admit closed forms:

$$\mu_q = (1+\lambda_{\text{neg}})p - \lambda_{\text{neg}}, \tag{5}$$

$$\sigma_q = (1+\lambda_{\text{neg}})\sqrt{p(1-p)}. \tag{6}$$

Thus, the normalized advantages for correct and incorrect trajectories become

$$\begin{aligned} A^+ &= \frac{(1+\lambda_{\text{neg}})(1-p)}{(1+\lambda_{\text{neg}})\sqrt{p(1-p)}+\varepsilon_{\text{std}}}, \\ A^- &= \frac{-(1+\lambda_{\text{neg}})p}{(1+\lambda_{\text{neg}})\sqrt{p(1-p)}+\varepsilon_{\text{std}}}. \end{aligned} \tag{7}$$

When ε_{std} is small relative to $(1+\lambda_{\text{neg}})\sqrt{p(1-p)}$, the geometry is dominated by p: rare successes ($p \ll 1$) receive large positive advantages, while rare failures ($p \approx 1$) receive large negative advantages. This yields an automatic *rare-event amplification* effect driven by group normalization rather than an explicit curriculum.

Example with $G = 8$. With $G = 8$ and $\varepsilon_{\text{std}} \ll \sqrt{p(1-p)}$, the induced advantage geometry depends mainly on p: hard groups with $p = 1/8$ yield $A^+ \approx \sqrt{7}$ and $A^- \approx -1/\sqrt{7}$, while easy groups with $p = 7/8$ yield $A^+ \approx 1/\sqrt{7}$ and $A^- \approx -\sqrt{7}$. This illustrates the core mechanism: hard prompts amplify rare successes (positive anchors), whereas easy prompts amplify rare failures (negative guidance), producing an adaptive curriculum without explicit difficulty heuristics.

3.2. Instance Selection via Positive–Negative Pairing

Core idea. Instead of selecting training prompts solely by a single scalar heuristic (e.g., historical-accuracy variance), we explicitly construct a *bidirectional* minibatch consisting of (i) one prompt that yields a stable positive anchor and (ii) one prompt that yields a stable negative warning. Concretely, we select a two-example training set $\mathcal{D}_\pm = \{q^+, q^-\}$, where q^+ is *hard-but-solvable* (rare successes exist) and q^- is *easy-but-brittle* (rare failures exist). Under WGRPO, these two regimes map directly to amplified tail-event teaching signals (Sec. 3.1).

Positive anchor: hard-but-solvable. We choose q^+ such that the current policy achieves a low but non-zero success rate:

$$p(q^+) \in \left[\tfrac{1}{G}, \tfrac{c}{G}\right], \tag{8}$$

so that $0 < k < G$ and each group typically contains a small number of correct rollouts. In this regime, WGRPO assigns large positive advantages to rare correct trajectories, concentrating updates on demonstrations of what the model *should* do.

Negative guidance: easy-but-brittle. We choose q^- such that the current policy achieves a high but not perfect success rate:

$$p(q^-) \in \left[1-\tfrac{c}{G},\, 1-\tfrac{1}{G}\right], \tag{9}$$

so that failures are rare but still occur. In this regime, WGRPO assigns large-magnitude negative advantages to rare failures, producing a sharp "do-not" signal that suppresses high-confidence failure modes while preserving alternative plausible solutions under the model prior.

Practical selection via lightweight probing. To instantiate positive–negative pairing with only two training prompts, we perform a simple probing stage on two candidate pools with different expected difficulty under the same base model. We use an "easy" candidate pool $\mathcal{C}^-$ and a "hard" candidate pool $\mathcal{C}^+$; in our experiments $\mathcal{C}^-$ is drawn from DeepScaleR-sub and $\mathcal{C}^+$ is drawn from AIME 2025, but the procedure is agnostic to the specific sources. For each candidate $q \in \mathcal{C}^+ \cup \mathcal{C}^-$, we estimate its success rate under the current policy by sampling M independent groups of size G and averaging:

$$\bar{p}(q) \;=\; \frac{1}{M}\sum_{m=1}^{M} \hat{p}_m(q).$$

To ensure non-degenerate within-group variance, we discard candidates with $\bar{p}(q) \notin [\delta, 1-\delta]$, where we use $\delta = 1/G$ by default. We then select one positive anchor and one negative guidance prompt by targeting the two WGRPO regimes:

$$\begin{aligned} q^+ &= \arg\min_{q\in\mathcal{C}^+} |\bar{p}(q) - p_{\text{hard}}|, \\ q^- &= \arg\min_{q\in\mathcal{C}^-} |\bar{p}(q) - p_{\text{easy}}|, \end{aligned} \tag{10}$$

where $p_{\text{hard}} \approx 1/G$ and $p_{\text{easy}} \approx 1 - 1/G$. This ensures that q^+ operates in a low-but-nonzero success regime that amplifies rare successes, while q^- operates in a high-but-not-perfect success regime that amplifies rare failures. Overall, the selection is deliberately simple, uses only on-policy probing (no historical training statistics), and directly instantiates the rare-event amplification mechanism of WGRPO with only two training examples.

4. Experimental Setup

Models. To study how different training-example selection strategies affect RLVR, we run our training pipeline on several representative open-weight LLMs from different families and scales. In particular, we train QWEN2.5-MATH-7B, QWEN2.5-MATH-7B-INSTRUCT, and LLAMA-3.1-8B-INSTRUCT.

Training dataset. The training examples we select come from AIME 2025 (Art of Problem Solving, 2025a) and