induce a position-dependent drift.

The bare-prompt forks were intended as a second check, comparing $\widehat{V}$ at the empty prefix against the source-rollout success rate. This check turned out to be *uninformative about $\widehat{V}$ as an estimator*. Both the source generation and the bare-prompt forks were issued to vLLM with the same seed, the same prompt, the same $n = 16$, and the same sampling parameters, so the two requests largely reproduce the same completions rather than resampling independently. Section 4 reports the resulting agreement, which is far closer than independent sampling would produce and therefore reflects request-level determinism rather than estimator accuracy. The check still verifies that prompt construction, detokenisation, and reward evaluation agree between the two code paths, and we report it only in that narrower sense. Trajectory-derived forks are unaffected, since each uses a distinct prefix and hence a distinct sampling stream.

States sampled from the same question and source response are correlated. Uncertainty intervals must therefore preserve this dependence. All standard errors reported below are computed across the $N = 71$ questions, treating each question as one observation, rather than treating the 7,952 trajectory-derived states as independent.

4 Results

4.1 The available improvement

The estimated ceiling over the $N = 71$ retained questions is

$$\widehat{\rho} = 0.287 \pm 0.020 \qquad (95\%\ \text{CI}\ [0.248,\ 0.327]),$$

so an exact state-value baseline would reduce the return-prediction MSE by roughly 29% relative to the question-level GRPO baseline. It varies widely by question — median 0.275, interquartile range [0.148, 0.417], exceeding 0.20 for 66% of questions and falling below 0.05 for only 8% — so the mean is not driven by a few outliers.

The variation is not concentrated at the end of the response. Table 1 shows $\mathrm{Var}[V(S_t)]$ growing steadily along the trajectory: by the midpoint, roughly 22% of the outcome variance is already determined by the prefix. A ceiling that only materialised in the final few tokens would be of little practical use for credit assignment; this one does not.

Relative position	Raw variance	MC noise	$\widehat{\sigma}^2_V$	$\widehat{\rho}$
0.10	0.0166	0.0119	0.0057	0.030
0.25	0.0277	0.0112	0.0166	0.084
0.40	0.0395	0.0104	0.0293	0.149
0.55	0.0537	0.0095	0.0444	0.224
0.70	0.0754	0.0080	0.0674	0.346
0.85	0.0940	0.0068	0.0872	0.462
0.95	0.1224	0.0053	0.1171	0.594

Table 1: Between-prefix value variance by relative position, corrected for Monte Carlo noise, computed within question and averaged over the 71 questions.

Of the total $\widehat{\sigma}^2_V = 0.0559 \pm 0.0044$, the decomposition of Section 3.5 attributes 0.0321 ± 0.0030 (57%) to variation *between* source trajectories, that is to how much the trajectory-average value differs from one response to another, and 0.0239 ± 0.0020 (43%) to variation *along* a trajectory.

Both components require state information beyond the position index. The martingale identity forces $\mathbb{E}[V(S_t)]$ to be constant across relative positions — which the means in Section 4.2 confirm — so a baseline that is a pure function of relative position carries no explanatory power at all: it can only reproduce a constant, which is what the GRPO baseline already is. The within-trajectory component is not a shared positional profile but trajectory-specific movement, and the larger between-trajectory component requires distinguishing concurrent responses to the same prompt from one another.

4.2 Validation

The martingale check is satisfied: the mean of $\widehat{V}(S)$ at the seven relative positions is 0.4195, 0.4078, 0.4100, 0.4187, 0.4246, 0.4147, 0.4231, with no systematic drift, and the overall mean 0.4169 agrees with the mean source-rollout success rate over retained questions, 0.4243.

The bare-prompt comparison gives a correlation of 0.968 with source-rollout success rates and a mean absolute difference of 0.0317, against roughly 0.139 expected under independent sampling with $M = G = 16$. As set out in Section 3.6, this reflects the shared vLLM seed rather than estimator accuracy, and we

report it only as confirmation that the two code paths construct prompts and evaluate rewards identically.

4.3 Scope

These results are conditional on the retained question set, the seven-position grid, the non-thinking Qwen3-4B policy at temperature 1.0, and the 3,072-token budget. They quantify a ceiling on *return-prediction MSE* reduction, which as noted in Section 2.2 is only a proxy for policy-gradient variance: it omits the score-norm weighting $\|\nabla_\theta \log \pi_\theta\|^2$, and for a gradient summed over a trajectory it also omits the cross-token covariance terms. Those terms matter specifically for per-token baselines. With a baseline that is constant along the trajectory, the summed gradient is a scalar multiple of $\nabla_\theta \log P_\theta(\tau)$; a baseline that varies along the trajectory reweights the summands and therefore changes the direction of that sum, not merely its magnitude. Establishing that a particular baseline realises a corresponding reduction in gradient variance consequently requires a separate measurement.