STABLEMAX-PPO: BEST-OF-N RLHF WITH ORDINAL REWARD DISTRIBUTIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Many language-model systems use Best-of-N sampling at deployment, while standard reinforcement learning from human feedback (RLHF) optimizes the expected reward of a single response. Existing Best-of-N training objectives often reduce each response to a scalar and discard the per-rating probabilities that determine the group maximum. This matters because two responses with the same predicted mean can have different probabilities of receiving a high rating. We introduce **StableMax-PPO**, which directly optimizes the expected best latent rating among N candidates using the predictive distribution of a frozen ordinal reward model trained from human ratings. Its variance is retained as a model-based diagnostic of annotator disagreement, not an optimization bonus. A one-sided CDF ambiguity set with a fixed radius gives the exact worst-case expected latent rating maximum and an exact response-level policy credit for PPO. To preserve typical responses and limit policy drift, we impose lower bounds on the predicted mean rating and on an independently trained quality score, and limit changes from the reference policy using a standard KL measure of policy change with rollback. Across three seeds at $N = 32$, **StableMax-PPO** achieves a competitive robust maximum with lower held-out KL than Mean-PPO, group relative policy optimization (GRPO), and scalar max-PPO baselines.

1 INTRODUCTION

Many language-model systems generate N responses and then return, rerank, or inspect the best candidate (Nakano et al., 2021; Cobbe et al., 2021). Standard RLHF instead optimizes the expected score of one response (Christiano et al., 2017; Stiennon et al., 2020; Ouyang et al., 2022). A higher mean score need not make a highly rated candidate more likely among these N responses. Mean-PPO and group relative policy optimization (GRPO) are general-purpose methods (Shao et al., 2024; DeepSeek-AI et al., 2025); recent $\max@k$, Best-of-N, and inference-aware policy gradients target performance across multiple responses (Chow et al., 2025; Tang et al., 2025; Walder & Karkhanis, 2025; Bagirov et al., 2025; Li et al., 2026; Poiani et al., 2026). Their scalar-reward formulations nevertheless reduce each response to one score. Equal means can conceal different probabilities of receiving high ratings, and thus different expected rating maxima. Upper-tail optimization can also amplify calibration error, degrade typical responses, or increase policy drift (Gao et al., 2023; Kirk et al., 2024).

We introduce StableMax-PPO, which converts rating probabilities into a group objective and a training signal for each response. A frozen ordinal reward model (Moment RM), trained from repeated HelpSteer2 helpfulness ratings (Wang et al., 2024b), predicts probabilities over $\{0, 1, 2, 3, 4\}$. Its mean measures typical model-scored quality; its variance is a model-based diagnostic of annotator disagreement, not a reward bonus. The cumulative distribution function (CDF) gives the probability of a rating at or below each threshold. Raising these CDFs within a fixed bound makes high ratings less likely and exactly attains the worst-case expected maximum over the declared product ambiguity set. The same probabilities give each response's exact marginal credit: how much it adds to the group's expected maximum.

The technical distinction is that CDF robustification changes both the group objective and its policy credit. PPO implements the update. A nominal-mean floor and a separately trained, frozen quality reward model (Quality RM) from preference pairs constrain typical model-scored quality. Reference-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

policy KL regularization measures how much token probabilities move away from the frozen reference policy; state-complete rollback controls accepted updates. Maximum, both quality measurements, and drift remain separate reported quantities.

Our contributions are:

- **Using rating distributions for Best-of-N alignment.** We retain all five rating probabilities from repeated human annotations to compute the expected maximum. The mean measures typical model-scored quality; the variance remains a disagreement diagnostic. The contribution is this interface for policy optimization, not a new ordinal regression model.
- **An exact robust objective and policy credit.** We derive the worst-case expected maximum over a one-sided product CDF ambiguity set and its leave-one-out credit (Theorems 1 and 2). The score-function estimator is unbiased before PPO clipping. A true-maximum lower bound holds under conditional coverage (Theorem 3). Value and credit cost $O(N|\mathcal{R}|)$ without sampling latent ratings.
- **A constrained PPO realization with explicit safeguards.** Adaptive dual weights target nominal-mean and independently trained quality floors. KL control regulates policy drift; state-complete rollback restores rejected policy and optimizer updates, leaving controller state unchanged. These safeguards do not imply a population KL guarantee.
- **A three-seed study with $N = 32$ responses per prompt.** At $N = 32$, StableMax-PPO attains robust maximum 3.458 and held-out KL 0.034, versus 3.424 and 0.070 for Mean-PPO. GRPO and scalar max-PPO attain higher maxima at higher drift. Matched comparators, component ablations, and independent model-based diagnostics characterize this operating point, not a complete Pareto frontier or superiority on fresh human judgments.

2 Related Work

Post-training and inference-time selection. Learning rewards from preferences underlies RLHF (Christiano et al., 2017; Stiennon et al., 2020; Ouyang et al., 2022); DPO reparameterizes a KL-regularized preference objective without online policy optimization (Rafailov et al., 2023). GRPO uses within-group reward normalization for post-training (Shao et al., 2024; DeepSeek-AI et al., 2025). At inference time, verifier-guided selection, answer aggregation, and adaptive compute allocation exploit multiple candidate responses (Cobbe et al., 2021; Wang et al., 2023; Snell et al., 2025). These procedures differ in what constitutes a successful group. Our setting fixes N and targets its expected ordinal rating maximum; it does not optimize majority voting or allocate compute across prompts.

Best-of-N and multi-response training objectives. Best-of-N-aware fine-tuning (Chow et al., 2025), generic inference-time objectives (Tang et al., 2025), pass@k policy gradients (Walder & Karkhanis, 2025), and max@k optimization (Bagirov et al., 2025) align training with multi-candidate evaluation. InfAlign derives reward transformations for inference procedures (Balashankar et al., 2024), while Compute Aligned Training derives procedure-dependent losses (Ousherovitch & Tewari, 2026). Li et al. (2026) address a different limitation: they extrapolate scalar reward tails when training uses fewer rollouts than deployment. Poiani et al. (2026) study policy structure and sample complexity for max@k in finite-horizon RL. Our contribution concerns the distribution *within each response*: conditional ordinal rating probabilities define an exact robust group value and its credit at the same N. It neither replaces these broader inference-aware frameworks nor claims to solve a mismatch between training and deployment procedures.

Ordinal, probabilistic, and distributional rewards. Ordinal regression provides ordered-category models (McCullagh, 1980), while HelpSteer2 supplies graded human annotations for reward modeling (Wang et al., 2024b). Uncertainty-aware and probabilistic reward models represent information beyond a point score (Lou et al., 2024; Sun et al., 2025); predictive variation and epistemic uncertainty are distinct concepts (Kendall & Gal, 2017). Proper scoring rules and calibration assess the predictive probabilities (Gneiting & Raftery, 2007; Guo et al., 2017). Distributional RL instead models return distributions over trajectories (Bellemare et al., 2017; Dabney et al., 2018). Here the frozen distribution is over annotations conditional on a completed response. Its CDF enters the Best-of-N objective, while its variance is neither an epistemic confidence measure nor an exploration bonus.

Robustness and reward-model overoptimization. Robust MDP formulations optimize against uncertainty sets over transition models (Iyengar, 2005; Nilim & El Ghaoui, 2005). We use a product ambiguity set over conditional rating CDFs; its exact solution relies on stochastic ordering and conditional independence rather than dynamic programming. Reward-model overoptimization motivates safeguards against optimizing a proxy beyond its reliable range (Gao et al., 2023), and reward-model ensembles can mitigate this problem (Coste et al., 2024). Our one-sided CDF envelope modifies the objective and credit together, but its true-rating lower bound requires conditional coverage. Neither calibration on held-out prompt clusters nor agreement between model-based evaluators establishes that coverage for every generated response.

Diversity and constrained policy optimization. RLHF can reduce output diversity (Kirk et al., 2024); deliberate representation-based exploration can improve multi-sample coverage (Tuyls et al., 2026). Our credit rewards incremental rating-threshold coverage, without adding a semantic-diversity or variance bonus. PPO and trust-region methods provide policy-update mechanisms (Schulman et al., 2017; 2015), and constrained policy optimization separates reward objectives from constraints (Achiam et al., 2017). We use adaptive quality-floor penalties and empirical KL control with rollback. This implementation does not inherit a population feasibility or monotonic-improvement guarantee from those methods.

3 Problem Setup

3.1 Setting and assumptions

We study response-level RLHF with a fixed N for Best-of-N generation. The policy parameter θ is trainable, whereas the ordinal Moment RM ω, scalar Quality RM η, and reference policy π_{ref} are frozen. For $x \sim \mathcal{D}$, responses factor as $\pi_\theta(Y \mid x) = \prod_h \pi_\theta(a_h \mid x, a_{<h})$, and a rollout contains $Y_{1:N} \overset{\text{iid}}{\sim} \pi_\theta(\cdot \mid x)$. We assume (A1) conditional independence of rating outcomes across candidates, (A2) frozen evaluators and radius, and (A3) differentiable policies with common support and integrable scores permitting differentiation under expectation. Each response receives one terminal evaluation; the group statistic is formed before any response-level credit is assigned.

Assumption (A1) concerns annotation outcomes conditional on completed responses; it does not assert token independence. It is what converts each threshold event for the maximum into a product of candidate CDFs. Assumption (A2) keeps both evaluators and the reference policy fixed, so the measured operating point changes only through the policy and controller. Assumption (A3) enables the likelihood-ratio differentiation in Theorem 2. Correlated evaluator errors are therefore outside the product model rather than an effect absorbed by the optimizer.

3.2 Ordinal evaluator and Best-of-N value

The Moment RM returns $p_{\omega,r}(x, y) = \Pr_\omega(R = r \mid x, y)$ for $\mathcal{R} = \{0, 1, 2, 3, 4\}$; write $\Delta(\mathcal{R})$ for the probability simplex on these five levels. Its mean $\mu_\omega = \sum_r r p_{\omega,r}$ measures target-RM quality, while $\sigma^2_\omega = \sum_r (r - \mu_\omega)^2 p_{\omega,r}$ is retained only as a model-based disagreement diagnostic; the independent scalar $Q_\eta(x, y)$ is a separate quality check. The category endpoints are fixed by the training data; the ordinal-probit parameterization is given in Appendix A. The five levels are fixed by the evaluator interface, not selected after observing policy outcomes.

For candidate j, let $F_{\omega,j}(t) = \sum_{r=0}^{t} p_{\omega,r}(x, Y_j)$. Under (A1), the nominal Best-of-N identity is $h_N^{\text{nom}} = \sum_{t=1}^{4} [1 - \prod_{j=1}^{N} F_{\omega,j}(t-1)]$, and $J_N^{\text{nom}}(\theta) = \mathbb{E}[h_N^{\text{nom}}]$. Its threshold products show why the full ordinal vector matters: a candidate is useful when the other responses leave a threshold uncovered, which a scalar maximum cannot represent from response means alone. The identity integrates out rating noise exactly; the top-rating event is only one of the four threshold terms.

This objective measures the expected best latent rating in a candidate group. It is not the expected rating of a response selected by its predicted mean: the maximum is taken over rating outcomes before integrating them out. We evaluate this group functional directly; mean-based response selection is reported only as a secondary diagnostic (Appendix B.7).

Fix a radius $\varepsilon \in [0,1]$ before policy training and use the one-sided envelope $\overline{F}_j(t) = \min\{1, F_{\omega,j}(t) + \varepsilon\}$, with $\overline{p}_{j,r} = \overline{F}_j(r) - \overline{F}_j(r-1)$. Define $\mathfrak{U}_\varepsilon(p_{\omega,1:N}) := \prod_{j=1}^N \{q_j \in \Delta(\mathcal{R}) : F_{q_j}(t) \le \overline{F}_j(t),\ t = 0,\dots,3\}$. Set $\overline{F}_j(-1) = 0, \overline{F}_j(4) = 1$. Increasing a CDF is adverse to an upper-tail maximum, so the envelope changes the deployment target and its policy credit together; it is not an auxiliary reward bonus. The radius fixed in the recorded runs is $\varepsilon_{\text{run}} = 0.0468083$. Those results characterize this fixed ambiguity set; they do not validate a population calibration guarantee. Appendix A gives the corrected calibration bound and distinguishes it from the recorded radius.

3.3 Robust constrained objective

The robust group value is

$$h_N^{\text{rob}} := \inf_{q_{1:N} \in \mathfrak{U}_\varepsilon(p_{\omega,1:N})} \mathbb{E}_q[\max_j R_j] = \sum_{t=1}^{4} \Big[1 - \prod_{j=1}^{N} \overline{F}_j(t-1)\Big]. \tag{1}$$

Our optimization objective is then

$$\max_\theta\ J_N^{\text{rob}}(\theta) := \mathbb{E}[h_N^{\text{rob}}(x, Y_{1:N})] \quad \text{s.t.} \quad M(\theta) \ge m_0,\ Q(\theta) \ge q_0,\ K(\theta) \le \delta, \tag{2}$$

where $M(\theta) := \mathbb{E}[\mu_\omega(x,Y)]$, $Q(\theta) := \mathbb{E}[Q_\eta(x,Y)]$, and

$$K(\theta) := \frac{\mathbb{E}_{x,Y\sim\pi_\theta}[\sum_{h=1}^{L(Y)} k_\theta(s_h)]}{\mathbb{E}_{x,Y\sim\pi_\theta}[L(Y)]},$$

where $L(Y) \ge 1$ is response length and $k_\theta(s) = \mathrm{KL}(\pi_\theta(\cdot \mid s) \| \pi_{\text{ref}}(\cdot \mid s))$. Thus K is a token-weighted population KL; rollout and held-out versions use different state distributions (Appendix C). The floors m_0, q_0 are fixed from the reference policy and keep target-RM and independent quality visible, while K measures policy motion separately. In PPO, the floors enter through dual credits, the KL term through the surrogate, and the distinct threshold δ_{hard} through batch rollback; the measured token KL is an acceptance statistic, not a population trust-region guarantee. The Quality RM is independently trained but participates in policy optimization through its floor. It is therefore a safeguard evaluator, not an untouched external test; the generated-output proxy evaluators provide additional diagnostics (Appendix B.9).

4 Method

We first transform the evaluator distribution into a worst-case Best-of-N functional, then derive response-level credits and incorporate them into PPO. The exact identities below characterize the unclipped target; the subsequent controller specifies how tentative updates are accepted.

4.1 Exact robust maximum and marginal credit

Theorem 1 (Exact Best-of-N robust maximum)**.** *Under (A1), for every fixed $(x, Y_{1:N})$, an optimizer over the product ambiguity set is $q_j = \overline{p}_j$ for every j, and the attained worst-case Best-of-N value is*

$$\inf_{q_{1:N} \in \mathfrak{U}_\varepsilon(p_{\omega,1:N})} \mathbb{E}_q[\max_j R_j] = \sum_{t=1}^{4} \left[1 - \prod_{j=1}^{N} \overline{F}_j(t-1)\right]. \tag{3}$$

The key point is simultaneous optimality. The maximum is coordinatewise increasing, so saturating every admissible CDF upper bound makes all candidates stochastically smallest at once; the adversary need not choose a different perturbation for each threshold. The full ambiguity-set problem therefore reduces to four threshold products, without rating sampling or enumerating 5^N joint outcomes.

For a candidate i, let $\overline{R}_i \sim \overline{p}_i$ and $\overline{M}_{-i} = \max_{j\ne i} \overline{R}_j$, with empty maximum zero when $N = 1$. Its leave-one-out marginal credit is

$$q_i^{\text{rob}} := \mathbb{E}[(\overline{R}_i - \overline{M}_{-i})_+ \mid x, Y_{1:N}] = \sum_{t=1}^{4} [1 - \overline{F}_i(t-1)] \prod_{j\ne i} \overline{F}_j(t-1). \tag{4}$$

The first factor is candidate i's probability of crossing threshold t; the product over $j \neq i$ is the probability that the other responses leave that threshold uncovered. This leave-one-out value captures incremental group credit rather than a winner indicator, so a response can matter without having the largest mean score. Prefix and suffix products compute all credits in $O(N|\mathcal{R}|)$ time per prompt using the same kernel as the group objective, without sampling latent ratings. It is a marginal, rather than additive, decomposition: ties or redundant candidates need not exhaust the group value, because the signal is intended to measure incremental threshold coverage.

Changing N changes the group functional, not merely the number of reward observations. The following identity applies likelihood-ratio differentiation (Williams, 1992; Sutton et al., 1999) to this fixed-N target.

Theorem 2 (Exact marginal policy credit)**.** *Under (A1)–(A3), the exact group objective has the score-function gradient*

$$\nabla_\theta J_N^{\mathrm{rob}}(\theta) = \mathbb{E}\left[\sum_{i=1}^{N} q_i^{\mathrm{rob}} \nabla_\theta \log \pi_\theta(Y_i \mid x)\right]. \tag{5}$$

Thus the response-level estimator is unbiased before PPO clipping.

The same cached probabilities define the robust target and every leave-one-out credit. The likelihood ratio distributes each response's credit over its tokens, without sampled ratings or a learned group-winner critic. PPO clipping, normalization, KL regularization, and rollback modify the realized update; the theorem does not assert that this practical update remains an unbiased gradient estimator of J_N^{rob}.

4.2 Constrained PPO update

For dual variables $\lambda_m, \lambda_q \geq 0$, the response-level credit used by PPO is

$$\widetilde{q}_i = q_i^{\mathrm{rob}} + \frac{\lambda_m}{N}\mu_\omega(x, Y_i) + \frac{\lambda_q}{N} Q_\eta(x, Y_i), \qquad A_i = N[\widetilde{q}_i - b_\psi(x)]. \tag{6}$$

The factors $1/N$ match the one-response constraints to the group objective; the prompt-only baseline preserves the score-function identity. We use the standard clipped PPO surrogate (Schulman et al., 2017) and full-categorical token KL. Writing $\rho_{i,h} = \pi_\theta(a_{i,h} \mid s_{i,h})/\pi_{\mathrm{old}}(a_{i,h} \mid s_{i,h})$, the update maximizes the standard per-token expression $\min\{\rho_{i,h}A_i, \mathrm{clip}(\rho_{i,h}, 1-\epsilon_{\mathrm{PPO}}, 1+\epsilon_{\mathrm{PPO}})A_i\}$ summed over action tokens and averaged over candidates, minus $\beta_{\mathrm{KL}}\widehat{\mathrm{KL}}$, and divides by the lagged RMS scale. The full implementation, including projected dual updates, is specified in Appendix C.

Before a tentative step, StableMax-PPO snapshots all policy parameters and the complete optimizer state. If the post-update KL on the same rollout states exceeds δ_{hard}, both snapshots are restored; baseline, dual, RMS, and controller states advance only after acceptance. Snapshotting the optimizer moments is necessary for Adam-like optimizers, and the resulting rollback count is comparable across baselines under the shared controller. The soft target δ and the hard threshold δ_{hard} serve different purposes: the former is the reference-policy constraint represented by the KL controller, whereas the latter is a batch-level reject rule. An accepted update can still differ from the population-constrained optimum, which is why held-out KL is measured independently rather than inferred from the rollback threshold.

4.3 StableMax-PPO

Here Opt denotes the complete policy-optimizer state, including momentum and adaptive-variance buffers. We compute $\widehat{\mathrm{KL}}$ as the batch mean of the full-categorical token KL on rollout states, not from sampled actions. The update uses q_i^{rob} and A_i from eqs. (4) and (6). Each iteration evaluates the frozen models on a fixed-size rollout and updates the policy and controller. The

models, CDF radius, floors, soft target δ, and hard rollback threshold remain fixed by the protocol.

Algorithm 1: StableMax-PPO update

Input: Prompts $x \sim \mathcal{D}$; initial trainable policy π_θ; frozen $p_\omega, Q_\eta, \pi_{\text{ref}}$; fixed $N, \varepsilon, m_0, q_0, \delta, \delta_{\text{hard}}$.

Output: Final accepted policy parameters θ, optimizer state Opt, and controller state $(b_\psi, \lambda_m, \lambda_q, \beta_{\text{KL}}, s)$.

Initialize Opt, b_ψ, λ_m, λ_q, β_{KL}, and lagged RMS scale s;
for $k = 1, \ldots, T_{\text{pol}}$ **do**
  Set $\pi_{\text{old}} \leftarrow \pi_\theta$ and roll out N responses per prompt; evaluate the frozen p_ω, Q_η;
  Compute q_i^{rob} from eq. (4) and A_i from eq. (6);
  Snapshot (θ, Opt) and take one clipped-PPO step with coefficient β_{KL};
  Measure $\widehat{\text{KL}}(\pi_\theta \| \pi_{\text{ref}})$ on the same rollout states;
  if $\widehat{\text{KL}} > \delta_{\text{hard}}$ **then**
    Restore (θ, Opt), discard the tentative update and its rollout, and do not advance $b_\psi, \lambda_m, \lambda_q, \beta_{\text{KL}}, s$;
  else
    Commit the transition; fit b_ψ, project the dual updates, and update β_{KL}, s from accepted data;

All quantities needed to reproduce the update are fixed above or in Appendix C; only the policy and controller states are updated, and rejected transitions advance none of them.

5 Theoretical Analysis

The previous section gives the exact target and its policy-gradient bridge. Here we state the guarantees that explain the role of the calibration radius and the update controller. Proofs are in Appendix A.

Theorem 3 (Calibration-robust lower bound)**.** *Under (A1), let F_j^* be the true conditional CDF. If $F_j^*(t) \le \overline{F}_j(t)$ for every candidate and $t \in \{0,1,2,3\}$, then the true conditional maximum satisfies $\mathbb{E}_*[\max_j R_j \mid x, Y_{1:N}] \ge h_N^{\text{rob}}(x, Y_{1:N})$. Without exact coverage, $\mathbb{E}_*[\max_j R_j \mid x, Y_{1:N}] \ge h_N^{\text{rob}}(x, Y_{1:N}) - 4N\xi(x, Y_{1:N})$, where $\xi = \max_{j\in[N],\, t\in\{0,1,2,3\}} [F_j^*(t) - \overline{F}_j(t)]_+$.*

This is a conditional envelope statement, not a distribution-free guarantee: even a valid population-average calibration bound does not establish pointwise coverage for every generated response. If the envelope misses by at most ξ, the four ordinal thresholds and N candidate CDFs contribute the explicit $4N\xi$ degradation term. The factor depends on the declared five-level support and product model, so it is a sensitivity certificate for this evaluator interface rather than a claim of universal calibration.

The factor $4N$ follows from the representation itself: at each of the four nontrivial thresholds, changing one candidate CDF changes the corresponding product by at most its CDF error. Summing these N-candidate sensitivities over thresholds gives the stated degradation, independently of the optimizer.

The radius also has a deterministic target-level interpretation independent of how it is estimated.

Proposition 1 (Calibration-radius monotonicity)**.** *For $0 \le \varepsilon_1 \le \varepsilon_2 \le 1$, $0 \le J_N^{\text{rob}}(\theta; \varepsilon_1) - J_N^{\text{rob}}(\theta; \varepsilon_2) \le 4N(\varepsilon_2 - \varepsilon_1)$, and at $\varepsilon = 0$, $J_N^{\text{rob}}(\theta; 0) = J_N^{\text{nom}}(\theta)$.*

The radius is therefore an interpretable target-level knob: increasing it makes the robust value more conservative with bounded sensitivity. Unlike a KL coefficient, it changes both the group target and its marginal credit, so the nominal-EV comparator isolates a change in the evaluator-defined objective. A radius change is consequently not equivalent to retuning PPO: it changes the policy-gradient direction as well as the objective level.

Let $\phi_i = \nabla_\theta \log \pi_\theta(Y_i \mid x)$. Consider a hypothetical estimator that samples latent ratings from the robust envelope, $G_{\text{samp}} = \sum_i (\overline{R}_i - \overline{M}_{-i})_+ \phi_i$, and the implemented credit estimator $G_{\text{RB}} =$

$\sum_i q_i^{\mathrm{rob}} \phi_i$. The former introduces evaluator-sampling noise that can be removed by conditional expectation (Casella & Robert, 1996).

Proposition 2 (Rao–Blackwell variance dominance)**.** *Assume* $\mathbb{E}\|G_{\mathrm{samp}}\|^2 < \infty$. *Conditioned on* $(x, Y_{1:N})$, $G_{\mathrm{RB}} = \mathbb{E}[G_{\mathrm{samp}} \mid x, Y_{1:N}]$. *Consequently, for every vector* v, $\mathrm{Var}(v^\top G_{\mathrm{RB}}) \leq \mathrm{Var}(v^\top G_{\mathrm{samp}})$.

The exact credit removes variance from drawing latent ordinal ratings while leaving rollout and policy-sampling variance unchanged. The PPO update thus uses the conditional expectation of the group credit without extra evaluator calls; the proposition makes no claim that total optimization variance vanishes. It identifies one avoidable source of noise and leaves response-sampling, optimization, and controller variability to the empirical analysis.

The last result concerns the realized optimizer state rather than the population objective.

Proposition 3 (Rollback correctness)**.** *If a tentative update is rejected, restoring the parameter and optimizer snapshots exactly returns the measured rollout-state policy and optimizer to their pre-update values. This is a batch-level invariant, not a population KL guarantee.*

The envelope protects the model-conditional target, the marginal identity makes it optimizable, and rollback controls the measured batch update. Restoring the complete snapshot is essential for optimizer and controller consistency, but does not imply a population trust region after an accepted update; maximum, quality, held-out KL, and rollback frequency are therefore reported separately. In particular, restoring policy weights without adaptive buffers could make a rejected transition affect the next step, so the invariant is stated for the complete optimizer/controller state.

6 Experiments

Our empirical question is whether the exact robust ordinal objective and its policy credit yield a reproducible maximum–drift trade-off under matched training and quality safeguards. We hold N, the prompt stream, seeds, optimizer, and rollback rule fixed, while comparators vary the deployment functional, calibration treatment, or safeguard. We report robust maximum together with independently modeled quality, held-out KL, and rollback frequency so that target performance and policy stability remain separately observable.

6.1 Experimental setup

The policy is Qwen2.5-1.5B-Instruct. The frozen ordinal Moment RM is a Qwen2.5-14B model trained on HelpSteer2 repeated helpfulness ratings, with the decontaminated preference replay specified in Appendix B (Wang et al., 2024b). The independent scalar Quality RM is a separately trained Qwen2.5-14B model fit on disjoint preference data. We use three confirmatory seeds $\{314, 2718, 1618\}$, $N = 32$, and 512 held-out prompts per seed. Primary baselines are Mean-PPO + Q, GRPO + Q, and Scalar Max@32-PPO + Q; full settings, ablations, and audit records are in Appendix B. We report robust maximum, held-out KL, and rollback frequency because they measure the two sides of the proposed operating point.

Repeated helpfulness ratings supervise the categorical output and its disagreement diagnostic; the independent Quality RM and generated-output proxies test transfer beyond the target evaluator. Shared prompts, seeds, the value of N, generation settings, and rollback thresholds keep every comparison on the same response-sampling problem. All confirmatory claims therefore concern this fixed $N = 32$ operating point, not scale or generalization across N.

The primary comparators are chosen to isolate the two axes in the objective. Mean-PPO keeps the conventional single-response target, GRPO keeps a group-relative scalar target, and Scalar max-PPO uses the same N with the ordinal distribution collapsed to response means. Matching their optimizer, quality safeguards, and rollback rule makes the comparison about the deployment functional and its credit assignment, rather than about a different amount of computation or a different update stabilizer. The additional evaluator and floor variants in Appendix B then vary one contract component at a time.

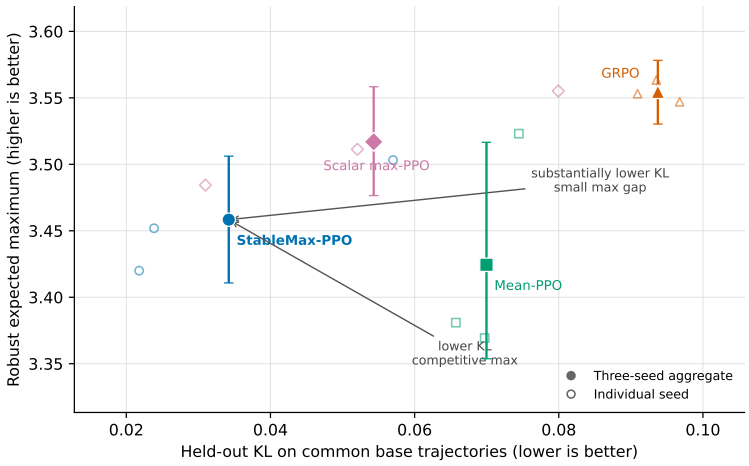

Figure 1: Robust maximum versus held-out KL. Filled markers are three-seed aggregates; open markers are individual seeds and bars are 95% intervals.

6.2 Main results

In Table 1, StableMax-PPO reaches a robust expected maximum of 3.458 and held-out KL of 0.034, compared with 3.424 and 0.070 for Mean-PPO. Under the shared rollback controller it rejects 31.0% of tentative updates, versus 58.6% for Mean-PPO and 82.3% for GRPO. GRPO and Scalar max-PPO attain higher robust maxima, so the evidence supports a lower-drift operating point rather than universal dominance. The primary endpoint is the same h_{32}^{rob} group functional used in training; it does not measure fresh human ratings of a selected response.

Figure 1 makes the operating point explicit: StableMax-PPO has lower held-out KL than all three trainable primary comparators while retaining a competitive robust maximum, whereas the max-oriented comparators remain higher on that endpoint. Rollback frequency is an optimization diagnostic, not a population guarantee.

Relative to Mean-PPO, the robust-maximum point estimate is higher by 0.034, but the paired hierarchical-bootstrap 95% interval for the difference is $[-0.069, 0.130]$ and does not establish superiority. Held-out KL decreases from 0.070 to 0.034 (a 51% relative reduction). Relative to Scalar max-PPO, it reduces held-out KL by 0.020 (37%), with a 0.059-point lower robust maximum (3.458 versus 3.517). Relative to GRPO, KL is 64% lower, with robust maxima of 3.458 versus 3.555. Percentages use unrounded aggregate records. These are the two axes that the method exposes; collapsing them into one weighted score would conceal the robust–stable operating point the experiment is designed to measure.

Held-out KL is a standard measure of how much the policy's next-token probabilities differ from the reference on common base-policy trajectories (Appendix C). Lower KL therefore means less measured change from the reference on these evaluation states. This is the measured drift-control benefit; quality is assessed separately by the declared floors and independent proxies.

Table 1: Primary aggregate results (three seeds, 512 prompts per seed). StableMax-PPO has lower aggregate KL than all three trainable primary comparators: 37% lower than scalar max-PPO, with a 0.059-point lower robust maximum. All trainable rows use the matched quality safeguards.

Method	Robust max	95% CI	Held-out KL	Rollbacks/900
StableMax-PPO	3.458	[3.411, 3.506]	0.034	279 (31.0%)
Mean-PPO	3.424	[3.354, 3.517]	0.070	527 (58.6%)
GRPO	3.555	[3.530, 3.578]	0.094	741 (82.3%)
Scalar max-PPO	3.517	[3.476, 3.558]	0.054	285 (31.7%)
Best-of-32 (base)	3.086	[3.045, 3.125]	–	–

The frozen mean and quality floors pass their declared tests, but the complete predeclared scientific-success criterion is not met (Appendix B.2). Higher maxima for GRPO and scalar max-PPO and the overlapping independent-proxy intervals delimit the claim: these runs demonstrate an objective-aligned, lower-drift policy, not a uniformly better policy or an established Pareto frontier.

6.3 Ablation and stability analysis

Removing the mean floor raises robust maximum from 3.458 to 3.508 and held-out KL from 0.034 to 0.048. Removing the quality floor gives $(3.467, 0.048)$, while removing CDF robustification gives $(3.421, 0.028)$, in (robust maximum, held-out KL) coordinates. Relative to either no-floor variant, the full method has about 29% lower aggregate KL and a lower robust maximum. The robust variant has higher values on both axes than the nominal ordinal variant, distinguishing the observed target trade-off from that of the floors. The complete aggregate and per-seed tables are in Appendix B. The three-seed intervals overlap and per-seed effects are heterogeneous, so these are sensitivity analyses rather than isolated causal effects. Removing CDF robustification changes the training objective; removing a floor changes the accepted-update trajectory. No component is presented as a universally improving reward term.

Together, the results answer the objective-design question at the evaluated operating point: the robust ordinal credit trains against the same Best-of-N statistic used for evaluation, while quality floors and held-out KL keep typical performance and policy movement visible. Independent generated-output diagnostics are reported in Appendix B.9; they are model-based diagnostics rather than fresh human judgments.

7 Limitations

Scope. The endpoint is induced by frozen evaluators, with no fresh human labels for generated outputs; model proxies may disagree with people, the 1.5B/14B capacity gap may induce evaluator shift, and three seeds do not establish scale or domain generalization. We fix $N = 32$, and the exact product form assumes conditionally independent rating outcomes, so correlated annotation or evaluator errors are outside the guarantee.

Future work. Blinded human ratings, additional N values and domains, larger policies, and correlated-error ambiguity sets are natural next tests. The last direction is especially important because it asks whether the threshold construction can retain tractable policy credit when evaluator errors are shared across candidates.

Reproducibility statement. The supplementary artifact contains the source, official ICLR style files, protocol, result tables, verification JSON, build commands, and experiment code in the accompanying anonymous archive. The main paper specifies the objective, algorithm, seeds, metrics, and aggregate results; appendices give proofs and the complete protocol. Model weights, credentials, private data, and other non-redistributable files are excluded.

AI use statement

Generative AI tools assisted with paper organization, methodological exposition, literature retrieval, reference checking, language editing, LaTeX and figure formatting, debugging of analysis code, proof development and correction, and interpretation of reported results. Language models serve as the experimental policies and evaluators described in this paper; editing assistants did not supply the reported measurements. The authors are responsible for verifying mathematical claims and proofs, checking implementation and numerical values against the frozen records, and reviewing all AI-assisted changes before submission.

Ethics statement. The study uses model-generated text and existing preference data. It collects no new human-subject data or personal information. Model-based preference proxies are reported as proxies, not as human judgments of newly generated outputs; possible reward-model exploitation and evaluator shift are disclosed in Section 7 and Appendix B.

REFERENCES

Joshua Achiam, David Held, Aviv Tamar, and Pieter Abbeel. Constrained policy optimization. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 22–31. PMLR, 2017. URL https://proceedings.mlr.press/v70/achiam17a.html.

Farid Bagirov, Mikhail Arkhipov, Ksenia Sycheva, Evgeniy Glukhov, and Egor Bogomolov. The best of N worlds: Aligning reinforcement learning with best-of-N sampling via max@k optimisation. *arXiv preprint arXiv:2510.23393*, 2025. URL https://arxiv.org/abs/2510.23393.

Ananth Balashankar, Ziteng Sun, Jonathan Berant, Jacob Eisenstein, Michael Collins, Adrian Hutter, Jong Lee, Chirag Nagpal, Flavien Prost, Aradhana Sinha, Ananda Theertha Suresh, and Ahmad Beirami. InfAlign: Inference-aware language model alignment. *arXiv preprint arXiv:2412.19792*, 2024. URL https://arxiv.org/abs/2412.19792.

Marc G. Bellemare, Will Dabney, and Rémi Munos. A distributional perspective on reinforcement learning. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 449–458. PMLR, 2017. URL https://proceedings.mlr.press/v70/bellemare17a.html.

George Casella and Christian P. Robert. Rao–blackwellisation of sampling schemes. *Biometrika*, 83 (1):81–94, 1996. doi: 10.1093/biomet/83.1.81. URL https://doi.org/10.1093/biomet/83.1.81.

Yinlam Chow, Guy Tennenholtz, Izzeddin Gur, Vincent Zhuang, Bo Dai, Aviral Kumar, Rishabh Agarwal, Sridhar Thiagarajan, Craig Boutilier, and Aleksandra Faust. Inference-aware fine-tuning for best-of-N sampling in large language models. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=77gQUdQhE7.

Paul F. Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep Reinforcement Learning from Human Preferences. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://papers.nips.cc/paper_files/paper/2017/hash/d5e2c0adad503c91f91df240d0cd4e49-Abstract.html.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training Verifiers to Solve Math Word Problems. *arXiv preprint arXiv:2110.14168*, 2021. URL https://arxiv.org/abs/2110.14168.

Thomas Coste, Usman Anwar, Robert Kirk, and David Krueger. Reward Model Ensembles Help Mitigate Overoptimization. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=dcjtMYkpXx.

Will Dabney, Mark Rowland, Marc G. Bellemare, and Rémi Munos. Distributional reinforcement learning with quantile regression. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*, pp. 2892–2901. AAAI Press, 2018. doi: 10.1609/aaai.v32i1.11791. URL https://doi.org/10.1609/aaai.v32i1.11791.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, et al. DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025. URL https://arxiv.org/abs/2501.12948.

Leo Gao, John Schulman, and Jacob Hilton. Scaling laws for reward model overoptimization. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 10835–10866. PMLR, 2023. URL https://proceedings.mlr.press/v202/gao23h.html.

Tilmann Gneiting and Adrian E. Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007. doi: 10.1198/016214506000001437. URL https://doi.org/10.1198/016214506000001437.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1321–1330. PMLR, 2017. URL https://proceedings.mlr.press/v70/guo17a.html.

Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963. doi: 10.1080/01621459.1963.10500830.

Garud N. Iyengar. Robust dynamic programming. *Mathematics of Operations Research*, 30(2): 257–280, 2005. doi: 10.1287/moor.1040.0129. URL https://doi.org/10.1287/moor.1040.0129.

Alex Kendall and Yarin Gal. What Uncertainties Do We Need in Bayesian Deep Learning for Computer Vision? In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://papers.nips.cc/paper_files/paper/2017/hash/2650d6089a6d640c5e85b2b88265dc2b-Abstract.html.

Robert Kirk, Ishita Mediratta, Christoforos Nalmpantis, Jelena Luketina, Eric Hambro, Edward Grefenstette, and Roberta Raileanu. Understanding the Effects of RLHF on LLM Generalisation and Diversity. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=PXD3FAVHJT.

Nathan Lambert, Valentina Pyatkin, Jacob Morrison, LJ Miranda, Bill Yuchen Lin, Khyathi Chandu, Nouha Dziri, Sachin Kumar, Tom Zick, Yejin Choi, Noah A. Smith, and Hannaneh Hajishirzi. RewardBench: Evaluating reward models for language modeling. *arXiv preprint arXiv:2403.13787*, 2024. URL https://arxiv.org/abs/2403.13787.

Muheng Li, Jian Qian, and Wenlong Mou. What should post-training optimize? a test-time scaling law perspective. *arXiv preprint arXiv:2605.10716*, 2026. URL https://arxiv.org/abs/2605.10716.

Chris Yuhao Liu, Liang Zeng, Jiacai Liu, Rui Yan, Jujie He, Chaojie Wang, Shuicheng Yan, Yang Liu, and Yahui Zhou. Skywork-Reward: Bag of tricks for reward modeling in LLMs. *arXiv preprint arXiv:2410.18451*, 2024. URL https://arxiv.org/abs/2410.18451.

Xingzhou Lou, Dong Yan, Wei Shen, Yuzi Yan, Jian Xie, and Junge Zhang. Uncertainty-aware reward model: Teaching reward models to know what is unknown. *arXiv preprint arXiv:2410.00847*, 2024. doi: 10.48550/arXiv.2410.00847. URL https://arxiv.org/abs/2410.00847.

Peter McCullagh. Regression models for ordinal data. *Journal of the Royal Statistical Society: Series B (Methodological)*, 42(2):109–127, 1980. doi: 10.1111/j.2517-6161.1980.tb01109.x. URL https://doi.org/10.1111/j.2517-6161.1980.tb01109.x.

Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, Xu Jiang, Karl Cobbe, Tyna Eloundou, Gretchen Krueger, Kevin Button, Matthew Knight, Benjamin Chess, and John Schulman. WebGPT: Browser-assisted question-answering with human feedback. *arXiv preprint arXiv:2112.09332*, 2021. URL https://arxiv.org/abs/2112.09332.

Arnab Nilim and Laurent El Ghaoui. Robust control of Markov decision processes with uncertain transition matrices. *Operations Research*, 53(5):780–798, 2005. doi: 10.1287/opre.1050.0216. URL https://doi.org/10.1287/opre.1050.0216.

Adam Ousherovitch and Ambuj Tewari. Compute Aligned Training: Optimizing for Test Time Inference. *arXiv preprint arXiv:2604.24957*, 2026. URL https://arxiv.org/abs/2604.24957.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pp. 27730–27744, 2022. doi: 10.52202/068431-2011. URL https://papers.nips.cc/paper_files/paper/2022/hash/b1efde53be364a73914f58805a001731-Abstract-Conference.html.

Riccardo Poiani, Martino Bernasconi, and Andrea Celli. Theoretical foundations of max@k reinforcement learning. *arXiv preprint arXiv:2607.17823*, 2026. URL https://arxiv.org/abs/2607.17823.

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. In *Advances in Neural Information Processing Systems*, volume 36, pp. 53728–53741, 2023. doi: 10.52202/075280-2338. URL https://papers.nips.cc/paper_files/paper/2023/hash/a85b405ed65c6477a4fe8302b5e06ce7-Abstract-Conference.html.

John Schulman, Sergey Levine, Pieter Abbeel, Michael Jordan, and Philipp Moritz. Trust region policy optimization. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 1889–1897. PMLR, 2015. URL https://proceedings.mlr.press/v37/schulman15.html.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017. URL https://arxiv.org/abs/1707.06347.

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

Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. Scaling LLM Test-Time Compute Optimally Can be More Effective than Scaling Parameters for Reasoning. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=4FWAwZtd2n.

Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F. Christiano. Learning to summarize with human feedback. In *Advances in Neural Information Processing Systems*, volume 33, pp. 3008–3021, 2020. URL https://papers.nips.cc/paper_files/paper/2020/hash/1f89885d556929e98d3ef9b86448f951-Abstract.html.

Wangtao Sun, Xiang Cheng, Xing Yu, Haotian Xu, Zhao Yang, Shizhu He, Jun Zhao, and Kang Liu. Probabilistic uncertain reward model. *arXiv preprint arXiv:2503.22480*, 2025. URL https://arxiv.org/abs/2503.22480.

Richard S. Sutton, David McAllester, Satinder Singh, and Yishay Mansour. Policy Gradient Methods for Reinforcement Learning with Function Approximation. In *Advances in Neural Information Processing Systems*, volume 12, 1999. URL https://papers.nips.cc/paper_files/paper/1999/hash/464d828b85b0bed98e80ade0a5c43b0f-Abstract.html.

Yunhao Tang, Kunhao Zheng, Gabriel Synnaeve, and Remi Munos. Optimizing language models for inference time objectives using reinforcement learning. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 59066–59085. PMLR, 2025. URL https://proceedings.mlr.press/v267/tang25o.html.

Jens Tuyls, Dylan J. Foster, Akshay Krishnamurthy, and Jordan T. Ash. Representation-Based Exploration for Language Models: From Test-Time to Post-Training. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=S4PCF1YxoR.

Christian Walder and Deep Karkhanis. Pass@K policy optimization: Solving harder reinforcement learning problems. *arXiv preprint arXiv:2505.15201*, 2025. URL https://arxiv.org/abs/2505.15201.

Haoxiang Wang, Wei Xiong, Tengyang Xie, Han Zhao, and Tong Zhang. Interpretable preferences via multi-objective reward modeling and mixture-of-experts. *arXiv preprint arXiv:2406.12845*, 2024a. URL https://arxiv.org/abs/2406.12845.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-Consistency Improves Chain of Thought Reasoning in Language Models. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=1PL1NIMMrw.

Zhilin Wang, Yi Dong, Olivier Delalleau, Jiaqi Zeng, Gerald Shen, Daniel Egert, Jimmy J. Zhang, Makesh Narsimhan Sreedhar, and Oleksii Kuchaiev. HelpSteer2: Open-source dataset for training top-performing reward models. *arXiv preprint arXiv:2406.08673*, 2024b. URL https://arxiv.org/abs/2406.08673.

Ronald J. Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning*, 8(3–4):229–256, 1992. doi: 10.1023/a:1022672621406. URL https://doi.org/10.1023/a:1022672621406.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024. URL https://arxiv.org/abs/2412.15115.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph Gonzalez, and Ion Stoica. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. In *Advances in Neural Information Processing Systems*, volume 36, pp. 46595–46623, 2023. doi: 10.52202/075280-2020. URL https://papers.nips.cc/paper_files/paper/2023/hash/91f18a1287b398d378ef22505bf41832-Abstract-Datasets_and_Benchmarks.html.

Appendix roadmap. Appendix A gives the ordinal parameterization and proofs of the exact objective, calibration, gradient, variance, and rollback claims. Appendix B records the complete experimental protocol and diagnostics; Appendix C gives implementation conventions for Algorithm 1; machine-readable hashes and verification records are summarized at the end of Appendix C.

A Detailed theoretical analysis

A.1 Ordinal-probit parameterization

Write $\Delta(\mathcal{R})$ for the probability simplex on $\mathcal{R} = \{0, 1, 2, 3, 4\}$. The frozen Moment RM emits a latent location $m_\omega(x, y)$ and positive scale $s_\omega(x, y)$. With frozen cutpoints $c_1 < \cdots < c_4$, its five-category probabilities are $p_{\omega,0} = \Phi((c_1 - m_\omega)/s_\omega)$, $p_{\omega,r} = \Phi((c_{r+1} - m_\omega)/s_\omega) - \Phi((c_r - m_\omega)/s_\omega)$ for $r = 1, 2, 3$, and $p_{\omega,4} = 1 - \Phi((c_4 - m_\omega)/s_\omega)$. The moments used by the policy are $\mu_\omega = \sum_{r=0}^{4} r p_{\omega,r}$ and $\sigma_\omega^2 = \sum_{r=0}^{4} (r - \mu_\omega)^2 p_{\omega,r}$. Repeated-rating likelihood and disagreement supervise the conditional scale; therefore this variance is a model-based signal of annotation disagreement, not epistemic uncertainty or a stand-alone reward. Proper scoring and calibration are required because the complete categorical distribution enters the Best-of-N objective (McCullagh, 1980; Kendall & Gal, 2017; Lou et al., 2024; Sun et al., 2025; Gneiting & Raftery, 2007; Guo et al., 2017).

This parameterization is included to make the evaluator interface reproducible, not to claim a new ordinal-link model. The methodological choice is what the policy does with the frozen output: the same five probabilities support the nominal mean, threshold products, robust envelope, and disagreement diagnostic. Keeping these uses separate is important because replacing the vector by its mean would make the Best-of-N objective and its calibration check unidentifiable.

Proposition 4 (Finite-support threshold decomposition)**.** *For conditionally independent ratings supported on $\mathcal{R} = \{0, 1, 2, 3, 4\}$, the identity $h_N^{\mathrm{nom}} = \sum_{t=1}^{4} [1 - \prod_{j=1}^{N} F_{\omega,j}(t-1)]$ is exact, lies in $[0, 4]$, and is nondecreasing under first-order stochastic improvement of any candidate. Its highest-threshold component is $1 - \prod_j (1 - p_{\omega,4}(x, Y_j))$.*

The proposition is the finite-support sanity check for the main construction. It also explains why the top-rating diagnostic in the artifact is not a substitute for the reported objective: it is only the $t = 4$ term, whereas the method trains against all four threshold events. Reporting the complete sum also exposes contributions from intermediate thresholds that the top-rating diagnostic alone omits.

Proof. For any integer-valued $Z \in \{0, \ldots, 4\}$, the tail-sum identity gives $\mathbb{E}[Z] = \sum_{t=1}^{4} \Pr\{Z \geq t\}$. Taking $Z = \max_j R_j$ and using (A1) gives $\Pr\{\max_j R_j < t \mid x, Y_{1:N}\} = \prod_{j=1}^{N} F_{\omega,j}(t-1)$.

Substitution proves the stated nominal identity. Each term lies in $[0, 1]$, so the sum lies in $[0, 4]$. First-order stochastic improvement decreases a candidate CDF and therefore cannot decrease any summand. At $t = 4$, $F_{\omega,j}(3) = 1 - p_{\omega,4}(x, Y_j)$, giving the highest-threshold expression above. □

Proposition 5 (Prompt-cluster population calibration)**.** *Let $G_{\mathrm{cal}} \geq 1$ validation prompt clusters be i.i.d. from a fixed cluster distribution. The entire evaluator, including checkpoint selection, cutpoints and any fitted temperature, must be fixed independently of these clusters. For cluster g, let $n_g \geq 1$ be its number of ratings, $R_{g\ell} \in \{0, 1, 2, 3, 4\}$ its ratings, and $F_{\omega,g\ell}(t)$ the corresponding evaluator CDF. Dependence among ratings within a cluster is unrestricted. For $t = 0, \ldots, 3$, define*

$$D_{g,t} = \frac{1}{n_g} \sum_{\ell=1}^{n_g} \{\mathbf{1}[R_{g\ell} \leq t] - F_{\omega,g\ell}(t)\}, \qquad \widehat{d}_t = \frac{1}{G_{\mathrm{cal}}} \sum_{g=1}^{G_{\mathrm{cal}}} D_{g,t}.$$

Conditional on the independently fitted evaluator, write $d_t = \mathbb{E}_g[D_{g,t}]$ for the population cluster mean. For $\alpha_{\mathrm{cal}} \in (0, 1)$, set

$$\varepsilon_{\mathrm{cal}} = \min\left\{1, \max_{t=0,\ldots,3} (\widehat{d}_t)_+ + \sqrt{\frac{2\log(4/\alpha_{\mathrm{cal}})}{G_{\mathrm{cal}}}}\right\}. \tag{7}$$

Then, over the random validation clusters, with probability at least $1 - \alpha_{\mathrm{cal}}$, $d_t \leq \varepsilon_{\mathrm{cal}}$ simultaneously for all four thresholds.

Proof. Each residual $D_{g,t}$ lies in $[-1, 1]$, an interval of width two. For fixed t, independence across clusters and Hoeffding's inequality (Hoeffding, 1963) therefore give, for $u > 0$,

$$\Pr\{d_t - \widehat{d}_t > u \mid \omega\} \le \exp(-G_{\mathrm{cal}} u^2/2).$$

Taking $u = \sqrt{2\log(4/\alpha_{\mathrm{cal}})/G_{\mathrm{cal}}}$ makes this probability at most $\alpha_{\mathrm{cal}}/4$. A union bound over four thresholds yields $d_t \le \widehat{d}_t + u$ simultaneously. Replacing $\widehat{d}_t$ by the largest positive part only increases the upper bound. Clipping at one remains valid because $d_t \le 1$. The conditional bound holds for every independently fitted evaluator and hence also after averaging over its training randomness. □

Scope and recorded experimental radius. This is a prompt-weighted population-average statement. It neither bounds individual response CDF errors nor transfers automatically to policy-generated responses. Freezing a model after fitting it on the same validation labels does not satisfy the independence premise; a fresh calibration split is required. Conditioning on validation responses to obtain width-one intervals would change the target to a fixed-design conditional mean and does not prove the population statement above.

The recorded experiments used the earlier rule

$$\varepsilon_{\mathrm{run}} = \min\left\{1, \max_t(\widehat{d}_t)_+ + \sqrt{\frac{\log(4/\alpha_{\mathrm{cal}})}{2G_{\mathrm{cal}}}}\right\},$$

with recorded value 0.04680826120821986. Its concentration term is half that in eq. (7); we make no population-confidence claim for this recorded radius. All empirical tables and figures retain this fixed value. The exact ambiguity-set value, credit, radius monotonicity and conditional-coverage results hold for every fixed $\varepsilon \in [0, 1]$, including $\varepsilon_{\mathrm{run}}$. They do not require that it was obtained from the population calibration estimator. Training or evaluating at a new radius would define a different experiment and is not reported here.

Proof of Theorem 1. The envelope $\overline{F}_j$ is nondecreasing, has boundary values zero and one, and therefore defines a valid categorical distribution. For every feasible $q_{1:N} \in \mathfrak{U}_\varepsilon(p_{\omega,1:N})$, we have

$$F_{q_j}(t) \le \overline{F}_j(t), \qquad t = 0, 1, 2, 3.$$

Hence q_j first-order stochastically dominates $\overline{p}_j$, or equivalently $\overline{p}_j$ is the stochastically smallest distribution in the ambiguity set. Since the maximum is coordinatewise nondecreasing in its ratings, replacing every feasible candidate distribution by $\overline{p}_j$ cannot increase its expectation. The envelope itself is feasible and thus attains the infimum. Applying Proposition 4 gives the stated closed form. □

Proof of Theorem 2. For nonnegative integer ratings, the positive-part marginal credit can be decomposed as

$$(\overline{R}_i - \overline{M}_{-i})_+ = \sum_{t=1}^{4} \mathbf{1}\{\overline{R}_i \ge t,\ \overline{M}_{-i} < t\}.$$

Taking conditional expectations and using the conditional independence assumption (A1) gives

$$\Pr\{\overline{R}_i \ge t,\ \overline{M}_{-i} < t \mid x, Y_{1:N}\} = [1 - \overline{F}_i(t-1)] \prod_{j \ne i} \overline{F}_j(t-1),$$

which proves Equation (4).

For the gradient identity, fix i and condition on (x, Y_{-i}). The robust group value decomposes as

$$h_N^{\mathrm{rob}} = \mathbb{E}[\overline{M}_{-i} \mid x, Y_{-i}] + q_i^{\mathrm{rob}}.$$

The first term is independent of Y_i, and the score-function identity implies that its product with $\nabla_\theta \log \pi_\theta(Y_i \mid x)$ has conditional expectation zero. Applying the likelihood-ratio derivative to all N i.i.d. responses and summing the remaining terms proves Equation (5). □

Proof of Theorem 3. Under (A1), let F_j^* denote the true conditional CDF. Under the pointwise coverage premise $F_j^*(t) \le \overline{F}_j(t)$ for every candidate and threshold, first-order stochastic dominance directly yields the lower-bound inequality stated in Theorem 3. The same independence assumption permits the product comparison used for the misspecified case.

For the misspecified case, define

$$\xi \triangleq \max_{j,t} [F_j^*(t) - \overline{F}_j(t)]_+.$$

At each threshold, the true and robust tail-sum terms can differ only through the corresponding CDF products. For arbitrary $a_j, b_j \in [0,1]$, the product difference is bounded by the sum of positive differences:

$$\prod_j a_j - \prod_j b_j \le \sum_j (a_j - b_j)_+.$$

Applying this inequality with $a_j = F_j^*(t-1)$ and $b_j = \overline{F}_j(t-1)$ at each of the four thresholds, the discrepancy is at most $N\xi$ per threshold, and therefore at most $4N\xi$ after summation. For the fixed group this gives

$$\mathbb{E}_*[\max_j R_j \mid x, Y_{1:N}] \ge h_N^{\mathrm{rob}}(x, Y_{1:N}) - 4N\xi(x, Y_{1:N}).$$

If prompts and responses are also averaged, both sides must be averaged: $\mathbb{E}_{x,Y}\mathbb{E}_*[\max_j R_j \mid x, Y] \ge J_N^{\mathrm{rob}}(\theta) - 4N\mathbb{E}_{x,Y}[\xi(x,Y)]$. □

Proof of Proposition 1. Fix a candidate group and threshold t, and define

$$a_j(\varepsilon) \triangleq \min\{1,\ F_{\omega,j}(t-1) + \varepsilon\}.$$

For $0 \le \varepsilon_1 \le \varepsilon_2 \le 1$, the clipping map is nondecreasing and 1-Lipschitz, which gives

$$0 \le a_j(\varepsilon_2) - a_j(\varepsilon_1) \le \varepsilon_2 - \varepsilon_1.$$

Using the telescoping product identity and the fact that $a_j(\varepsilon) \in [0,1]$, we have

$$0 \le \prod_j a_j(\varepsilon_2) - \prod_j a_j(\varepsilon_1) \le \sum_j \big(a_j(\varepsilon_2) - a_j(\varepsilon_1)\big) \le N(\varepsilon_2 - \varepsilon_1).$$

Substituting this bound into the four-term tail-sum expression for h_N^{rob}, and then taking expectations over prompts and responses, proves Proposition 1.

At $\varepsilon = 0$, the robust envelope equals the nominal CDF, so the exact infimum in Theorem 1 equals $J_N^{\mathrm{nom}}(\theta)$. □

Proof of Proposition 2. By the conditional credit definition in Equation (4), we have

$$\mathbb{E}[(\overline{R}_i - \overline{M}_{-i})_+ \mid x, Y_{1:N}] = q_i^{\mathrm{rob}},$$

for every i. The score terms ϕ_i are fixed under this conditioning, so linearity gives

$$\mathbb{E}[G_{\mathrm{samp}} \mid x, Y_{1:N}] = G_{\mathrm{RB}}.$$

For any fixed vector v, the law of total variance yields

$$\mathrm{Var}(v^\top G_{\mathrm{samp}}) = \mathbb{E}\big[\mathrm{Var}(v^\top G_{\mathrm{samp}} \mid x, Y_{1:N})\big] + \mathrm{Var}(v^\top G_{\mathrm{RB}}),$$

which directly implies the stated variance inequality. □

Proof of Proposition 3. Let $(\theta_0, \mathsf{Opt}_0)$ denote the snapshot taken before a tentative update. If the measured rollout-state KL exceeds δ_{hard}, the algorithm executes the restoration $(\theta, \mathsf{Opt}) \leftarrow (\theta_0, \mathsf{Opt}_0)$. By the semantics of the implementation's state-copy operation, both objects are restored exactly to their pre-update values. Because controller-state updates occur only in the accepted branch, a rejected transition also leaves the baseline, dual variables, KL coefficient, and RMS scale unchanged. This statement is conditional on the measured batch and therefore makes no population-level KL claim. □

Proposition 6 ($N = 1$ reduction and complexity)**.** *For $N = 1$,*

$$h_1^{\mathrm{rob}} = q_1^{\mathrm{rob}} = \sum_{r=0}^{4} r\overline{p}_{1,r}.$$

Thus the upper-tail reward component reduces exactly to the robust mean reward. For general N, all group values and marginal credits can be computed in $O(N|\mathcal{R}|)$ time using threshold-wise prefix and suffix products.

This reduction serves as a useful implementation sanity check: at $N = 1$, the method must recover the robust mean and cannot manufacture a group advantage. For $N > 1$, the additional terms exactly capture the response's incremental chance of opening a new threshold above the other candidates. The test suite validates both regimes, ensuring that any improvement in the group objective cannot be attributed to a special-case implementation artifact.

Proof. With a single candidate, the empty opponent maximum is zero by convention, so $(\overline{R}_1 - 0)_+ = \overline{R}_1$, and the first equality follows directly. For the complexity analysis, note that evaluating each of the four nontrivial thresholds requires one forward and one backward product scan over the N candidates. Hence the total time is $O(N \cdot |\mathcal{R}|)$, where $|\mathcal{R}| = 5$. □

B Full experimental protocol

B.1 Experimental questions

The experiments are designed to answer four predeclared questions:

1. Does StableMax-PPO attain a competitive held-out robust $\mathbb{E}[\max_{j \le 32} R_j]$ objective relative to mean optimization and a direct scalar max@32 policy gradient, while reducing held-out KL?
2. Are gains preserved under the nominal exact ordinal maximum and the top-rating probability $\Pr\{\exists j : R_j = 4\}$?
3. Do nominal expected rating, independent Quality RM score, and reference-policy KL remain within their predeclared safeguards?
4. Do independent generated-output proxies and existing human-labeled preference diagnostics provide evidence against target-RM exploitation?

No result is used to remove a comparator, redefine a threshold, or select a more favorable protocol.

B.2 Protocol audit and scientific interpretation

The artifact verifier passed for seeds 314, 2718, and 1618. The terminal record preserves every predeclared proxy, non-inferiority, safeguard, and KL outcome for auditability. We interpret these fields as protocol diagnostics that characterize the measured operating points, not as a universal method ranking or a proof that reward-model exploitation is absent.

The diagnostic vector is read as a decision record rather than a single leaderboard score. A method can satisfy KL and quality safeguards while losing a max comparison, or improve the target reward while degrading an independent proxy. Reporting the vector preserves these distinctions and prevents a favorable endpoint from hiding a failure mode that the method was designed to measure.

The aggregate terminal record reports `scientific_success=false`: the mean and independent-quality floors and the held-out KL comparisons against Mean-PPO and GRPO pass, but the declared robust-maximum gates and Qwen proxy gate do not all pass. In particular, the paired robust-maximum difference from Mean-PPO is 0.0339 with a 95% interval of $[-0.0686, 0.1304]$. The corresponding intervals against GRPO and Scalar max-PPO are $[-0.1336, -0.0609]$ and $[-0.1024, -0.0101]$. Passing artifact verification therefore establishes protocol consistency, not scientific success.

The recorded runs fix $\varepsilon_{\mathrm{run}} = 0.04680826120821986$ using the earlier calibration rule with $\alpha_{\mathrm{cal}} = 0.05$. As explained in Appendix A, this is a fixed ambiguity-set parameter, not a certified population

calibration radius. We retain all recorded outcomes and do not substitute the corrected formula into completed runs. The released summaries record the radius; the original cluster-level calibration data and report remain external, so their sample count and split independence cannot be reconstructed from this release.

B.3 Models, data, and split discipline

The policy is Qwen2.5-1.5B-Instruct (Yang et al., 2024). The frozen Moment RM is a Qwen2.5-14B ordinal model trained on HelpSteer2 repeated helpfulness ratings, with the decontaminated preference replay recorded in its training manifest (Wang et al., 2024b). It exports $p_{\omega,0:4}$ directly; exposing the category probabilities does not retrain or alter its accepted checkpoint. An independently trained Qwen2.5-14B scalar Quality RM, fit separately on disjoint preference data, supplies Q_η and has no variance head. Its frozen 4,096-pair confirmation accuracy is 94.60% with a Wilson lower 95% bound of 93.87%.

All policy, reward-model, evaluator, and lockbox splits are content-hashed. Before partitioning RewardBench, we remove prompt and response overlap with the policy and Quality-RM data (Lambert et al., 2024). The remaining calibration, policy-diagnostic, and reserve partitions are also mutually prompt- and response-disjoint. They contain 1,024, 1,024, and 427 pairs, respectively. The sealed Skywork diagnostic contains 4,096 pairs and is excluded from Quality-RM training and selection (Liu et al., 2024).

The Moment RM and policy prompts remain drawn from the same broad HelpSteer2 domain. Disjoint rows prevent direct leakage but do not establish out-of-domain generalization; this is treated as a limitation rather than hidden by the split terminology.

B.4 Frozen training protocol

The primary setting is $N = 32$. Each trainable confirmatory method uses three seeds $\{314, 2718, 1618\}$, 300 policy steps, two prompts per step, 32 responses per prompt, learning rate 2×10^{-5}, one PPO epoch, clip 0.2, LoRA rank 8 with scale 32, sampling temperature 1, $\text{top}_p = 1$, maximum prompt length 384, and maximum response length 64. Seed 42 is used only for development and real-GPU smoke tests and is excluded from all confirmatory intervals, gates, and aggregate claims. The target KL is 0.02 and the hard rollback level is 0.04. All methods share the same prompt stream, value of N, optimizer class, policy architecture, and checkpoint-selection rule.

The non-inferiority floors are frozen before any trainable policy outcome is observed. The reference sample contains 2,048 base-policy responses from 512 independent policy-training prompts, with four responses per prompt. For $Z \in \{\mu_\omega, Q_\eta\}$, let $\bar{Z}_g$ be the four-response prompt mean and let L_Z be the fifth percentile of 10,000 bootstrap means obtained by resampling prompts. The frozen floor is $z_0 = L_Z - 0.1 s_Z$, where s_Z is the response-level standard deviation in the same reference sample. The bootstrap unit, one-sided level, 10,000 draws, sample counts, generation seed, bootstrap seed, and practical margin are all predeclared. A separate prompt-disjoint sample of 512 responses from 128 prompts reports standardized reference-policy shift. It is diagnostic only and cannot accept or reject training.

B.5 Comparators and ablations

The ten trainable methods are:

1. **StableMax-PPO**: the complete method in Algorithm 1.
2. **Mean-PPO + Q**: PPO on μ_ω with the same mean, Quality, and KL protocol.
3. **GRPO + Q**: group-relative standardized μ_ω credit with matched safeguards (Shao et al., 2024; DeepSeek-AI et al., 2025).
4. **Scalar Max@32-PPO + Q**: exact leave-one-out scalar max credit $[\mu_i - \max_{j \neq i} \mu_j]_+$ with matched safeguards (Tang et al., 2025; Walder & Karkhanis, 2025; Bagirov et al., 2025).
5. **Entropic-PPO + Q**: PPO on $\text{CE}_\tau = -\tau \log \sum_r p_{\omega,r} e^{-r/\tau}$, using $\tau = 1$.
6. **Nominal Ordinal EV-PPO + Q**: exact J_{32}^{nom} without the CDF ambiguity set.

Table 2: Complete aggregate table. Values are three-seed robust expected maximum, with 95% confidence intervals, held-out KL, and rollback counts.

Method	Robust max	95% CI	Held-out KL	Rollbacks/900
StableMax-PPO	3.458	[3.411, 3.506]	0.034	279 (31.0%)
Mean-PPO	3.424	[3.354, 3.517]	0.070	527 (58.6%)
GRPO	3.555	[3.530, 3.578]	0.094	741 (82.3%)
Scalar max-PPO	3.517	[3.476, 3.558]	0.054	285 (31.7%)
Entropic PPO	3.400	[3.353, 3.447]	0.037	302 (33.6%)
Nominal EV-PPO	3.421	[3.377, 3.467]	0.028	121 (13.4%)
StableMax-PPO (no mean floor)	3.508	[3.426, 3.574]	0.048	230 (25.6%)
StableMax-PPO (no quality floor)	3.467	[3.402, 3.530]	0.048	329 (36.6%)
Gaussian EV-PPO	3.558	[3.533, 3.581]	0.061	367 (40.8%)
Top-4 PPO	3.418	[3.353, 3.499]	0.042	277 (30.8%)
Best-of-32 (base)	3.086	[3.045, 3.125]	–	–

Table 3: Per-seed primary endpoints at $N = 32$. Each cell reports robust expected maximum / held-out KL on the same 512-prompt evaluation split. The KL values are the common-base-trajectory means from the held-out evaluator, matching the aggregate endpoint in Table 2.

Method	Seed 314	Seed 2718	Seed 1618
StableMax-PPO	3.503 / 0.057	3.452 / 0.024	3.420 / 0.022
Mean-PPO	3.369 / 0.070	3.523 / 0.074	3.381 / 0.066
GRPO	3.563 / 0.094	3.547 / 0.097	3.553 / 0.091
Scalar max-PPO	3.511 / 0.052	3.555 / 0.080	3.484 / 0.031

7. **StableMax-PPO without mean floor**: ablates nominal-mean non-inferiority.
8. **StableMax-PPO without Quality RM**: ablates the independent quality safeguard and is excluded from constrained-feasibility claims.
9. **Gaussian EV-PPO + Q**: the previous heteroscedastic Gaussian expected-maximum objective, retained as a model-assumption ablation.
10. **Top-4 PPO + Q**: exact optimization of $\Pr\{\exists j : R_j = 4\}$, isolating the highest threshold from the complete ordinal maximum.

An eleventh, evaluation-only **Best-of-32** row samples 32 responses from the untrained base policy and ranks them by μ_ω. It is an inference-time reference, not a compute-matched training baseline; its generation and scoring cost is reported separately.

The comparator set is organized to separate three changes that are otherwise easy to conflate: the deployment functional (mean, scalar maximum, top-rating, or full ordinal maximum), the CDF robustification, and the stabilizers. The nominal ordinal and Gaussian variants hold the policy optimizer fixed while changing the evaluator assumption; the no-floor variants hold the target fixed while removing one safeguard. This is why the ablations support a controlled interpretation of the observed operating points without claiming that any single component is universally beneficial.

The per-seed view makes the operating-point claim testable without relying on the aggregate interval alone: StableMax-PPO has lower held-out KL than Mean-PPO and GRPO in every seed, and lower KL than Scalar max-PPO in two of the three seeds (2718 and 1618). Its robust maximum remains competitive across the same paired prompts. The robust values are the checked-in per-seed CSV records and the KL values are the held-out common-base means; all entries are rounded only for display, with no additional aggregation or seed selection performed here.

B.6 Ablation and stability analysis

The component rows in Table 2 isolate the operating point controls. Removing the mean floor increases the aggregate robust maximum from 3.458 to 3.508, but also raises held-out KL from 0.034 to 0.048. Removing the independent Quality floor gives $(3.467, 0.048)$, while removing CDF robustification gives $(3.421, 0.028)$, in (robust maximum, KL) coordinates. The three-seed intervals

overlap and the per-seed effects are heterogeneous, so these are sensitivity comparisons rather than isolated causal effects. CDF-envelope removal changes the evaluator-defined target; floor removal changes the accepted-update trajectory. The records therefore support a controlled operating-point interpretation, not a claim that any safeguard is a universally improving reward term.

B.7 Evaluation endpoints

Every method is evaluated on 512 frozen test prompts with newly generated groups of 32 responses. The primary paired prompt-level endpoint is h_{32}^{rob}, exactly matching the StableMax-PPO training objective. Secondary model-based endpoints are h_{32}^{nom}, U_{32}^{nom}, candidate-level μ_ω, ordinal variance, $p_{\omega,4}$, independent Quality score, reference KL, constraint residuals, accepted updates, and exact rollback counts.

The confirmatory evaluation fixes $N = 32$ for every method; no cross-N sensitivity result is claimed. The group-level maximum uses all 32 candidates and does not require a selected response. A deployment-style $\arg\max_i \mu_\omega(x, Y_i)$ response is reported only as a secondary selector diagnostic. Held-out KL is computed on common base trajectories, so the paired comparison measures policy motion under the same prompt protocol rather than differences in test-set difficulty.

For completeness, the top-category diagnostic used in the cached evaluation is

$$U_N^{\mathrm{nom}}(\theta) := \mathbb{E}_{x,Y_{1:N}}\left[1 - \prod_{j=1}^{N}(1 - p_{\omega,4}(x, Y_j))\right]. \tag{8}$$

It is the frozen-model probability that at least one candidate receives rating four. It is not a Gaussian tail approximation and is not the primary estimand; the complete ordinal maximum remains the reported objective.

B.8 Statistical analysis and success criteria

All seeds are reported separately. The frozen analysis uses 10,000 draws of a seed-then-prompt percentile bootstrap: sample three seeds with replacement, then independently resample 512 prompt indices within each sampled seed. Paired method contrasts use within-seed prompt differences before resampling. The resulting 2.5th and 97.5th percentiles are the reported intervals. Because prompts are shared across seeds but resampled independently within them, these are intervals under the frozen nested-resampling convention, not an exact crossed seed–prompt population confidence guarantee. Three seeds further limit precision of inference over training randomness.

The recorded success gates require the robust-max difference interval's lower endpoint to be > 0 against Scalar max-PPO, ≥ -0.05 against Mean-PPO, and ≥ -0.10 against GRPO. They also require the held-out-KL difference interval's upper endpoint to be < 0 against both Mean-PPO and GRPO. The candidate nominal-mean and Quality-score interval lower endpoints must reach their respective frozen floors. Each generated-output proxy must have tie-adjusted preference mean ≥ 0.50 and interval lower endpoint ≥ 0.45. All nine gates must pass jointly; the recorded decision is false. These are the unchanged protocol criteria, not thresholds chosen after seeing the results. We make no superiority claim from the positive Mean-PPO robust-max point estimate.

Additional diagnostics average paired differences over seeds within each prompt before sign-flip randomization. Such tests require sign-exchangeability under their null. The historical output key `paired_rank_biserial` actually records the sign-based effect $(n_+ - n_-)/(n_+ + n_-)$, excluding zero differences (defined as zero if all differences vanish); it does not weight signed ranks. We retain that field name for record compatibility and interpret it as a sign-dominance statistic. Individual intervals and auxiliary tests are descriptive and are not simultaneous confidence guarantees across all methods or endpoints.

B.9 Anti-exploitation and human-grounded diagnostics

Newly generated outputs are compared with Mean-PPO outputs by two frozen evaluators: the Qwen2.5-14B-Instruct judge and `RLHFlow/ArmoRM-Llama3-8B-v0.1` (Zheng et al., 2023; Wang et al., 2024a). For each seed, the diagnostic aligns up to 512 held-out prompt groups, retains the

response selected by $\arg\max_i \mu_\omega(x, Y_i)$ from each 32-response group, and evaluates the matched pair. Qwen judgments use randomized presentation order and a reverse order check; ArmoRM supplies an independent pointwise score. Each evaluator must first pass its frozen RewardBench calibration gate (between 512 and 1,024 calibration pairs, with the frozen split and hashes recorded by the protocol). Pairwise judgments are model-based proxies, not human labels. The complete group-level diagnostic is run separately on all 32 candidates and reports robust/nominal maximum, probability of any rating-4 response, and the paired group delta; selected-response diagnostics additionally report length, character count, repeated-trigram rate, unique-word ratio, refusal behavior, and upper-tail strata.

Table 4: Independent generated-output proxy comparisons against Mean-PPO. Preference is tie-adjusted, so 0.5 denotes parity; intervals are three-seed hierarchical-bootstrap 95% intervals.

Evaluator	Preference	95% CI
Qwen2.5-14B judge	0.494	[0.478, 0.510]
ArmoRM-Llama3-8B-v0.1	0.502	[0.458, 0.544]

Both point estimates are near parity and both intervals contain 0.5. These intervals establish neither a quality difference nor equivalence to Mean-PPO; in particular, the Qwen proxy does not pass its declared gate. A target-RM increase accompanied by proxy degradation would be evidence consistent with reward-model exploitation; these results do not show that signature, but the model-based design cannot establish its absence.

RewardBench and Skywork chosen–rejected lockboxes contain previously collected human labels. For each policy, length-normalized chosen–rejected policy log-probability accuracy and margin are reported as preference-retention diagnostics. They are not Best-of-32 endpoints and do not label newly sampled policy responses. In particular, a new response cannot be declared human preferred merely because it was generated for a prompt that also appears in a fixed preference benchmark.

No new human annotation is collected in this protocol. Therefore, even if all anti-exploitation gates pass, the permitted claim is *no evidence of exploitation under the measured model-based and existing-label diagnostics*. The experiment cannot establish absence of reward hacking or direct human preference for generated outputs. A blinded group-rating packet is emitted so that fresh human evaluation can be added later without regenerating or selecting outputs.

B.10 Reproducibility and reporting

Code, model adapters, data partitions, floor samples, calibration reports, and protocol files are content-hashed before training. A real-GPU smoke test must execute every declared objective, verify finite applicable metrics, verify constraint activation, and test exact parameter and optimizer rollback. Formal training and evaluation run only after these checks pass. Technical failures are versioned and corrected with regression tests. Unfavorable scientific outcomes are reported without changing seeds, thresholds, comparators, evaluators, or success criteria.

The 1.5B policy is evaluated by 14B reward and proxy models. This capacity gap may improve discrimination but may also create systematic evaluator shift. The three-seed design measures optimization variability; it does not replace replication at other model scales or fresh blinded labels.

C Implementation Details for StableMax-PPO

This appendix records implementation conventions for Algorithm 1, rather than a second algorithm. Appendix B specifies models, seeds, and thresholds; the machine-readable supplement provides aggregate and per-seed tables, terminal records, and SHA-256 hashes.

For each prompt x_b, the implementation caches behavior-policy token log probabilities and response masks before the tentative policy update. The five ordinal probabilities are converted to prefix CDFs. For $t = 0, \ldots, 3$, the robust envelope is $\bar{F}_{b,i}(t) = \min\{1, F_{\omega,b,i}(t) + \varepsilon\}$, with $\bar{F}_{b,i}(-1) = 0$ and $\bar{F}_{b,i}(4) = 1$. The robust probabilities are adjacent differences of this envelope. Prefix and suffix products compute all group values and leave-one-out credits in $O(BN|\mathcal{R}|)$ time without enumerating rating tuples.

The constrained credit used by the PPO surrogate is $\tilde{q}_{b,i} = q^{\mathrm{rob}}_{b,i} + \lambda_m \mu_\omega(x_b, Y_{b,i})/N + \lambda_q Q_\eta(x_b, Y_{b,i})/N$, with $A_{b,i} = N[\tilde{q}_{b,i} - b_\psi(x_b)]$. The prompt baseline is fitted only from accepted transitions. Dual variables use projected updates after acceptance; rejected steps leave all controller state unchanged. The lagged RMS scale is computed from accepted credits only, so a rejected outlier cannot change the next normalization factor.

Before the optimizer step, the implementation snapshots all trainable policy parameters and the complete optimizer state, including momentum and adaptive variance buffers. It evaluates full-categorical token KL on the same rollout states after the tentative step. If the hard threshold is exceeded, both snapshots are restored and no baseline, dual, KL-controller, or RMS state is advanced. This is an exact batch-level rollback invariant; it does not imply a population KL bound.

Table 5: Implementation invariants checked by the artifact.

Component	Invariant	Test coverage
Ordinal kernel	Five probabilities form a valid distribution; CDF boundaries and threshold identities hold.	`evrl_ordinal`
Credit kernel	Prefix/suffix products match the exact credit and the $N = 1$ reduction.	`evrl_reduction`
Rollback state	Policy parameters and optimizer buffers are restored together after rejection.	`publication_v8`
Controller state	Baseline, dual, KL, and RMS states advance only on accepted updates.	`publication_v8`
Protocol integrity	Finite metrics, content hashes, and verification records are checked before aggregation.	v8 verification

C.1 Loss normalization, KL estimands, and controller updates

Let B be the number of prompts, $L_{b,i}$ a sampled response length, and $T = \sum_{b,i} L_{b,i}$ the number of action tokens. With $\ell_{b,i,h} = \min\{\rho_{b,i,h} A_{b,i}, \mathrm{clip}(\rho_{b,i,h}, 1-\epsilon_{\mathrm{PPO}}, 1+\epsilon_{\mathrm{PPO}}) A_{b,i}\}$, the minimized loss in the recorded implementation is

$$\mathcal{L}_k = \frac{-\frac{1}{BN}\sum_{b,i}\sum_{h=1}^{L_{b,i}} \ell_{b,i,h} + \frac{\beta_k}{T}\sum_{b,i}\sum_{h=1}^{L_{b,i}} k_\theta(s_{b,i,h})}{s_k + 10^{-8}}.$$

The policy term sums tokens within a response and averages responses; the KL term averages all action tokens. States are sampled from π_{old}, and their distribution is held fixed during the surrogate step. Thus the penalty is not an unbiased gradient of the population $K(\theta)$. The post-step rollback statistic uses these same states with updated policy probabilities. Held-out KL instead averages token KL within each common base-policy response, then averages those response means within prompts and across prompts and seeds. Its state distribution and length weighting differ from the training statistic, so held-out KL may exceed δ_{hard} without contradicting rollback correctness.

After acceptance, for $Z \in \{\mu_\omega, Q_\eta\}$ and its floor z_0, $\lambda_Z \leftarrow \mathrm{clip}(\lambda_Z + 0.05(z_0 - \widehat{Z}), 0, 20)$, where $\widehat{Z}$ averages the accepted rollout's candidate scores and both duals start at one. Let $c_{b,i} = \widetilde{q}_{b,i} - b_\psi(x_b)$. The lagged scale starts at $s_0 = 1$, with $v_0 = 1$, and updates as $v_{k+1} = 0.99 v_k + 0.01(BN)^{-1}\sum_{b,i} c_{b,i}^2$, $s_{k+1} = \sqrt{v_{k+1} + 10^{-8}}$. The baseline fits the prompt mean of $\widetilde{q}$ by squared loss, with AdamW learning rate 10^{-3} and hidden width 128. Policy and baseline gradient norms are clipped at one. The KL coefficient starts at 0.02, doubles when accepted-batch KL exceeds 1.5δ, halves when it falls below $\delta/1.5$, and is clipped to $[10^{-5}, 10]$. A rejected transition changes none of these states. The recorded protocol uses one PPO epoch per tentative update.

All floating-point reductions use deterministic tensor order where supported. The supplementary code contains tests for ordinal identities, credit reductions, finite metrics, protocol hashes, and rollback restoration.