STAR-GRPO: CANONICAL ANCHORING AND RELIABILITY-FIRST ADVANTAGES AGAINST REPRESENTATION-DEPENDENT REWARD HACKING

Anonymous authors
Paper under double-blind review

ABSTRACT

Reward hacking occurs when policy optimization exploits a brittle reward interface or an overly permissive proxy objective, improving the training score without improving the underlying response quality. This phenomenon is amplified in group-relative policy optimization: an unsupported reward can shift the group baseline and alter the updates of other rollouts, while post-hoc or purely relative weighting cannot represent group-wide uncertainty. We propose *Self-Tuned Anchored Reliability Group-Relative Policy Optimization* (STAR-GRPO), a reliability-first advantage estimator based on paired assessments of the same rollout. STAR separates the quality signal from its learning influence: score disagreement determines rollout reliability, relative reliability enters a self-tuned robust location–scale fit before group normalization, and absolute group reliability attenuates the resulting bounded advantage. The analysis establishes coordinate and second-moment bounds, characterizes exact centering through the weighted location equation, and gives reliability-dependent attenuation guarantees for outlying rewards. We evaluate STAR-GRPO in two complementary reward-hacking regimes. In token-interface exploitation, STAR prevents runaway optimization of the deployed-interface score while improving the canonical quality signal. In rubric-proxy overoptimization for medical reasoning, STAR improves independent semantic evaluation, narrows the proxy–judge discrepancy, and reduces overclaim while optimizing the same task proxy. Together, these results show that reliability-first normalization offers a principled way to limit unsupported reward influence on both group baselines and policy updates, while retaining the task reward as the optimization target.

1 INTRODUCTION

Reinforcement learning from human feedback optimizes reward proxies for human preferences (Christiano et al., 2017; Ouyang et al., 2022; Bai et al., 2022). As optimization proceeds, a policy can improve the proxy without a corresponding improvement in external quality (Skalse et al., 2022; Gao et al., 2023). An auxiliary assessment of the same response can reveal disagreement with the training score. The optimization question is then how to use this information: *how should reliability enter group-relative advantages so that a suspicious reward has limited influence on both its own update and the baseline assigned to other responses?*

Ordinary group-relative policy optimization (GRPO) centers and scales rewards within each prompt group (Shao et al., 2024). This coupling creates two distinct obstacles. First, an unreliable high reward changes the group mean and scale before any post-hoc weight is applied; reducing its own advantage does not undo the advantages already assigned to the other rollouts. Second, relative weighting alone cannot express common low confidence: multiplying every weight by the same constant leaves normalized weights unchanged. An advantage construction should therefore use reliability inside normalization and preserve its absolute magnitude in the update.

STAR-GRPO implements these two requirements while separating *what to optimize* from *how strongly to update*. A prespecified quality path selects or combines the training and anchor scores. Their discrepancy supplies an optimization reliability coefficient. Relative weights enter a pseudo-Huber fit of prompt-specific locations and a minibatch-shared scale; a bounded score then receives

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.

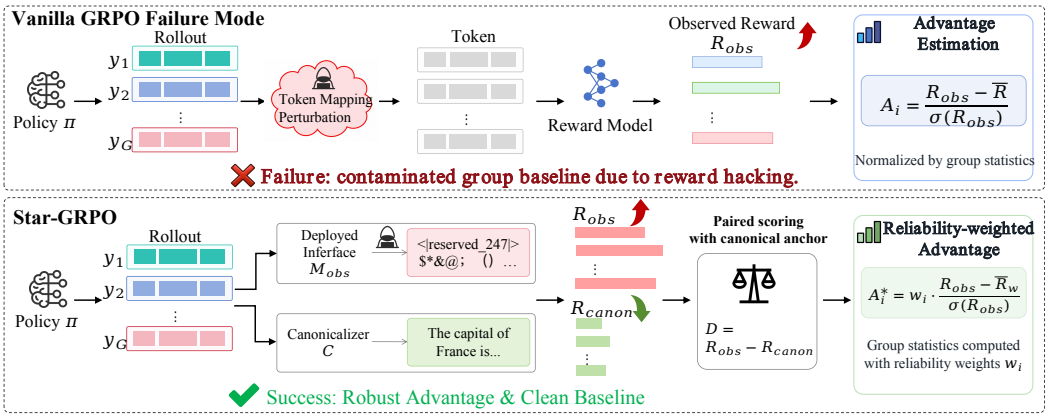


Figure 1: Overview of token-interface reward hacking and STAR-GRPO. A deployed representation and a canonical rendering of the same rollout are scored in parallel; their discrepancy determines reliability, which enters group statistics *before* normalization. The final STAR advantage combines a bounded robust score with rollout- and group-level reliability, preventing unsupported rewards from dominating either the baseline or the policy update.

an absolute group-reliability factor. The resulting advantage satisfies $|A_i| \leq w_i$ for any finite fitted location and positive scale. Exact fitting is needed for centering, not for this magnitude control.

We instantiate this principle in two reward-hacking regimes with different semantics but the same optimization structure. In *token-interface exploitation*, direct token-index mappings can obtain high reward even when the corresponding decoded response is not supported by the canonical text interface (Zhang et al., 2026); STAR pairs the deployed and canonical views and uses their discrepancy to regulate the update. In *rubric-proxy overoptimization*, a rubric-conditioned training judge can reward surface-level criterion satisfaction more strongly than an independent semantic evaluator; STAR keeps the proxy as the optimization target while using a rubric-free semantic anchor to determine how strongly each rollout should influence learning. This separation between the score being optimized and the evidence supporting that score is the central design principle of STAR-GRPO.

Our contributions are threefold:

- We introduce *reliability-first* group-relative advantages: rollout reliability enters the robust group fit before normalization, while an explicit absolute-reliability factor preserves group-wide confidence in the final update.
- We establish deterministic coordinate and second-moment bounds, characterize exact centering through the weighted location equation, and derive reliability-to-gradient attenuation guarantees. A one-outlier analysis shows linear attenuation in the rollout reliability, in contrast to post-hoc weighting and ordinary weighted mean–standard-deviation normalization.
- We demonstrate the same mechanism across two distinct reward-hacking regimes. STAR suppresses representation-dependent token-interface exploitation while improving canonical quality; in rubric-proxy medical reasoning, the independent-judge score rises from 0.2706 to 0.3174, the proxy–judge gap decreases from 0.2935 to 0.2318, and the overclaim fraction decreases from 0.2464 to 0.2204.

Figure 1 summarizes the core idea: paired scoring exposes unsupported reward, reliability enters group statistics before normalization, and the resulting bounded advantage controls both local and group-wide learning influence.

2 Related Work

Reward overoptimization and reward-model limitations motivate assessing progress beyond the training score (Gao et al., 2023; Casper et al., 2023). TOMPA studies token-interface attacks (Zhang

et al., 2026); representation engineering supplies internal shortcut signals for modifying GRPO advantages (Wu and Tang, 2026). STAR instead derives reliability from paired reward assessments of each rollout.

Reward-model ensembles support conservative optimization (Coste et al., 2024) and uncertainty penalties (Zhai et al., 2024); robust reward training addresses preference artifacts (Liu et al., 2024). Conformal Feedback Alignment weights preference optimization using answer-level reliability (Chen et al., 2026). STAR's distinction is to place reliability inside the group fit and retain absolute attenuation afterward. Its robust estimation and calibration tools are established methods, detailed in Appendix A.

Normalization and evaluation choices introduce further sources of bias. Dr. GRPO analyzes response-length and question-difficulty biases in GRPO (Liu et al., 2025), while DAPO studies token-level loss aggregation and training stability (Yu et al., 2025). These works motivate specifying the loss reduction separately from the advantages. LLM judges also exhibit position, verbosity, and self-enhancement biases (Zheng et al., 2023). To separate the training proxy from evaluation, RQ2 uses a third judge that is never used in policy updates, providing an independent measurement of whether proxy optimization transfers to semantic quality.

3 Problem Setup and Threat Model

Both settings supply a pair of scores for the same rollout: a deployed or proxy score and an anchor score. We formalize the token-interface instantiation first, where the same reward model evaluates two representations and the discrepancy has a direct interface interpretation. RQ2 supplies a rubric proxy and a separate semantic anchor to the same advantage construction; its score reference and statistical scope are specified in Section 6.2.

For prompt X_b, the old policy samples G rollouts $O_{b,i} \sim \pi_{\theta_{\text{old}}}(\cdot \mid X_b), b \in [B], \quad i \in [G]$. The deployed mapping M_{obs} and a frozen canonicalizer C produce

$$R_{b,i}^{\text{obs}} = R_\phi(X_b, M_{\text{obs}}(O_{b,i})), \quad R_{b,i}^{\text{canon}} = R_\phi(X_b, C(O_{b,i})), \quad D_{b,i} = R_{b,i}^{\text{obs}} - R_{b,i}^{\text{canon}}. \tag{1}$$

The canonicalizer decodes with the policy tokenizer, applies prespecified normalization and chat-template rules, and retokenizes with the reward tokenizer. Positive D measures deployed-interface reward unsupported by this anchor. Each rollout has a prespecified context $H_{b,i} \in [H]$, such as tokenizer pair, task family, or response-length bin. Trusted benign outputs are divided before score inspection into discrepancy-fitting data $\mathcal{I}_{\text{fit}}^D$ and conformal-calibration data $\mathcal{I}_{\text{cal}}^D$. Policy training, attack development, independent clean audits, and final evaluation use disjoint identifiers. Context definitions, rare-context merges, and unseen-context fallback rules are frozen before calibration. Fix numerical regularizers $\varepsilon_D \geq 0$ and $\varepsilon_w \geq 0$; the positive scale constraints keep all displayed score denominators strictly positive. A fixed $\kappa \in [0, 1]$ determines how much deployed reward enters optimization:

$$R_{\kappa,b,i} = R_{b,i}^{\text{canon}} + \kappa D_{b,i}, \quad r_{b,i} = h(R_{\kappa,b,i}) \in [-R_{\max}, R_{\max}], \tag{2}$$

where h is fixed, monotone, bounded, and 1-Lipschitz. Unless stated otherwise, STAR uses $\kappa = 1$; $\kappa = 0$ selects the canonical reward path. The quality path is part of the specified STAR instantiation. RQ1 additionally reports a canonical-only diagnostic control using the same bounded canonical quality path; all RQ1 reward-model evaluations use a 4,096-token input limit.

The threat model targets reward inflation that is expressed as disagreement between two views of the same rollout, including positive cross-interface shifts, heterogeneous clean discrepancy scales, heavy-tailed quality rewards before h, and groups containing many attacked rollouts. The observed interface must be reachable by the deployed system; arbitrary mappings are reserved for audit stress tests. The following result formalizes why paired views supply information that is unavailable from a single scalar reward.

Theorem 3.1. *(i) For any probability law P on $\mathbb{R}$ and $B_0 > 0$, there are a clean model and an attacked model with the same observable law $R^{\text{obs}} \sim P$, while the attacked canonical reward is smaller by B_0. Hence no detector based only on R^{obs} uniformly distinguishes the two models. (ii) If an attack adds the same random shift to every evaluated view, all pairwise reward differences are unchanged and every discrepancy-only test has type-I plus type-II error at least one.*

The argument in Appendix B shows what a single scalar reward cannot reveal without additional assumptions. Paired views supply evidence about differential shifts; they do not by themselves establish that every observed discrepancy is an attack.

4 Reliability-First Group-Relative Advantages

STAR first specifies the score pair and quality path, then estimates reliability, fits a robust group baseline, and forms policy advantages. We describe the split-calibrated interface instantiation below. RQ2 retains the advantage construction with a fixed discrepancy reference. Algorithm 1 gives the procedure used for the formal analysis; experiment-specific solver and loss choices are documented separately.

4.1 Trusted Discrepancy Calibration

For trusted discrepancies $D_{c,1:n_c}$ in context c, choose $z_{D,c} > 0$ and $0 < s_{\min,c} < s_{\max,c} < \infty$. Following self-tuned robust estimation (Sun, 2024), STAR fits a context location and scale with

$$\ell_{n,z}(u,s) = \frac{ns}{z^2}\left(\sqrt{1+\frac{z^2u^2}{ns^2}}-1\right)+\frac{s}{2}, \tag{3}$$

$$(\widehat{m}_{D,c},\widehat{s}_{D,c}) \in \operatorname*{arg\,min}_{\substack{m\in\mathbb{R}\\ s\in[s_{\min,c},s_{\max,c}]}} \frac{1}{n_c}\sum_{j=1}^{n_c}\ell_{n_c,z_{D,c}}(D_{c,j}-m,s). \tag{4}$$

The convex perspective loss jointly estimates location and influence scale, within the robust M-estimation framework (Huber, 1964). Scale-boundary hits are logged as diagnostics. On the disjoint calibration split, define $S_j^D = \frac{D_j-\widehat{m}_{D,H_j}}{\widehat{s}_{D,H_j}+\varepsilon_D}$. For $n_{\text{cal}} = |\mathcal{I}_{\text{cal}}^D|$ and $k_\alpha = \lceil (n_{\text{cal}}+1)(1-\alpha)\rceil$, let $\widehat{q}_{1-\alpha}$ be the k_αth calibration order statistic, or $+\infty$ if $k_\alpha > n_{\text{cal}}$. A rollout receives

$$S_{b,i}^D = \frac{D_{b,i}-\widehat{m}_{D,H_{b,i}}}{\widehat{s}_{D,H_{b,i}}+\varepsilon_D}, \quad w_{b,i} = \exp\{-\lambda[S_{b,i}^D-\widehat{q}_{1-\alpha}]_+\}. \tag{5}$$

We take $w = 1$ when $\widehat{q}_{1-\alpha} = +\infty$. The threshold controls when downweighting begins; $\lambda > 0$ controls its rate and is selected without reusing the conformal split. The coefficient w controls optimization influence; it is not a posterior probability that the response is clean. With a fixed reference, as in RQ2, the same weight map is defined without the coverage claim of split calibration.

4.2 Reliability-First Robust Advantages

For group b, compute total reliability W_b. A group with $W_b = 0$ is skipped before evaluating the relative weights or effective size. For $W_b > 0$, define

$$W_b = \sum_{i=1}^{G} w_{b,i}, \quad \bar{w}_b = \frac{W_b}{G}, \quad p_{b,i} = \frac{w_{b,i}}{W_b}, \quad G_{\text{eff},b} = \frac{W_b^2}{\sum_i w_{b,i}^2}. \tag{6}$$

The primary algorithm declares a group admissible when $W_b \geq W_{\min} > 0$ and $G_{\text{eff},b} \geq G_{\min} > 1$. Non-admissible groups have zero reward advantage and do not enter the shared-scale fit. The recorded STAR runs use this group-admission rule. In the primary objective, keeping the policy-loss denominator equal to the original number of groups makes abstention reduce, rather than renormalize, the reward update. If no group is admissible, the optimizer step is skipped. The recorded STAR runs use token-mean loss reduction, specified in Appendix F.

Let $\mathcal{B}_{\text{adm}}$ be the admissible groups and $B_{\text{adm}} = |\mathcal{B}_{\text{adm}}|$. For $b \in \mathcal{B}_{\text{adm}}$, one prompt-specific location μ_b and a shared scale $v \in [v_{\min}, v_{\max}]$ minimize

$$\ell_b^A(u,v) = \frac{G_{\text{eff},b}v}{z_A^2}\left(\sqrt{1+\frac{z_A^2u^2}{G_{\text{eff},b}v^2}}-1\right)+\frac{v}{2},$$

$$\mathcal{L}_A(\mu_{1:B},v) = \frac{1}{B_{\text{adm}}}\sum_{b\in\mathcal{B}_{\text{adm}}}\sum_{i=1}^{G} p_{b,i}\ell_b^A(r_{b,i}-\mu_b,v). \tag{7}$$

Algorithm 1 STAR-GRPO with the reference numerical protocol

Require: Frozen fit/calibration objects; old and reference policies; $G, \kappa, z_A, \lambda, W_{\min}, G_{\min}$.

1: Sample G rollouts per prompt from the frozen old policy.
2: Evaluate deployed and canonical views; compute D, $r = h(R_\kappa)$, S^D, and w.
3: Compute $W_b, \bar{w}_b, G_{\mathrm{eff},b}$; set reward advantages to zero for non-admissible groups.
4: If any group is admissible, fit (7) on those groups to the recorded KKT tolerance and compute (9).
5: **if** no group is admissible or the solver/KKT check fails **then**
6: Skip the complete optimizer step and record the failure mode.
7: **else**
8: Stop gradients through all reward-side quantities and optimize the standard clipped GRPO objective with the original BG reduction; do not recenter, whiten, or rescale A^{STAR}.
9: **end if**
10: Log contexts, both rewards, $D, w, \bar{w}, G_{\mathrm{eff}}$, abstention, solver residuals, boundary hits, and reward-gradient norms.

Sharing v pools normalization information across the minibatch while retaining prompt-specific centers, which stabilizes scale estimation for small prompt groups. Absolute group reliability is then reintroduced explicitly in Equation (9), so the strength of each group update remains coupled to the amount of supported reward evidence. At a fitted solution $(\widehat{\mu}_{1:B}, \widehat{v})$, put

$$x_{b,i} = \frac{z_A(r_{b,i} - \widehat{\mu}_b)}{\sqrt{G_{\mathrm{eff},b}\widehat{v}}}, \quad \varphi_{b,i} = \frac{x_{b,i}}{\sqrt{1 + x_{b,i}^2}}, \quad \nu_b = \left(\frac{1}{G}\sum_i w_{b,i}^2\right)^{1/2}, \tag{8}$$

and define the STAR advantage

$$A_{b,i}^{\mathrm{STAR}} = \bar{w}_b \, \frac{w_{b,i}}{\nu_b + \varepsilon_w} \, \varphi_{b,i}. \tag{9}$$

Placing reliability in the fit reduces a suspicious rollout's weight when estimating the baseline; a multiplier applied only after ordinary normalization would leave that baseline unchanged. The factor $p_{b,i}$ provides this relative weighting. The factor $\bar{w}_b$ preserves absolute reliability: for a fixed admitted set and $\varepsilon_w = 0$, multiplying all weights in a group by $c \in (0, 1]$ multiplies its advantages by c. Thus relative weights determine which rollouts inform the baseline, while absolute reliability determines update strength. STAR requires two scoring calls per rollout: two interfaces to one model in RQ1, or two evaluators in RQ2. Each first-order or Newton pass costs $O(BG)$; prompt locations parallelize, and the shared scale is one dimensional.

5 Theoretical Guarantees

The construction admits three levels of analysis. Magnitude bounds follow directly from the advantage formula; centering additionally requires a location equation; statistical reliability requires assumptions about the paired assessments and reference data. All update bounds concern the initial reward-side direction at the old policy, with reward-side quantities held fixed. Proofs appear in Appendices B–D.

Theorem 5.1 (Magnitude control and centering)**.** *Fix weights $w_{b,i} \in [0, 1]$ and an admissible set. For any finite locations and positive scale, Equation* (9) *satisfies*

$$\frac{1}{G}\sum_i (A_{b,i}^{\mathrm{STAR}})^2 \le \bar{w}_b^2, \qquad |A_{b,i}^{\mathrm{STAR}}| \le w_{b,i}.$$

If the weighted location equation $\sum_i p_{b,i}\varphi_{b,i} = 0$ holds, then $\sum_i A_{b,i}^{\mathrm{STAR}} = 0$. Separately, objective (7) *is jointly convex and has a unique minimizer if every admitted prompt has two distinct positive-weight rewards. Let $g_{b,i}^{\mathrm{avg}} = |O_{b,i}|^{-1}\sum_t \nabla_\theta \log \pi_{\theta_{\mathrm{old}}}(o_{b,i,t} \mid X_b, o_{b,i,<t})$ satisfy $\|g_{b,i}^{\mathrm{avg}}\|_2 \le L_\pi$. Then*

$$\left\|U_b^{\mathrm{STAR}}\right\|_2 = \left\|\frac{1}{G}\sum_i A_{b,i}^{\mathrm{STAR}} g_{b,i}^{\mathrm{avg}}\right\|_2 \le \bar{w}_b L_\pi, \qquad \left\|\frac{A_{b,i}^{\mathrm{STAR}} g_{b,i}^{\mathrm{avg}}}{G}\right\|_2 \le \frac{L_\pi w_{b,i}}{G}.$$

These direction bounds do not require an exact fit.

The proof uses $|\varphi| \le 1$ and $\bar{w}_b \le \nu_b$. For an approximate fit, the centering error is exactly $|\sum_i A^{\mathrm{STAR}}_{b,i}| = \bar{w}_b W_b e^{\varphi}_b/(\nu_b + \varepsilon_w)$, where $e^{\varphi}_b = |\sum_i p_{b,i}\varphi_{b,i}|$. This identity makes numerical centering directly auditable, while the coordinate and second-moment bounds hold for any finite fitted location and positive scale.

Proposition 5.2 (One outlier: ordering and attenuation)**.** *Let $G \ge 3$, $n = G - 1$, $M \in (0, R_{\max}]$, $r = (0, \dots, 0, M)$, and $w = (1, \dots, 1, \epsilon)$ for $\epsilon \in (0, 1]$. Use population-form standard deviations without a numerical stabilizer for the two comparators. Ordinary GRPO followed by multiplication by w gives each clean rollout advantage $-1/\sqrt{n}$. Using the weighted mean $\mu_w = \sum_i w_i r_i / \sum_i w_i$ and weighted standard deviation σ_w instead gives*

$$\widetilde{A}_i = w_i(r_i - \mu_w)/\sigma_w, \qquad \widetilde{A}_{\mathrm{clean}} = -\sqrt{\epsilon/n}, \qquad \widetilde{A}_{\mathrm{outlier}} = \sqrt{n\epsilon}.$$

For an admitted STAR group satisfying the exact location equation,

$$|A^{\mathrm{STAR}}_{\mathrm{outlier}}| \le \epsilon, \qquad |A^{\mathrm{STAR}}_{\mathrm{clean}}| \le \epsilon/n.$$

If $W_{\min} < n$, $G_{\min} < n$, and the scale lies in fixed $[v_{\min}, v_{\max}] \subset (0, \infty)$, exact STAR minimizers also satisfy $\widehat{\mu}_\epsilon \to 0$ as $\epsilon \downarrow 0$.

STAR therefore converts a rollout reliability of ϵ into an $O(\epsilon)$ coordinate bound for both the suspicious rollout and, through exact centering, the induced clean-rollout perturbation. This is precisely the reliability-first behavior sought by the method: unsupported mass vanishes before it can dominate group normalization.

Theorem 5.3 (Conditional reliability control)**.** *Fix $\alpha \in (0, 1)$ and condition on $\mathcal{I}^D_{\mathrm{fit}}$. If clean calibration pairs and one new clean pair are exchangeable, then*

$$\mathbb{P}(w < 1 \mid Z = 0, \mathcal{I}^D_{\mathrm{fit}}) \le \alpha, \qquad \mathbb{E}[1 - w \mid Z = 0, \mathcal{I}^D_{\mathrm{fit}}] \le \alpha.$$

Here $Z = 0$ denotes a clean sample and $Z = 1$ an attacked sample. If $\Delta \ge 0$, $\beta \in [0, 1]$, and $\mathbb{P}\{S^D \ge \widehat{q}_{1-\alpha} + \Delta \mid Z = 1\} \ge 1 - \beta$, then $\mathbb{E}[w \mid Z = 1] \le \eta_A := \beta + (1 - \beta)e^{-\lambda\Delta}$.

The first claim is the standard finite-sample marginal split-calibration guarantee (Vovk et al., 2005; Angelopoulos and Bates, 2023). Corollary C.7 extends this control to adaptive round t under the stated conditional total-variation premise, giving a clean downweighting probability of at most $\alpha + \mathbb{E}\rho_t$. For separated attacked scores, the same weight map yields the exponential attenuation in Theorem 5.3.

Corollary 5.4 (From reliability to initial influence)**.** *Under the attacked-score condition of Theorem 5.3 and the policy-score bound of Theorem 5.1, with skipped contributions set to zero,*

$$\mathbb{E}\Big[\Big\| A^{\mathrm{STAR}}_{b,i} g^{\mathrm{avg}}_{b,i}/G \Big\|_2 \mid Z_{b,i} = 1\Big] \le L_\pi \eta_A / G.$$

For the token-mean reduction used in the experiments, Corollary D.6 gives the corresponding length-weighted control. For response lengths L_i and $T = \sum_i L_i > 0$, $\|U_{\mathrm{token}}\|_2 \le L_\pi \sum_i L_i w_i / T$, directly linking the realized optimization reduction to the rollout reliabilities.

6 Experiments

We evaluate STAR-GRPO in two qualitatively different reward-hacking regimes that share the same structural failure: the optimized score becomes larger than the support provided by an alternative view of the same rollout. *RQ1* studies representation-dependent reward inflation caused by a token-interface mismatch. *RQ2* studies semantic proxy overoptimization, where a rubric-conditioned training judge increasingly disagrees with an independent evaluator. The two settings test the same reliability-first principle with different score-pair semantics. Detailed model, calibration, optimizer, and evaluation configurations are given in Appendix F.

6.1 RQ1: Mitigating Token-Interface Reward Hacking

Setup. We use Llama-3.2-1B-Instruct as the policy and Skywork-Reward-V2-Qwen3-8B (Liu et al., 2026) as the reward model, with 10,000 WildChat training prompts (Zhao et al., 2024) and 100 curated NoveltyBench validation prompts (Zhang et al., 2025). Each validation prompt is sampled

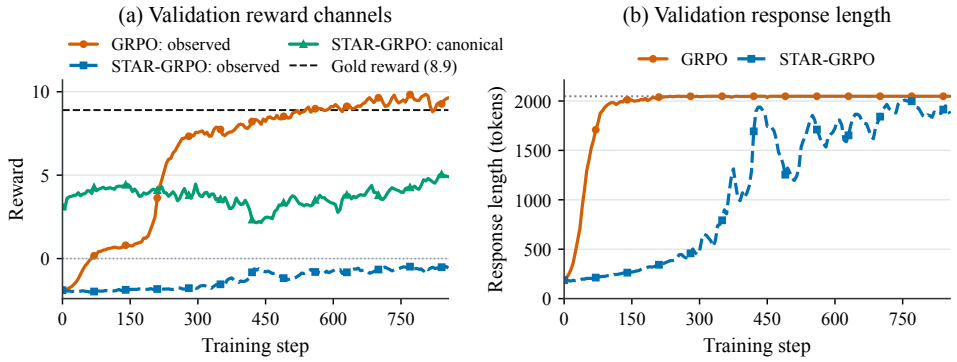


Figure 2: RQ1 validation trajectories through step 855. (a) TOMPA-GRPO rapidly increases the deployed-interface reward, whereas STAR tracks both deployed and canonical views and optimizes the canonical quality path. The dashed 8.9 line is a reference-answer score on the observed scale. (b) TOMPA-GRPO reaches the 2,048-token response limit early in training, while STAR maintains a paired-score training signal throughout optimization.

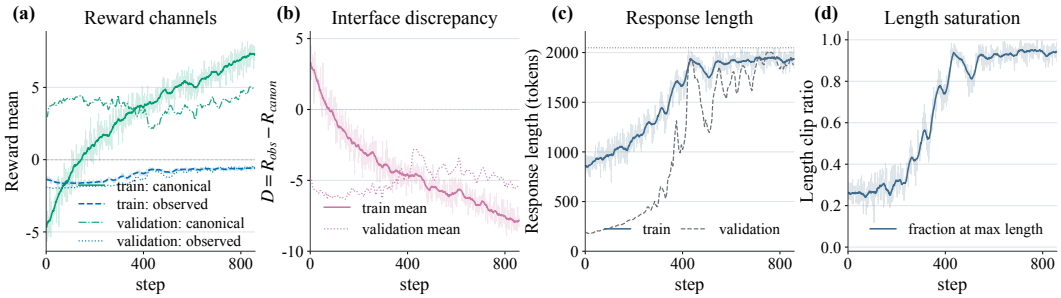


Figure 3: Canonical-primary STAR in RQ1: (a) observed and canonical rewards, (b) interface discrepancy, (c) response length, and (d) training length-clip ratio. Training and validation traces follow the legend styles.

eight times. Training uses learning rate 10^{-6}, batch size 64, eight rollouts per prompt, one PPO epoch, bfloat16, and a reference-policy KL coefficient of 0.001. Prompt and response limits are 512 and 2,048 policy tokens, and all reward-model evaluations use a 4,096-token input limit.

TOMPA-GRPO optimizes the deployed-interface score R^{obs} through the identity token map $\Phi(j) = j$, which transfers response token IDs directly to the reward-model vocabulary. STAR evaluates the same rollout through both the deployed interface and a canonical decode–normalize–retokenize path, producing

$$D = R^{\text{obs}} - R^{\text{canon}}. \tag{10}$$

The STAR quality path is canonical-primary, $r = 10\tanh(R^{\text{canon}}/10)$ ($\kappa = 0$), while D determines how strongly each rollout is allowed to influence the robust group statistics and the final advantage. This directly targets representation-dependent reward inflation: optimization follows canonical quality, while the deployed/canonical mismatch controls reliability.

Mitigating interface exploitation. TOMPA-GRPO exhibits the characteristic token-interface failure mode: its observed reward rises from -1.894 to 9.641, and the mean response length reaches the 2,048-token cap by step 235. In contrast, STAR improves the canonical validation reward from 3.121 to 4.902, an absolute gain of 1.781 (57.1% relative), while the observed-interface reward ends at -0.570. The optimized signal therefore moves in the canonical direction rather than following the exploitable deployed-interface score. Figure 2 makes this separation visible over the full trajectory: the TOMPA reward channel accelerates toward the hacked high-score regime, whereas STAR preserves a canonical quality signal throughout training.

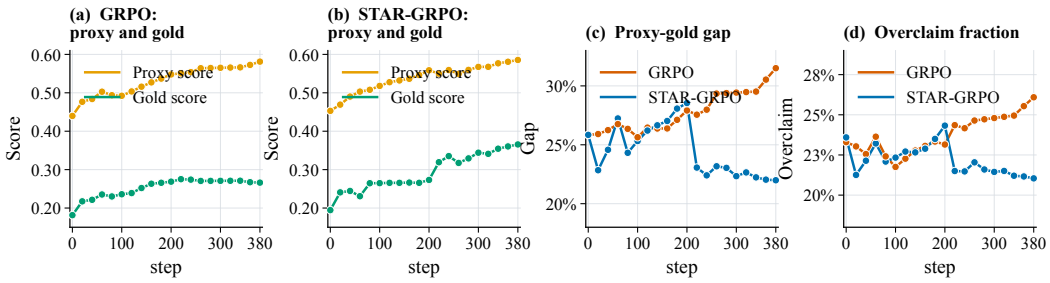


Figure 4: RQ2 validation trajectories. (a,b) Training-proxy and independent-judge scores for GRPO and STAR-GRPO; (c) proxy–judge gap; (d) overclaim fraction. The independent judge is used only for evaluation and is distinct from STAR's rubric-free training anchor.

Selective reliability. STAR's average rollout reliability remains high, between 0.9847 and 1.0000, and the mean effective group size stays between 7.955 and 8.000 out of eight. At the same time, the minimum individual reliability reaches 0.105 over the recorded training trajectory (Figure 3). This is the intended operating regime: STAR leaves the bulk of supported rollouts nearly unchanged while strongly attenuating the small subset whose deployed score is poorly supported by the canonical view. The canonical-only diagnostic control and additional endpoint statistics are reported in Appendix F.1.1; they confirm that the canonical channel provides a meaningful anchor, while STAR turns that anchor disagreement into a general reliability mechanism that also applies when the proxy itself must remain the optimization target, as in RQ2.

6.2 RQ2: Mitigating Rubric-Proxy Overoptimization

Setup. We train Qwen3-4B on RubricHub-Medical (Li et al., 2026) and evaluate on HealthBench-Hard (Arora et al., 2025). Both GRPO and STAR optimize the same GPT-4o-mini rubric-conditioned proxy score. STAR additionally obtains a rubric-free semantic assessment from Gemini-2.5-Flash-Lite and forms $D^{\text{rubric}} = R^{\text{proxy}} - R^{\text{anchor}}$ to estimate reliability. The scalar quality path remains $r = R^{\text{proxy}}$ ($\kappa = 1$): the anchor does not replace the task reward; it determines how strongly the proxy score is trusted for learning. Claude-Sonnet-4-6 is used only as an independent validation judge and never enters the policy update.

Both methods use the same policy model, training data, 16 rollouts per prompt, PPO schedule, token-mean loss aggregation, and validation protocol. STAR uses fixed discrepancy center 0, scale 0.1, reliability threshold 1.645, and decay 2. Invalid score pairs are mapped to zero reliability and therefore excluded from the reward-side fit, consistent with the same reliability-first rule used for valid but weakly supported proxy scores.

Evaluation. The primary external metric is the independent-judge score. We additionally report the proxy–judge gap, criterion pass rate, and overclaim fraction. The proxy–judge gap measures the extent to which optimization progress on the training proxy is not reproduced by the independent evaluator; overclaim is the fraction of positive-weight rubric criteria credited by the proxy but not by the independent judge. Both validation judges score the same generated response against the same evaluation rubric, and Table 7 uses the latest common checkpoint satisfying the fixed validation-availability criterion.

Independent quality improves as proxy overclaim decreases. STAR reaches an independent-judge score of 0.3174 compared with 0.2706 for GRPO, corresponding to a 17.3% relative increase. At the same checkpoint, the proxy–judge gap decreases from 0.2935 to 0.2318 (21.0% relative reduction), while the overclaim fraction decreases from 0.2464 to 0.2204 (10.6% relative reduction). Criterion pass rate simultaneously increases from 0.4752 to 0.5049. Importantly, these external gains occur while STAR's proxy score decreases from 0.5640 to 0.5492. This is the desired signature of reduced reward hacking: STAR is not simply driving the proxy harder; it converts more of the optimized proxy signal into quality that is reproduced by an independent evaluator.

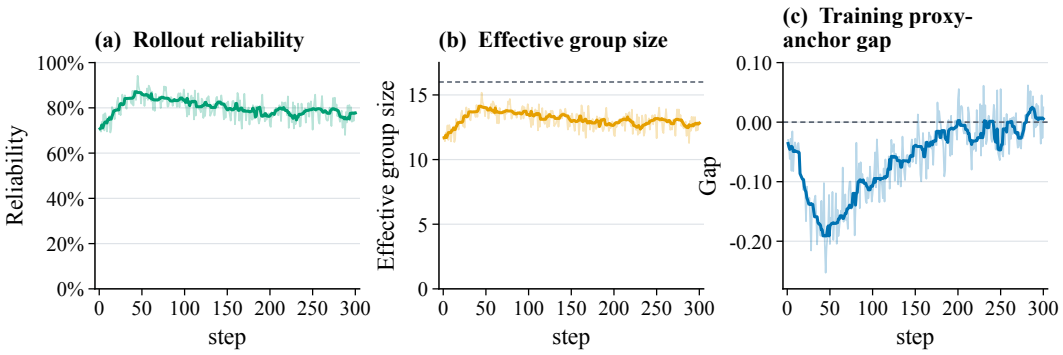


Figure 5: STAR training diagnostics in RQ2: rollout reliability, effective group size, and proxy–anchor gap. Bold curves are 11-step rolling medians of per-update metadata. The training anchor differs from the independent evaluation judge.

Reliability is active throughout training. The 11-step rolling median rollout reliability remains approximately 0.70–0.90 and ends near 0.78, while mean effective group size ranges from 11.3 to 15.2 out of 16 (Figure 5). The training proxy–anchor gap ranges from approximately -0.253 to 0.062 and moves toward zero, and non-admissible groups reach approximately 10.9%. These diagnostics show that the mechanism is active rather than degenerate: reliability continuously reshapes the contribution of individual rollouts and, when support becomes too weak, converts uncertainty into abstention instead of a potentially misleading reward update. The extra reward-side computation is lightweight relative to policy rollout: one anchor assessment per rollout plus an $O(BG)$ robust fit with a one-dimensional shared scale.

Two reward-hacking regimes, one optimization principle. RQ1 and RQ2 stress STAR in complementary conditions. RQ1 targets *representation-dependent* reward inflation, where the deployed token interface itself creates an exploitable score channel; STAR shifts learning toward canonical quality and prevents the hacked observed reward from becoming the optimization target. RQ2 targets *semantic proxy* overoptimization, where the proxy must remain the task reward; STAR therefore uses an independent semantic view only to modulate influence. Across both regimes, paired reward assessments provide the missing signal, and reliability-first normalization turns that signal into controlled group statistics and bounded policy advantages.

7 Conclusion

Reward hacking exposes a structural weakness of group-relative policy optimization: an unsupported reward can distort the group baseline and thereby change the advantages assigned to other rollouts. STAR-GRPO addresses this problem using paired assessments of each rollout to estimate reliability before normalization. Reliability shapes the robust group fit, while an absolute group-reliability factor attenuates the bounded advantages used for policy updates. The resulting construction provides coordinate and second-moment control, an auditable condition for exact centering, and reliability-dependent attenuation of reward-side influence.

Across two distinct reward-hacking settings, the experiments show that STAR limits the influence of poorly supported scores. In token-interface exploitation, it suppresses runaway deployed-interface reward while improving canonical reward. In rubric-proxy medical reasoning, it improves independent-judge quality and reduces both proxy–judge disagreement and overclaim while retaining the proxy as the task reward. These results suggest that an auxiliary assessment can improve the robustness of group-relative learning by regulating how strongly a reward shapes the baseline and policy update, without requiring the same assessment to serve as the optimization target.

Ethics Statement. STAR supports reliable reward-based training by monitoring paired scoring views. Deployment also calls for independent semantic evaluation, appropriate human oversight, and protection of sensitive logged outputs and model interfaces. Attack descriptions and evaluation artifacts should be released with safeguards proportionate to the systems they may affect.

Reproducibility Statement. Equations (2)–(9) fully specify the STAR advantage construction. The appendix provides complete proofs, the reference numerical protocol, the paired-score interfaces, calibration and reliability parameters, optimizer settings, and the realized RQ1/RQ2 configurations needed to reproduce the reported experiments.

AI Use Statement. Generative AI tools were used to assist with manuscript restructuring, literature discovery, methodological critique, mathematical formulation and checking, proof editing, and LaTeX preparation. They were not used to generate experimental observations or to independently determine empirical conclusions. The authors retain responsibility for checking mathematical claims, proofs, citations, and manuscript content against the stated assumptions, derivations, experimental artifacts, and primary sources, and for the final paper.

REFERENCES

A. N. Angelopoulos and S. Bates. Conformal prediction: A gentle introduction. *Foundations and Trends in Machine Learning*, 16(4):494–591, 2023. https://doi.org/10.1561/2200000101.

R. K. Arora, J. Wei, R. S. Hicks, P. Bowman, J. Quiñonero-Candela, F. Tsimpourlas, M. Sharman, M. Shah, A. Vallone, A. Beutel, J. Heidecke, and K. Singhal. HealthBench: Evaluating Large Language Models Towards Improved Human Health. arXiv preprint arXiv:2505.08775, 2025. https://arxiv.org/abs/2505.08775.

Y. Bai, A. Jones, K. Ndousse, A. Askell, A. Chen, N. DasSarma, D. Drain, S. Fort, D. Ganguli, T. Henighan, et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. arXiv preprint arXiv:2204.05862, 2022. https://arxiv.org/abs/2204.05862.

R. F. Barber, E. J. Candès, A. Ramdas, and R. J. Tibshirani. Conformal prediction beyond exchangeability. *The Annals of Statistics*, 51(2), 2023. https://doi.org/10.1214/23-aos2276.

S. Casper, X. Davies, C. Shi, T. K. Gilbert, J. Scheurer, J. Rando, R. Freedman, T. Korbak, D. Lindner, P. Freire, et al. Open problems and fundamental limitations of reinforcement learning from human feedback. arXiv preprint arXiv:2307.15217, 2023. https://arxiv.org/abs/2307.15217.

O. Catoni. Challenging the empirical mean and empirical variance: A deviation study. *Annales de l'Institut Henri Poincare, Probabilites et Statistiques*, 48(4):1148–1185, 2012. https://doi.org/10.1214/11-AIHP454.

T. Chen, X. Liu, V. Nandam, K.-R. Liou, and H. Wei. Conformal feedback alignment: Quantifying answer-level reliability for robust LLM alignment. arXiv preprint arXiv:2601.17329, 2026. https://arxiv.org/abs/2601.17329.

P. F. Christiano, J. Leike, T. Brown, M. Martic, S. Legg, and D. Amodei. Deep reinforcement learning from human preferences. In *Advances in Neural Information Processing Systems*, 30, 2017. https://papers.nips.cc/paper_files/paper/2017/hash/d5e2c0adad503c91f91df240d0cd4e49-Abstract.html.

T. Coste, U. Anwar, R. Kirk, and D. Krueger. Reward Model Ensembles Help Mitigate Overoptimization. In *International Conference on Learning Representations*, 2024. https://arxiv.org/abs/2310.02743.

A. D'Amour, K. Heller, D. Moldovan, B. Adlam, B. Alipanahi, A. Beutel, C. Chen, J. Deaton, J. Eisenstein, M. D. Hoffman, et al. Underspecification presents challenges for credibility in modern machine learning. *Journal of Machine Learning Research*, 23(226):1–61, 2022. https://jmlr.org/papers/v23/20-1335.html.

Y. Dubois, B. Galambosi, P. Liang, and T. B. Hashimoto. Length-Controlled AlpacaEval: A Simple Way to Debias Automatic Evaluators. arXiv preprint arXiv:2404.04475, 2024. https://arxiv.org/abs/2404.04475.

L. Gao, J. Schulman, and J. Hilton. Scaling laws for reward model overoptimization. In *Proceedings of the 40th International Conference on Machine Learning*, pages 10835–10866, 2023. https://arxiv.org/abs/2210.10760.

P. J. Huber. Robust Estimation of a Location Parameter. *The Annals of Mathematical Statistics*, 35(1), pages 73–101, 1964. https://doi.org/10.1214/aoms/1177703732.

N. Lambert, V. Pyatkin, J. Morrison, LJ Miranda, B. Y. Lin, K. Chandu, N. Dziri, S. Kumar, T. Zick, Y. Choi, N. A. Smith, and H. Hajishirzi. RewardBench: Evaluating Reward Models for Language Modeling. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 1755–1797, 2025. https://aclanthology.org/2025.findings-naacl.96/.

G. Lecue and M. Lerasle. Robust machine learning by median-of-means: Theory and practice. *The Annals of Statistics*, 48(2):906–931, 2020. https://doi.org/10.1214/19-AOS1828.

S. Li, J. Zhao, H. Ren, Z. Wei, Y. Zhou, J. Yang, S. Liu, K. Zhang, and W. Chen. RubricHub: A Comprehensive and Highly Discriminative Rubric Dataset via Automated Coarse-to-Fine Generation. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 31320–31344, 2026. https://aclanthology.org/2026.acl-long.1445/.

Z. Liu, C. Chen, W. Li, P. Qi, T. Pang, C. Du, W. S. Lee, and M. Lin. Understanding R1-Zero-Like Training: A Critical Perspective. arXiv preprint arXiv:2503.20783, 2025. https://arxiv.org/abs/2503.20783.

C. Y. Liu, L. Zeng, Y. Xiao, J. He, J. Liu, C. Wang, R. Yan, W. Shen, F. Zhang, J. Xu, Y. Liu, and Y. Zhou. Skywork-Reward-V2: Scaling Preference Data Curation via Human-AI Synergy. In *International Conference on Learning Representations*, 2026. https://arxiv.org/abs/2507.01352.

T. Liu, W. Xiong, J. Ren, L. Chen, J. Wu, R. Joshi, Y. Gao, J. Shen, Z. Qin, T. Yu, et al. RRM: Robust reward model training mitigates reward hacking. arXiv preprint arXiv:2409.13156, 2024. https://arxiv.org/abs/2409.13156.

G. Lugosi and S. Mendelson. Mean estimation and regression under heavy-tailed distributions: A survey. *Foundations of Computational Mathematics*, 19(5):1145–1190, 2019. https://doi.org/10.1007/s10208-019-09427-x.

L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, et al. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, 2022. https://arxiv.org/abs/2203.02155.

J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov. Proximal policy optimization algorithms. arXiv preprint arXiv:1707.06347, 2017. https://arxiv.org/abs/1707.06347.

Z. Shao, P. Wang, Q. Zhu, R. Xu, J. Song, X. Bi, H. Zhang, M. Zhang, Y. K. Li, Y. Wu, et al. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. arXiv preprint arXiv:2402.03300, 2024. https://arxiv.org/abs/2402.03300.

J. Skalse, N. Howe, D. Krasheninnikov, and D. Krueger. Defining and Characterizing Reward Gaming. In *Advances in Neural Information Processing Systems*, 35, 2022. https://proceedings.neurips.cc/paper_files/paper/2022/hash/3d719fee332caa23d5038b8a90e81796-Abstract-Conference.html.

Q. Sun. Do we need to estimate the variance in robust mean estimation? arXiv preprint arXiv:2107.00118, version 5, 2024. https://arxiv.org/abs/2107.00118v5.

V. Vovk, A. Gammerman, and G. Shafer. *Algorithmic Learning in a Random World*. Springer, 2005. https://doi.org/10.1007/b106715.

R. Wu and R. Tang. From rebound to remedy: Understanding and mitigating reward hacking via representation engineering. arXiv preprint arXiv:2604.01476, 2026. https://arxiv.org/abs/2604.01476.

Q. Yu, Z. Zhang, R. Zhu, Y. Yuan, X. Zuo, Y. Yue, W. Dai, T. Fan, G. Liu, L. Liu, X. Liu, H. Lin, Z. Lin, B. Ma, G. Sheng, Y. Tong, C. Zhang, M. Zhang, W. Zhang, H. Zhu, J. Zhu, J. Chen, J. Chen, C. Wang, H. Yu, Y. Song, X. Wei, H. Zhou, J. Liu, W.-Y. Ma, Y.-Q. Zhang, L. Yan, M. Qiao, Y. Wu, and M. Wang. DAPO: An Open-Source LLM Reinforcement Learning System at Scale. arXiv preprint arXiv:2503.14476, 2025. https://arxiv.org/abs/2503.14476.

Y. Zhai, H. Zhang, Y. Lei, Y. Yu, K. Xu, D. Feng, B. Ding, and H. Wang. Uncertainty-penalized reinforcement learning from human feedback with diverse reward LoRA ensembles. arXiv preprint arXiv:2401.00243, 2024. https://arxiv.org/abs/2401.00243.

Y. Zhang, H. Diddee, S. Holm, H. Liu, X. Liu, V. Samuel, B. Wang, and D. Ippolito. NoveltyBench: Evaluating Language Models for Humanlike Diversity. arXiv preprint arXiv:2504.05228, 2025. https://arxiv.org/abs/2504.05228.

Y. Zhang, M. Huo, M. Zhu, M. Zhang, and N. Jiang. Beyond semantic manipulation: Token-space attacks on reward models. arXiv preprint arXiv:2604.02686, 2026. https://arxiv.org/abs/2604.02686.

W. Zhao, X. Ren, J. Hessel, C. Cardie, Y. Choi, and Y. Deng. WildChat: 1M ChatGPT Interaction Logs in the Wild. In *International Conference on Learning Representations*, 2024. https://arxiv.org/abs/2405.01470.

L. Zheng, W.-L. Chiang, Y. Sheng, S. Zhuang, Z. Wu, Y. Zhuang, Z. Lin, Z. Li, D. Li, E. P. Xing, H. Zhang, J. E. Gonzalez, and I. Stoica. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. arXiv preprint arXiv:2306.05685, 2023. https://arxiv.org/abs/2306.05685.

A Extended Related Work

Learning a reward from human preferences separates feedback collection from policy optimization (Christiano et al., 2017; Ouyang et al., 2022; Bai et al., 2022). That separation creates a proxy objective whose maximization need not preserve the intended reward ordering. Formal work characterizes when proxy and true rewards can disagree under optimization (Skalse et al., 2022); scaling experiments quantify overoptimization under PPO and best-of-n sampling (Gao et al., 2023). Underspecification explains why equivalent training-domain performance need not imply equivalent deployment behavior (D'Amour et al., 2022), and surveys identify broader limitations of RLHF (Casper et al., 2023). Our identification results address the narrower information boundary of detecting shifts from paired scores.

Reward-model ensembles mitigate overoptimization through conservative objectives (Coste et al., 2024), while diverse LoRA ensembles support uncertainty-penalized rewards (Zhai et al., 2024). Robust reward-model training uses a causal formulation and data augmentation to reduce prompt-independent preference artifacts (Liu et al., 2024). Conformal Feedback Alignment constructs answer-level reliability for DPO- and PPO-style alignment (Chen et al., 2026). These approaches motivate reliability-aware learning but intervene at different points. STAR's contribution is orthogonal: it uses a setting-specific score discrepancy to weight the group baseline *before* normalization and then preserves the absolute reliability level in the final advantages, directly targeting contamination of group-relative statistics.

Group-relative optimization removes the learned critic by forming advantages within each prompt group (Shao et al., 2024). Dr. GRPO identifies how per-response length reduction and reward-standard-deviation normalization change optimization weights (Liu et al., 2025). DAPO's token-level objective assigns equal weight to valid tokens, so longer responses contribute more to the aggregate direction (Yu et al., 2025). These observations motivate our explicit distinction between equal-response and token-mean reductions. STAR changes the reward-side advantage construction while retaining the specified PPO likelihood-ratio and clipping mechanics (Schulman et al., 2017); the initial-direction bounds do not imply guarantees for the full optimization trajectory.

Evaluation introduces a separate reliability problem. RewardBench tests preference discrimination across chat, reasoning, and safety cases (Lambert et al., 2025). LLM-as-a-judge studies document position, verbosity, and self-enhancement biases (Zheng et al., 2023); length-controlled AlpacaEval explicitly adjusts for verbosity as a confounder (Dubois et al., 2024). These findings motivate interpreting response length as a diagnostic and paired disagreement as an operational signal. RQ2 uses RubricHub's structured training criteria (Li et al., 2026) and the physician-designed HealthBench evaluation rubrics (Arora et al., 2025), with our specified proxy and independent judges. Using these prompts and rubrics does not make our reported scores an evaluation under the original benchmarks' complete protocols. Similarly, RQ1 uses prompts from WildChat and NoveltyBench (Zhao et al., 2024; Zhang et al., 2025), but reports reward and length dynamics rather than NoveltyBench's diversity metric.

The statistical tools serve distinct roles. Robust M-estimation limits the effect of extreme residuals (Huber, 1964); bounded-influence and median-based estimators address heavy tails (Catoni, 2012; Lugosi and Mendelson, 2019; Lecue and Lerasle, 2020). Self-tuned pseudo-Huber estimation jointly adapts location and robustification scale (Sun, 2024). Split conformal methods provide finite-sample marginal guarantees under exchangeability (Vovk et al., 2005; Angelopoulos and Bates, 2023). Conformal prediction beyond exchangeability studies coverage loss under distribution drift (Barber et al., 2023); our adaptive-round statement instead uses its explicitly stated conditional total-variation premise. Neither that premise nor clean exchangeability is supplied automatically by on-policy training.

B Core Proofs and Formal Foundations

The following proof map links the main-text statements to their detailed derivations. We then fix notation and establish the identification results.

B.1 Proof Map and Formal Setup

The following proof map separates the main statements by the assumptions they use. The identification and calibration results concern observable score information; the magnitude and initial-direction bounds are deterministic conditional on the realized reward and weight arrays.

Proof of Theorem 3.1. Part (i) is Theorem B.2. Part (ii) is Theorem B.4, whose total-variation argument also covers randomized discrepancy-only tests after conditioning on their auxiliary randomness. □

Proof of Theorem 5.3. The clean and separated-attack claims are Theorem C.9. The adaptive extension discussed after the theorem is Corollary C.7: its conditional total-variation premise transfers the static rank bound and averaging gives the outer marginal guarantee. □

Proof of Theorem 5.1. The magnitude bounds and the conditional zero-sum statement follow from Theorem D.3. Theorem D.1 establishes convexity and uniqueness separately. Applying the magnitude bounds to the old-policy score vectors gives Theorem D.5; this step uses neither optimality of the fit nor exact centering. □

Proof of Corollary 5.4. A skipped rollout contributes zero. For an admissible rollout, Theorem D.5 bounds its contribution by $L_\pi w_{b,i}/G$. Taking the attacked conditional expectation and invoking (36) gives the displayed result. □

The proof of Proposition 5.2 is given with the STAR advantage calculation in Section D.1.3, where the limiting fit and the vanishing outlier contribution can be checked together. With these dependencies fixed, we now specify the rollout model, the trusted-data split, and the probability conventions used by every later derivation.

A minibatch contains prompts $X_1, \dots, X_B$. For each prompt, the old policy samples G rollouts

$$O_{b,i} \sim \pi_{\theta_{\text{old}}}(\cdot \mid X_b), \quad b \in [B],\ i \in [G].$$

The observed interface M_{obs} is the representation actually scored by the deployed pipeline. The canonical channel C decodes the policy sequence into user-visible text, applies a prespecified normalization, and retokenizes with the reward tokenizer. Equation (1) defines the two scores. Their difference is the minimal observable needed to separate quality from interface support, so we define it before specifying the trusted calibration model.

Definition B.1 (Representation discrepancy)**.** The representation discrepancy of rollout (b, i) is

$$D_{b,i} = R^{\text{obs}}_{b,i} - R^{\text{canon}}_{b,i}.$$

Positive discrepancy measures observed-interface reward unsupported by the canonical channel.

Direct token-index mapping is not assumed to preserve semantics. It is an interface intervention. The online method uses the actual deployed interface as M_{obs}; arbitrary counterfactual mappings are used for auditing unless they are reachable in deployment.

Let $H_{b,i} \in [H]$ be a finite prespecified context. A context may encode reward-model identity, policy–reward tokenizer pair, task class, language, response-length bin, or frozen policy epoch. Contexts are fixed before the held-out conformal calibration split is examined. Rare contexts follow a deterministic fallback hierarchy.

A trusted benign dataset is split into

$$\mathcal{I}^D_{\text{fit}} \quad \text{and} \quad \mathcal{I}^D_{\text{cal}}.$$

The first fits context-specific discrepancy locations and scales; the second calibrates a single pooled standardized quantile. Policy-training rollouts and final test examples are disjoint from both splits. Fix $\varepsilon_D \geq 0$ and $\varepsilon_w \geq 0$; positive scale constraints ensure every score denominator is strictly positive.

For clean outputs in context c, write

$$D = m_{D,c} + s_{D,c}\epsilon_{D,c}, \quad \mathbb{E}\epsilon_{D,c} = 0, \quad \mathbb{E}\epsilon^2_{D,c} = 1, \tag{11}$$

with $0 < s_{D,c} < \infty$. Finite variance is used for the self-tuned scale interpretation, but conformal validity itself only requires exchangeability after the score map has been fitted.

For approximate cross-context pooling, let $F_{D,c}$ be the CDF of $\epsilon_{D,c}$ and assume

$$\sup_x |F_{D,c}(x) - F_D(x)| \le \eta_{\text{het}}. \tag{12}$$

The special case $\eta_{\text{het}} = 0$ is a common standardized residual law.

A latent label $Z \in \{0, 1\}$ is introduced only for detection-power analysis. We use three nested alternatives.

1. *Standardized shift:*
$$S^D \mid Z = 1, H = c = \epsilon_{1,c} + \Delta_c, \quad \Delta_c > 0.$$
2. *Stochastic shift:* if $F_{0,c}$ and $F_{1,c}$ are clean and attacked standardized-score CDFs,
$$F_{1,c}(t) \le F_{0,c}(t - \Delta_c).$$
3. *Quantile separation:*
$$F_{1,c}^{-1}(\beta) > F_{0,c}^{-1}(1 - \alpha).$$

Mean separation alone is not enough for power without variance or tail control.

The canonical view is an integrity anchor rather than the only permitted quality source. Fix $\kappa \in [0, 1]$ before validation and define

$$R_{\kappa,b,i} = R_{b,i}^{\text{canon}} + \kappa(R_{b,i}^{\text{obs}} - R_{b,i}^{\text{canon}}), \quad r_{b,i} = h(R_{\kappa,b,i}), \tag{13}$$

where $h : \mathbb{R} \to [-R_{\max}, R_{\max}]$ is fixed, monotone, and 1-Lipschitz. The default $\kappa = 1$ retains the deployed reward, whereas $\kappa = 0$ selects the canonical reward path. The value is fixed before calibration to specify the robustness–fidelity trade-off.

Let $w_{b,i} \in (0, 1]$ be the reliability weights defined below. Put

$$W_b = \sum_{i=1}^{G} w_{b,i}, \quad \bar{w}_b = \frac{W_b}{G}, \quad p_{b,i} = \frac{w_{b,i}}{W_b}, \quad G_{\text{eff},b} = \frac{W_b^2}{\sum_i w_{b,i}^2}. \tag{14}$$

A group is admissible when $W_b \ge W_{\min} > 0$ and $G_{\text{eff},b} \ge G_{\min} > 1$. The primary fallback is skip: the group's reward advantage is zero, it is excluded from the shared-scale fit, and the policy loss retains the original group denominator. This is the fallback used in the realized experiments. If every group is non-admissible, the complete optimizer step is skipped. The analysis below uses one final set of conventions to keep three sources of error separate.

All conformal guarantees are conditional on the fitting split $\mathcal{I}_{\text{fit}}^D$ because the standardized score map is learned there. Unless explicitly stated otherwise, probabilities in Sections C.1.1–C.2 integrate over the held-out calibration split and a new example. Conditional statements about STAR advantages treat rewards and reliability weights as fixed arrays.

For a generalized inverse, we use

$$F^{-1}(u) = \inf\{x : F(x) \ge u\}, \quad u \in [0, 1].$$

The convention $F^{-1}(0) = -\infty$ and $F^{-1}(1) = \sup\{x : F(x) < 1\}$ is used when needed. Ties in conformal scores are handled conservatively by strict rejection $S_{\text{new}} > \widehat{q}$.

There are three distinct discrepancies in the analysis:

1. *representation discrepancy* $D = R^{\text{obs}} - R^{\text{canon}}$;
2. *calibration estimation error* in $(\widehat{m}_{D,c}, \widehat{s}_{D,c}, \widehat{q})$;
3. *online weight error* $w - w^{\text{or}}$ relative to latent clean indicators.

The first is observed, the second is statistical, and the third is an oracle comparison. Human-utility error is not identified.

The theoretical group update uses one score vector $g_{b,i}$ per rollout. In token-level GRPO, this vector can be interpreted as the average of token score vectors. If every token score is bounded by L_{tok} and the rollout score is length-normalized, then $L_\pi = L_{\text{tok}}$ is valid. If score vectors are summed rather than averaged, L_π scales with response length and must be tracked explicitly.

B.2 Identifiability Boundaries

The identification results establish the information provided by an additional scoring view and characterize invariance under common shifts across views.

A scalar reward cannot reveal a counterfactual score that was never evaluated. The next result formalizes this missing-information obstruction and motivates paired scoring.

Theorem B.2. *Let P be any probability law on $\mathbb{R}$ and let $B > 0$. There exist two latent models with identical observable law $R^{\mathrm{obs}} \sim P$ such that in the first model every example is clean and in the second every example is representation-dependently attacked with canonical reward smaller by B. Consequently, no measurable detector based only on R^{obs} can uniformly distinguish the two model classes.*

Proof. In Model A draw $Y \sim P$, set $Z = 0$, $R^{\mathrm{obs}} = Y$, and $R^{\mathrm{canon}} = Y$. In Model B draw the same $Y \sim P$, set $Z = 1$, $R^{\mathrm{obs}} = Y$, and $R^{\mathrm{canon}} = Y - B$. The marginal law of the only observed variable R^{obs} is P in both models, but the attack labels and canonical rewards differ. Let $\delta : \mathbb{R} \to \{0, 1\}$ be any detector. Its distribution is identical under the two models because its argument has the same law. Hence it cannot have both zero false-positive probability in Model A and zero false-negative probability in Model B. The same argument applies to randomized detectors after conditioning on their auxiliary randomness. □

Remark B.3. Any scalar transformation of R^{obs}, including clipping, a median, a median-of-means aggregate, Huberization, or mean–standard-deviation normalization, remains a function of the same non-identifying observation.

Paired scoring resolves the single-view ambiguity only when an attack changes cross-view differences. The following result states the complementary failure boundary used throughout the later threat model.

Theorem B.4. *Let Q be any detector score measurable with respect to reward differences among a collection of views $(R_0, \ldots, R_K)$. If the clean and attacked laws of Q are identical, every test based on Q has type-I plus type-II error at least one. In particular, if an attack adds the same random shift B to all views,*

$$R_k^{\mathrm{attack}} = R_k^{\mathrm{clean}} + B, \quad k = 0, \ldots, K,$$

then all pairwise differences are unchanged and no discrepancy-only method can detect the attack.

Proof. Let P_0 and P_1 denote the clean and attacked laws of Q. For a test $\delta(Q) \in \{0, 1\}$, the sum of errors is

$$P_0\{\delta = 1\} + P_1\{\delta = 0\} = 1 - \big(P_1\{\delta = 1\} - P_0\{\delta = 1\}\big) \geq 1 - \mathrm{TV}(P_0, P_1).$$

If $P_0 = P_1$, this lower bound equals one. Under a common additive shift across all views, every difference $R_k - R_\ell$ is unchanged pointwise, so any score measurable with respect to such differences has identical clean and attacked laws. □

Together, Theorems B.2 and B.4 identify the exact scope of STAR. A canonical anchor supplies information absent from a single reward, but only attacks that create a detectable cross-interface discrepancy can be attenuated. The method is intentionally silent about shared semantic misspecification.

C Calibration and Reliability Guarantees

Trusted paired examples define the context-specific discrepancy fit and the conformal threshold. The analysis connects these quantities to clean-sample preservation, attack attenuation, and distributional drift.

C.1 Self-Tuned Context Calibration

C.1.1 Robust Context Fit

For context c, let $D_{c,1}, \ldots, D_{c,n_c}$ be trusted fitting discrepancies. Choose a confidence parameter $z_{D,c} > 0$ and scale bounds $0 < s_{\min,c} < s_{\max,c} < \infty$. Define

$$\ell_{n,z}(u, s) = \frac{ns}{z^2}\left(\sqrt{1 + \frac{z^2u^2}{ns^2}} - 1\right) + \frac{s}{2}, \tag{15}$$

and

$$(\widehat{m}_{D,c}, \widehat{s}_{D,c}) \in \underset{m,\, s_{\min,c} \le s \le s_{\max,c}}{\arg\min} \frac{1}{n_c}\sum_{j=1}^{n_c} \ell_{n_c, z_{D,c}}(D_{c,j} - m, s). \tag{16}$$

The coefficient $1/2$ makes the population scale approach the ordinary standard deviation in the large-sample regime under the self-tuned parameterization (Sun, 2024).

Put

$$a_{c,j}(m, s) = \frac{z_{D,c}^2 (D_{c,j} - m)^2}{n_c s^2}.$$

At an interior optimum, direct differentiation gives

$$0 = \sum_{j=1}^{n_c} \frac{D_{c,j} - \widehat{m}_{D,c}}{\widehat{s}_{D,c}\sqrt{1 + a_{c,j}(\widehat{m}_{D,c}, \widehat{s}_{D,c})}}, \tag{17}$$

$$0 = \frac{1}{n_c}\sum_{j=1}^{n_c}\left[\frac{n_c}{z_{D,c}^2}\left(\frac{1}{\sqrt{1 + a_{c,j}(\widehat{m}_{D,c}, \widehat{s}_{D,c})}} - 1\right) + \frac{1}{2}\right]. \tag{18}$$

The location score has bounded magnitude:

$$\left|\frac{u}{s\sqrt{1 + z^2u^2/(ns^2)}}\right| \le \frac{\sqrt{n}}{z}. \tag{19}$$

The offline fit must be globally auditable before its output can define a frozen conformal score map. The following proposition supplies that optimization property and the uniqueness condition used by the calibration protocol.

Proposition C.1 (Convexity of the offline fit)**.** *For fixed $n, z > 0$, the map $(m, s) \mapsto \ell_{n,z}(d - m, s)$ is jointly convex on $\mathbb{R} \times (0, \infty)$. Hence the empirical objective in* (16) *is jointly convex. If the positively weighted data contain at least two distinct values and the minimizer is interior, the objective is strictly convex and the minimizer is unique.*

Proof. Define the convex scalar function

$$\rho(t) = \frac{n}{z^2}\left(\sqrt{1 + \frac{z^2t^2}{n}} - 1\right).$$

Its second derivative is

$$\rho''(t) = \left(1 + \frac{z^2t^2}{n}\right)^{-3/2} > 0.$$

The first term of (15) is the perspective $s\rho(u/s)$, which is convex on $s > 0$. Composition with the affine residual $u = d - m$ preserves convexity, and adding $s/2$ preserves convexity. Strict convexity follows from strict convexity of ρ together with non-collinearity of residual directions generated by at least two distinct observations; equivalently, the sum of Hessian rank-one terms has full rank in (m, s) at an interior point. □

The next result isolates the self-tuning mechanism without overclaiming a finite-sample variance estimate.

Proposition C.2 (Population scale adaptation)**.** *Let U be a nondegenerate mean-zero random variable with variance $\sigma^2 < \infty$. Fix $n, z > 0$ with $z^2 < 2n$. Suppose the population scale equation*

$$\mathbb{E}\left[\frac{1}{\sqrt{1 + z^2U^2/(ns^2)}}\right] = 1 - \frac{z^2}{2n} \tag{20}$$

has an interior solution $s^\circ > 0$. Let $\tau^\circ = \sqrt{n}\, s^\circ / z$ and

$$\sigma^2_{\tau^\circ} = \mathbb{E}\left[U^2 \mathbf{1}\{|U| \le \tau^\circ\}\right].$$

Then

$$2\left(1 - \frac{1}{\sqrt{2}}\right)\sigma^2_{\tau^\circ} \le (s^\circ)^2 \le \sigma^2. \tag{21}$$

Consequently, if $\sigma^2_{\tau^\circ} \ge c_{\mathrm{tv}}\sigma^2$ for some $c_{\mathrm{tv}} > 0$, then

$$\sqrt{2(1 - 2^{-1/2})c_{\mathrm{tv}}}\,\sigma \le s^\circ \le \sigma.$$

Proof. Set $x = U^2/(\tau^\circ)^2$ and $g(x) = 1 - (1 + x)^{-1/2}$. Equation (20) is

$$\mathbb{E}g(x) = \frac{z^2}{2n}.$$

For every $x \ge 0$, concavity of the square root or direct differentiation yields $g(x) \le x/2$. Hence

$$\frac{z^2}{2n} \le \frac{\mathbb{E}U^2}{2(\tau^\circ)^2} = \frac{z^2\sigma^2}{2n(s^\circ)^2},$$

which gives $(s^\circ)^2 \le \sigma^2$.

For $0 \le x \le 1$, the ratio $g(x)/x$ is decreasing and therefore at least $g(1) = 1 - 2^{-1/2}$. Thus

$$\frac{z^2}{2n} = \mathbb{E}g(x) \ge \left(1 - \frac{1}{\sqrt{2}}\right)\mathbb{E}\left[x\mathbf{1}\{x \le 1\}\right] = \left(1 - \frac{1}{\sqrt{2}}\right)\frac{z^2\sigma^2_{\tau^\circ}}{n(s^\circ)^2}.$$

Rearranging proves the lower bound. □

Remark C.3. The lower bound depends on truncated variance. Uniform scale comparability therefore requires a regularity condition ensuring that a fixed truncation retains a constant fraction of total variance. Scale-boundary diagnostics complement this condition in practice.

C.1.2 Finite-Sample Calibration and Transfer

The robust fit is not needed for rank validity, but its estimation error matters when pooled scores are interpreted across contexts. We therefore expose the required one-context deviation event and record the external result that can instantiate it.

Fix $\delta_D \in (0, 1)$. For every retained context c, the fitting discrepancies are independent draws with location $m_{D,c}$, finite positive standard deviation $s_{D,c}$, and sample size $n_c \ge 5$ large enough for the source theorem below. Set

$$z^2_{D,c} = \log(n_c H/\delta_D) < 2n_c.$$

After mapping $v_0, V_0, \mu^\star, \sigma, \delta$ in Sun (2024, Theorems 3.4–3.5, arXiv v5) to $s_{\min,c}, s_{\max,c}, m_{D,c}, s_{D,c}, \delta_D/H$, respectively, assume all of those theorems' scale-bound, truncated-variance, curvature, and sample-size premises hold. Equivalently for the present analysis, assume there are deterministic constants $C_{m,c} > 0$ and $0 < L_{s,c} \le U_{s,c} < \infty$ such that

$$\mathbb{P}\left(|\widehat{m}_{D,c} - m_{D,c}| > C_{m,c}s_{D,c}\sqrt{\frac{\log(n_cH/\delta_D)}{n_c}} \text{ or } \widehat{s}_{D,c} \notin [L_{s,c}, U_{s,c}]\right) \le \frac{\delta_D}{H}.$$

Rare contexts handled by the frozen fallback hierarchy are excluded from this simultaneous statement.

The role of the next result is to lift verified one-context guarantees to the finite frozen context collection used by STAR.

Proposition C.4 (Simultaneous context fitting)**.** *Under the preceding contextwise deviation conditions, with probability at least $1-\delta_D$, simultaneously for all retained $c \in [H]$,*

$$|\widehat{m}_{D,c} - m_{D,c}| \le C_{m,c} s_{D,c} \sqrt{\frac{\log(n_c H/\delta_D)}{n_c}}, \tag{22}$$

$$L_{s,c} \le \widehat{s}_{D,c} \le U_{s,c}. \tag{23}$$

If the instantiated base theorem gives constant-multiple scale bounds under a fixed-fraction truncated-variance condition, the same constants hold simultaneously here.

Proof. Apply the assumed one-context statement at failure probability δ_D/H and take a union bound over the retained contexts. The final sentence merely preserves any explicitly instantiated constant-multiple implication of the base theorem; it introduces no additional probability event. □

Fix $\alpha \in (0,1)$. After fitting, compute for each $j \in \mathcal{I}^D_{\mathrm{cal}}$

$$S^D_j = \frac{D_j - \widehat{m}_{D,H_j}}{\widehat{s}_{D,H_j} + \varepsilon_D}. \tag{24}$$

Let $n_{\mathrm{cal}} = |\mathcal{I}^D_{\mathrm{cal}}| \ge 1$ and $k_\alpha = \lceil (n_{\mathrm{cal}}+1)(1-\alpha) \rceil$. If $k_\alpha \le n_{\mathrm{cal}}$, let $\widehat{q}_{1-\alpha}$ be the k_αth order statistic; otherwise set it to $+\infty$.

The fitting split determines a fixed score map; the disjoint calibration split then supplies finite-sample rank validity independently of estimator consistency.

Theorem C.5. *Conditional on $\mathcal{I}^D_{\mathrm{fit}}$, suppose the clean calibration pairs (D_j, H_j) and an independent new clean pair $(D_{\mathrm{new}}, H_{\mathrm{new}})$ are exchangeable. Then*

$$\mathbb{P}\left(S^D_{\mathrm{new}} > \widehat{q}_{1-\alpha} \,\middle|\, \mathcal{I}^D_{\mathrm{fit}}\right) \le \alpha. \tag{25}$$

The probability integrates over the calibration split and the new pair.

Proof. Condition on the fitting split. If $k_\alpha = n_{\mathrm{cal}} + 1$, then $\widehat{q}_{1-\alpha} = +\infty$ and the rejection event is empty. Otherwise attach independent continuous auxiliary variables to the $n_{\mathrm{cal}}+1$ scores and rank the resulting pairs lexicographically. Exchangeability makes the new pair's randomized rank uniform on $\{1, \dots, n_{\mathrm{cal}}+1\}$. The strict event $S^D_{\mathrm{new}} > \widehat{q}_{1-\alpha}$ implies that this rank exceeds k_α; ties can only remove strict rejections. Its probability is therefore at most

$$\frac{n_{\mathrm{cal}} + 1 - k_\alpha}{n_{\mathrm{cal}} + 1} \le \alpha.$$

□

Self-tuning improves the balance of pooled calibration when contexts differ mainly by location and scale, while exchangeability supplies conformal validity. The next result quantifies this transfer at a fixed threshold.

Proposition C.6 (Approximate context equalization)**.** *Condition on $\mathcal{I}^D_{\mathrm{fit}}$ and let $(D_{\mathrm{new}}, H_{\mathrm{new}})$ be an independent fresh clean pair. Suppose the clean residual CDF in context c satisfies* (12) *and the reference CDF F_D has density bounded by M_D. On the $\mathcal{I}^D_{\mathrm{fit}}$-measurable event where*

$$\left|\frac{\widehat{m}_{D,c} - m_{D,c}}{s_{D,c}}\right| \le a_c, \quad \left|\frac{\widehat{s}_{D,c}}{s_{D,c}} - 1\right| \le b_c < 1,$$

the following inequality holds simultaneously for every $q \in \mathbb{R}$:

$$\left|\mathbb{P}\left\{S^D_{\mathrm{new}} > q \,\middle|\, \mathcal{I}^D_{\mathrm{fit}}, H_{\mathrm{new}} = c, Z_{\mathrm{new}} = 0\right\} - [1 - F_D(q)]\right| \le \eta_{\mathrm{het}} + M_D\left(a_c + |q| b_c + \frac{|q|\varepsilon_D}{s_{D,c}}\right). \tag{26}$$

Proof. Condition throughout on the fitting split. Write $D = m_{D,c} + s_{D,c}\epsilon_c$. The event $S^D > q$ is equivalent to

$$\epsilon_c > \frac{\widehat{m}_{D,c} - m_{D,c}}{s_{D,c}} + q\frac{\widehat{s}_{D,c} + \varepsilon_D}{s_{D,c}}.$$

Relative to q, the threshold displacement is at most $a_c + |q|b_c + |q|\varepsilon_D/s_{D,c}$. The density bound converts this displacement into a reference-CDF difference, and (12) contributes η_{het}. □

The static rank guarantee does not condition on an adaptively trained policy. To transfer it to round t, the following corollary makes the feedback dependence explicit through a conditional distributional-stability premise.

Corollary C.7. *Let $\mathcal{G} = \sigma(\mathcal{I}^D_{\mathrm{fit}}, \mathcal{I}^D_{\mathrm{cal}})$ and let $\mathcal{H}_t \supseteq \mathcal{G}$ be the pre-rollout training history. Let $P^0_{\mathrm{ref}}(\cdot \mid \mathcal{I}^D_{\mathrm{fit}})$ denote the law of an independent clean reference pair exchangeable with the calibration pairs. If, almost surely,*

$$\mathrm{TV}\big(P^0_t(\cdot \mid \mathcal{H}_t), P^0_{\mathrm{ref}}(\cdot \mid \mathcal{I}^D_{\mathrm{fit}})\big) \le \rho_t,$$

then the outer marginal probability satisfies

$$\mathbb{P}_t\{S^D > \widehat{q}_{1-\alpha}\} \le \alpha + \rho_t.$$

If ρ_t is random, the right-hand side is replaced by $\alpha + \mathbb{E}\rho_t$.

Proof. For every realization of $\mathcal{H}_t$, total-variation duality gives

$$P^0_t(S^D > \widehat{q}_{1-\alpha} \mid \mathcal{H}_t) \le P^0_{\mathrm{ref}}(S^D > \widehat{q}_{1-\alpha} \mid \mathcal{G}) + \rho_t.$$

Average over the history. The reference draw is independent of the calibration split conditional on the fitting split, so Theorem C.5 bounds the first expectation by α. Averaging the drift term proves the claim. □

C.1.3 Supporting Calculus and Boundary Conditions

For completeness, the derivatives that underlie the convexity and numerical diagnostics can be written in closed form. For $u = d - m$, $a = z^2u^2/(ns^2)$, and

$$\ell(u, s) = \frac{ns}{z^2}(\sqrt{1+a} - 1) + \frac{s}{2},$$

the derivatives are

$$\partial_m \ell = -\frac{u}{s\sqrt{1+a}}, \tag{27}$$

$$\partial_s \ell = \frac{n}{z^2}\left(\frac{1}{\sqrt{1+a}} - 1\right) + \frac{1}{2}, \tag{28}$$

$$\partial^2_{mm} \ell = \frac{1}{s(1+a)^{3/2}}, \tag{29}$$

$$\partial^2_{ms} \ell = \frac{u}{s^2(1+a)^{3/2}}, \tag{30}$$

$$\partial^2_{ss} \ell = \frac{u^2}{s^3(1+a)^{3/2}}. \tag{31}$$

Therefore the Hessian equals

$$\nabla^2_{m,s} \ell = \frac{1}{s^3(1+a)^{3/2}} \begin{bmatrix} s^2 & su \\ su & u^2 \end{bmatrix} = \frac{1}{s^3(1+a)^{3/2}} \begin{bmatrix} s \\ u \end{bmatrix} \begin{bmatrix} s & u \end{bmatrix}, \tag{32}$$

which is positive semidefinite and rank one for one observation. Two observations with distinct residuals yield linearly independent vectors (s, u_j) and a positive-definite sum.

The same calculus clarifies when the population scale exists. For nondegenerate U, define

$$H(s) = \mathbb{E}\left[(1 + z^2U^2/(ns^2))^{-1/2}\right].$$

The function is continuous and nondecreasing in s. If $\mathbb{P}(U = 0) = 0$, then $H(s) \to 0$ as $s \downarrow 0$ and $H(s) \to 1$ as $s \to \infty$. Thus for $0 < z^2/(2n) < 1$, equation (20) has a solution. If U has an atom at zero, existence holds when $\mathbb{P}(U = 0) < 1 - z^2/(2n)$. Strict monotonicity follows whenever $\mathbb{P}(U \neq 0) > 0$, yielding uniqueness.

For a large-sample interpretation, suppose s_n° solves (20) with $z_n^2 = o(n)$ and the family s_n° stays away from zero. Then $\tau_n^\circ = \sqrt{n} s_n^\circ / z_n \to \infty$, and dominated convergence gives $\sigma^2_{\tau_n^\circ} \to \sigma^2$. The population scale calculation above therefore sandwiches $(s_n^\circ)^2$ between a fixed multiple of a quantity converging to σ^2 and σ^2. The sharper conclusion $s_n^\circ \to \sigma$ follows from a first-order expansion of $g(x)$ and uniform integrability; this is the self-tuned interpretation established in the source robust-estimation theory.

At a finite scale boundary, the equality is replaced by a KKT condition. For empirical objective $L_c(m, s)$ on $[s_{\min,c}, s_{\max,c}]$, the location equation always holds at an optimizer because m is unconstrained, while

$$\begin{aligned} \partial_s L_c(\widehat{m}, \widehat{s}) &= 0 && \text{if } s_{\min,c} < \widehat{s} < s_{\max,c}, \\ \partial_s L_c(\widehat{m}, s_{\min,c}) &\geq 0 && \text{if } \widehat{s} = s_{\min,c}, \\ \partial_s L_c(\widehat{m}, s_{\max,c}) &\leq 0 && \text{if } \widehat{s} = s_{\max,c}. \end{aligned}$$

Every experiment must report the proportion of contexts or minibatches hitting a boundary.

C.2 Conformal Reliability, Separation, and Drift

The calibration analysis has three components: a rank guarantee for a new clean pair, a tail-separation premise for attacked pairs, and a total-variation transfer condition for adaptive policy rounds.

To distinguish validity from power, first consider how accurately the empirical threshold localizes a population quantile. Let calibration scores be $S_1, \ldots, S_n$ and $S_{(1)} \leq \cdots \leq S_{(n)}$. Put $k = \lceil (n+1)(1-\alpha) \rceil$. If $k = n+1$, the finite-sample threshold is $+\infty$, so a nontrivial threshold requires $\alpha \geq 1/(n+1)$ under this convention.

Let $\widehat{F}_n$ be the empirical CDF and suppose

$$\sup_x |\widehat{F}_n(x) - F_0(x)| \leq \epsilon_n.$$

At $q = S_{(k)}$, $\widehat{F}_n(q) \geq k/n = u_k$, so $F_0(q) \geq u_k - \epsilon_n$, which implies

$$q \geq F_0^{-1}((u_k - \epsilon_n)_+).$$

For any $x < q$, at most $k-1$ observations are at or below x, so $\widehat{F}_n(x) \leq (k-1)/n < u_k$. If $F_0(x) > u_k + \epsilon_n$, then $\widehat{F}_n(x) > u_k$, a contradiction. Under the strict-increase condition stated in Section C.2, this implies $q \leq F_0^{-1}((u_k + \epsilon_n) \wedge 1)$. DKW gives the uniform event with probability at least $1 - 2e^{-2n\epsilon_n^2} = 1 - \delta$.

If $F_0 = \Phi$ and attacked standardized scores are $N(\Delta, 1)$, then on the DKW event

$$\mathrm{TPR} \geq 1 - \Phi\left(\Phi^{-1}((u_k + \epsilon_n) \wedge 1) - \Delta\right).$$

This formula is a power statement, not a validity statement. It describes how much standardized separation is needed after accounting for finite calibration size.

Suppose attacked scores have mean Δ and variance at most σ_1^2, and the threshold is deterministically bounded by $q_+ < \Delta$. Cantelli's inequality gives

$$\mathbb{P}(S \leq q_+ \mid Z = 1) = \mathbb{P}(S - \Delta \leq -(\Delta - q_+)) \leq \frac{\sigma_1^2}{\sigma_1^2 + (\Delta - q_+)^2}.$$

Thus TPR is at least $(\Delta - q_+)^2/[\sigma_1^2 + (\Delta - q_+)^2]$. Under finite variance, the corresponding bound is polynomial rather than exponential.

The simultaneous context-equalization bound above holds simultaneously over thresholds on the fitting event. Hence, on the event $\{|\widehat{q}_{1-\alpha}| \leq Q\}$, for a fresh clean pair independent of the calibration

split conditional on $\mathcal{I}^D_{\text{fit}}$, write $\mathcal{C}_c$ for conditioning on $(\mathcal{I}^D_{\text{fit}}, \mathcal{I}^D_{\text{cal}}, H_{\text{new}} = c, Z_{\text{new}} = 0)$. Then

$$\left|\mathbb{P}\big(S^D_{\text{new}} > \widehat{q}_{1-\alpha} \,\big|\, \mathcal{C}_c\big) - [1 - F_D(\widehat{q}_{1-\alpha})]\right| \le \eta_{\text{het}} + M_D \left(a_c + Qb_c + \frac{Q\varepsilon_D}{s_{D,c}}\right).$$

Relating the reference tail $1 - F_D(\widehat{q}_{1-\alpha})$ to α, or to the actual pooled clean tail, requires the empirical-quantile localization above and its failure probability; no exact group-conditional conformal guarantee is implied. We now convert this calibrated score into the rollout reliability used by STAR.

For a training rollout, calculate

$$S^D_{b,i} = \frac{D_{b,i} - \widehat{m}_{D,H_{b,i}}}{\widehat{s}_{D,H_{b,i}} + \varepsilon_D} \tag{33}$$

and, for $\lambda > 0$, define

$$w_{b,i} = \exp\{-\lambda[S^D_{b,i} - \widehat{q}_{1-\alpha}]_+\}. \tag{34}$$

When $\widehat{q}_{1-\alpha} = +\infty$, the convention is $w_{b,i} = 1$. The calibration split determines where down-weighting begins; λ determines its post-threshold slope and is selected using separate validation data.

The basic shape of this map is used both in the attack attenuation theorem and in the local sensitivity calculation.

Proposition C.8 (Regularity of the soft reliability map)**.** *For every rollout,* $0 < w_{b,i} \le 1$. *If* $S^D_{b,i} \le \widehat{q}_{1-\alpha}$, *then* $w_{b,i} = 1$. *The weight is nonincreasing and globally* λ*-Lipschitz in* $S^D_{b,i}$, *and differentiable away from the threshold.*

Proof. Below the threshold the map is constant. Above it, $w(s) = e^{-\lambda(s-\widehat{q})}$, so $w'(s) = -\lambda w(s)$ and $|w'(s)| \le \lambda$. Continuity at the threshold proves global Lipschitzness. □

The conformal event can now be translated into a statement about the operational reliability weight. The separation condition is a tail condition, not a claim that a mean shift alone gives detection power.

Theorem C.9. *Under Theorem C.5, for a new clean rollout,*

$$\mathbb{P}(w < 1 \mid Z = 0, \mathcal{I}^D_{\text{fit}}) \le \alpha, \quad \mathbb{E}[1 - w \mid Z = 0, \mathcal{I}^D_{\text{fit}}] \le \alpha. \tag{35}$$

Let $\Delta \ge 0$ *and* $\beta \in [0, 1]$. *If an attacked rollout satisfies*

$$\mathbb{P}\{S^D \ge \widehat{q}_{1-\alpha} + \Delta \mid Z = 1\} \ge 1 - \beta,$$

then

$$\mathbb{E}[w \mid Z = 1] \le \beta + (1 - \beta)e^{-\lambda\Delta}. \tag{36}$$

Proof. The event $w < 1$ equals $S^D > \widehat{q}_{1-\alpha}$, and $0 \le 1 - w \le \mathbf{1}\{w < 1\}$. For the attack statement, let $E = \{S^D \ge \widehat{q}_{1-\alpha} + \Delta\}$. On E, $w \le e^{-\lambda\Delta}$; on E^c, $w \le 1$. Splitting the expectation proves the bound. □

Remark C.10 (Majority and fully attacked groups). The discrepancy baseline is external to the current group, so a common positive shift is not redefined as normal. Relative weights alone would nevertheless cancel if all group members were equally suspicious. STAR addresses this second failure through the absolute factor $\bar{w}_b$ and the primary $W_{\min}$ abstention rule.

For comparison, a hard-gating baseline is

$$w^{\text{hard}}_{b,i} = \mathbf{1}\{S^D_{b,i} \le \widehat{q}_{1-\alpha}\}.$$

It provides an interpretable accepted set but creates discontinuous active-set changes. The primary STAR method uses soft individual weights together with deterministic group skipping at $W_{\min}$ or $G_{\min}$.

Quantitative power requires more than exchangeability. Conditional on the fitted score map, suppose the calibration scores are iid from a continuous CDF F_0, the attacked test score is independent with

CDF F_1, and F_0 is strictly increasing over the relevant quantile neighborhood. With $u_k = k_\alpha / n_{\mathrm{cal}}$ and

$$\epsilon_n = \sqrt{\frac{\log(2/\delta)}{2n_{\mathrm{cal}}}},$$

when $k_\alpha \le n_{\mathrm{cal}}$, the Dvoretzky–Kiefer–Wolfowitz localization derived above implies, with probability at least $1-\delta$,

$$\mathbb{P}_{F_1}\{S^D > \widehat{q}_{1-\alpha}\} \ge 1 - F_1\big(F_0^{-1}((u_k+\epsilon_n)\wedge 1)\big)\,. \tag{37}$$

Under stochastic shift $F_1(t) \le F_0(t-\Delta)$, replace F_1 on the right by $F_0(\cdot - \Delta)$. This remains a power statement rather than a validity guarantee.

The same map supplies several implementation diagnostics.

By the preceding Lipschitz property, if estimated discrepancy scores have error $|\widehat{S} - S^\star| \le \epsilon_S$, then

$$|\widehat{w} - w^\star| \le \lambda \epsilon_S.$$

This deterministic relation is useful in the local oracle comparison.

Calibration drift has an equally direct diagnostic implication. Under Corollary C.7, for deterministic ρ_t,

$$\mathbb{P}_t(w < 1 \mid Z = 0) \le \alpha + \rho_t, \quad \mathbb{E}_t[1 - w \mid Z = 0] \le \alpha + \rho_t.$$

For random ρ_t, both right-hand sides are replaced by $\alpha + \mathbb{E}\rho_t$. If drift is not controlled, calibration diagnostics rather than nominal conformal levels must be used.

Reliability also determines the effective sample size. For nonnegative weights, $1 \le G_{\mathrm{eff}} \le G$ whenever at least one weight is positive. The upper bound is Cauchy–Schwarz:

$$(\sum_i w_i)^2 \le G \sum_i w_i^2.$$

The lower bound follows from $(\sum_i w_i)^2 \ge \sum_i w_i^2$. Equality $G_{\mathrm{eff}} = G$ occurs for equal weights; $G_{\mathrm{eff}} = 1$ occurs when only one weight is nonzero. Soft weights are strictly positive mathematically, but numerical underflow can make them zero, motivating stable log-weight computation.

Finally, the trusted external reference is essential. Suppose every discrepancy in a group is $D_i = M + \xi_i$, where M is a large attack shift and ξ_i has small spread. Any translation-equivariant group location estimator gives $\widehat{m}_D \approx M$, and standardized within-group discrepancies depend mainly on ξ_i. Hence weights based on $(D_i - \widehat{m}_D)/\widehat{s}_D$ approach clean-looking values even as $M \to \infty$. External trusted calibration removes this translation invariance with respect to the attacked group.

D Reliability-First Advantages and Optimization

The weighted robust objective yields zero-sum bounded advantages and a controlled initial reward-side direction at the old policy.

D.1 Weighted Robust Objective and Advantage Geometry

Reliability enters the location–scale fit before normalization, and the leading group factor carries absolute reliability into the update. The following derivations establish convexity, KKT equations, zero sum, coordinate bounds, and common-weight scaling.

D.1.1 Weighted Location–Scale Fit

GRPO typically uses few rollouts per prompt. Estimating an independent location and scale from $G \in \{4, 8, 16\}$ observations is unstable, especially after downweighting. STAR therefore assigns one robust location μ_b to each prompt and one scale v to the minibatch. The model is descriptive rather than a claim that all prompts have identical population variance: the shared scale is an update normalization fitted from BG weighted residuals.

Let $\mathcal{B}_{\mathrm{adm}} \subseteq [B]$ denote the fixed admissible group set and $B_{\mathrm{adm}} = |\mathcal{B}_{\mathrm{adm}}| > 0$. For $b \in \mathcal{B}_{\mathrm{adm}}$, define $p_{b,i}$ and $G_{\mathrm{eff},b}$ by (14). Fix $z_A > 0$ and $0 < v_{\min} < v_{\max}$. The weighted objective is

$$\mathcal{L}_A(\mu_{\mathcal{B}_{\mathrm{adm}}}, v) = \frac{1}{B_{\mathrm{adm}}} \sum_{b \in \mathcal{B}_{\mathrm{adm}}} \sum_{i=1}^{G} p_{b,i} \left[\frac{G_{\mathrm{eff},b} v}{z_A^2} \left(\sqrt{1 + \frac{z_A^2 (r_{b,i} - \mu_b)^2}{G_{\mathrm{eff},b} v^2}} - 1 \right) + \frac{v}{2} \right]. \tag{38}$$

Define

$$(\widehat{\mu}_{\mathcal{B}_{\mathrm{adm}}}, \widehat{v}) \in \operatorname*{arg\,min}_{\substack{\mu_b \in \mathbb{R},\ b \in \mathcal{B}_{\mathrm{adm}} \\ v_{\min} \le v \le v_{\max}}} \mathcal{L}_A. \tag{39}$$

All reward, discrepancy, reliability, and fitted quantities are stop-gradient inputs to policy optimization.

Let

$$a_{b,i}(\mu_b, v) = \frac{z_A^2 (r_{b,i} - \mu_b)^2}{G_{\mathrm{eff},b} v^2}.$$

Because every μ_b is unconstrained, every minimizer satisfies

$$\sum_{i=1}^{G} p_{b,i} \frac{r_{b,i} - \widehat{\mu}_b}{\widehat{v} \sqrt{1 + a_{b,i}(\widehat{\mu}_b, \widehat{v})}} = 0, \quad b \in \mathcal{B}_{\mathrm{adm}}. \tag{40}$$

If $\widehat{v} \in (v_{\min}, v_{\max})$, it also satisfies

$$\frac{1}{B_{\mathrm{adm}}} \sum_{b \in \mathcal{B}_{\mathrm{adm}}} \left[\frac{G_{\mathrm{eff},b}}{z_A^2} \sum_{i=1}^{G} p_{b,i} \left(\frac{1}{\sqrt{1 + a_{b,i}(\widehat{\mu}_b, \widehat{v})}} - 1 \right) + \frac{1}{2} \right] = 0. \tag{41}$$

At a scale boundary, this equality is replaced by the corresponding KKT inequality.

The perspective form makes the fit globally auditable once the weights and active groups are frozen. The next result ensures that the solver targets a single statistical object rather than a local nonconvex surrogate.

Theorem D.1. *Conditional on a fixed admissible set and fixed nonnegative weights with positive total mass in each admitted group,* (38) *is jointly convex in* $(\mu_{\mathcal{B}_{\mathrm{adm}}}, v)$ *on* $v > 0$*. If every admissible prompt has at least two distinct positive-weight quality rewards, the objective is strictly convex and the constrained minimizer is unique.*

Proof. For each (b, i), the bracketed term is $v\rho_b((r_{b,i} - \mu_b)/v) + v/2$, where

$$\rho_b(t) = \frac{G_{\mathrm{eff},b}}{z_A^2} \left(\sqrt{1 + \frac{z_A^2 t^2}{G_{\mathrm{eff},b}}} - 1 \right)$$

is strictly convex. Its perspective is jointly convex. In the Hessian representation of Appendix D.1.2, two distinct residuals in each prompt identify its location coordinate and the shared scale direction, so the weighted Hessian sum is positive definite. Restriction to the convex scale interval preserves uniqueness. □

The reliability w is a function of $R^{\mathrm{obs}} - R^{\mathrm{canon}}$, while $r = h(R_\kappa)$ generally depends on both views. The convexity and advantage results are deterministic conditional on observed (r, w) and require no independence. Interpreting a fitted location as an unselected population parameter would require an additional weighted estimating-equation or cross-fitting condition; STAR does not identify it with human utility.

D.1.2 Objective Geometry and Numerical Conditions

For a fixed prompt b, set $n_b = G_{\mathrm{eff,b}}$, $u_{b,i} = r_{b,i} - \mu_b$, and $a_{b,i} = z_A^2 u_{b,i}^2/(n_b v^2)$. The Hessian of one bracketed term in (38) with respect to (μ_b, v) is

$$\frac{1}{v^3 (1 + a_{b,i})^{3/2}} \begin{bmatrix} v^2 & v u_{b,i} \\ v u_{b,i} & u_{b,i}^2 \end{bmatrix}.$$

Multiplication by $p_{b,i}$ preserves positive semidefiniteness. In the full parameter vector $(\mu_b : b \in \mathcal{B}_{\mathrm{adm}}, v)$, each observation contributes a rank-one vector supported on coordinate b and the shared-scale coordinate. Strict convexity requires enough distinct residuals in every prompt to identify each μ_b and aggregate variation to identify v.

At an interior optimum, (41) can be rewritten as

$$\frac{1}{B_{\mathrm{adm}}} \sum_{b \in \mathcal{B}_{\mathrm{adm}}} \frac{G_{\mathrm{eff,b}}}{z_A^2} \left[1 - \sum_{i=1}^{G} p_{b,i} \frac{1}{\sqrt{1 + a_{b,i}}} \right] = \frac{1}{2}.$$

A second-order expansion $1 - (1+a)^{-1/2} \approx a/2$ gives

$$\widehat{v}^2 \approx \frac{1}{B_{\mathrm{adm}}} \sum_{b \in \mathcal{B}_{\mathrm{adm}}} \sum_{i=1}^{G} p_{b,i} (r_{b,i} - \widehat{\mu}_b)^2.$$

This is an interpretation, not an exact identity. Large residuals contribute subquadratically through the exact equation.

For fixed v, each location objective is strictly convex and its derivative is monotone, so bisection is globally safe. For fixed locations, the scale objective is convex on $v > 0$ and one dimensional. Alternating exact minimization decreases the joint objective and converges to a global minimizer because the objective is convex and level sets are compact under the scale bounds and bounded rewards. A practical implementation may use a warm start from the previous minibatch.

When effective reliable mass is insufficient, the unique primary rule is *skip*: set the group's reward advantages to zero, exclude it from the shared-scale fit, retain its entries in the loss tensor, and keep the original BG denominator. This is the equal-response reduction in the reference protocol. Both recorded STAR runs use zero reward advantages for skipped groups and retain their token entries in the token-mean denominator, as analyzed in Corollary D.6; this differs from the equal-response reference reduction.

D.1.3 STAR Advantage and Deterministic Bounds

For an admissible group, define

$$\varphi_{b,i} = \frac{z_A}{\sqrt{G_{\mathrm{eff},b}}} \frac{r_{b,i} - \widehat{\mu}_b}{\widehat{v} \sqrt{1 + z_A^2 (r_{b,i} - \widehat{\mu}_b)^2 / (G_{\mathrm{eff},b} \widehat{v}^2)}}. \tag{42}$$

For small residuals the score is approximately linear, while large residuals saturate. The following elementary bound is what makes the subsequent geometry distribution free.

Lemma D.2 (Bounded self-tuned score)**.** *For every admissible rollout,* $|\varphi_{b,i}| \le 1$.

Proof. Let $x = z_A (r_{b,i} - \widehat{\mu}_b) / (\sqrt{G_{\mathrm{eff},b}} \widehat{v})$. Then $\varphi = x / \sqrt{1 + x^2}$. □

Let

$$\nu_b = \left(\frac{1}{G} \sum_{j=1}^{G} w_{b,j}^2 \right)^{1/2}.$$

The STAR advantage is

$$A_{b,i}^{\mathrm{STAR}} = \bar{w}_b \frac{w_{b,i}}{\nu_b + \varepsilon_w} \varphi_{b,i}. \tag{43}$$

Relative weights $p_{b,i}$ act inside the fitted location and scale. The factor $\bar{w}_b$ carries the group's absolute reliability to the policy update, closing the common-scale cancellation present when only w_i / ν_b is used.

The next theorem distinguishes magnitude control from centering. Magnitude control uses the formula alone. At an exact fit, the unconstrained location coordinate satisfies its first-order equation even when the shared scale is at a boundary.

Theorem D.3. *Suppose group b is admissible, $w_{b,i} \in [0,1]$, and $\varepsilon_w \ge 0$. Evaluate Equation* (43) *at any finite location and positive scale. Then:*

1. Conditional zero sum: *if* $\sum_i p_{b,i}\varphi_{b,i} = 0$*, then* $\sum_i A^{\mathrm{STAR}}_{b,i} = 0$*. This additional condition is not needed for the remaining statements.*

2. Empirical second moment:

$$\frac{1}{G}\sum_{i=1}^{G}(A^{\mathrm{STAR}}_{b,i})^2 \le \bar{w}_b^2\left(\frac{\nu_b}{\nu_b+\varepsilon_w}\right)^2 \le \bar{w}_b^2.$$

3. Individual bound:

$$|A^{\mathrm{STAR}}_{b,i}| \le \bar{w}_b\frac{w_{b,i}}{\nu_b+\varepsilon_w} \le w_{b,i} \le 1.$$

4. Clean reduction: *if every* $w_{b,i} = 1$*, then* $\bar{w}_b = \nu_b = 1$ *and STAR reduces to the self-tuned anchored-quality score divided by* $1+\varepsilon_w$*.*

5. Separated attack attenuation: *if* $w_{b,i} \le e^{-\lambda\Delta}$*, then* $|A^{\mathrm{STAR}}_{b,i}| \le e^{-\lambda\Delta}$*.*

Proof. For the first statement only, the location-equation premise gives $\sum_i p_{b,i}\varphi_{b,i} = 0$. Since $p_{b,i} = w_{b,i}/W_b$, multiplying by the common factor $\bar{w}_b/(\nu_b+\varepsilon_w)$ proves zero sum. The bounded-score calculation above gives

$$\frac{1}{G}\sum_{i}(A^{\mathrm{STAR}}_{b,i})^2 \le \bar{w}_b^2\frac{G^{-1}\sum_i w_{b,i}^2}{(\nu_b+\varepsilon_w)^2}.$$

Finally, RMS dominates the arithmetic mean, so $\bar{w}_b \le \nu_b$. This proves the individual bound; the last two statements follow by substitution. □

For a numerical solve, let $e_b^{\varphi} = |\sum_i p_{b,i}\varphi_{b,i}|$ be the reported location-equation residual. Then

$$\left|\sum_i A^{\mathrm{STAR}}_{b,i}\right| = \frac{\bar{w}_b W_b}{\nu_b+\varepsilon_w}\, e_b^{\varphi}.$$

Thus exact zero sum requires an exact location equation, while the magnitude bounds continue to hold for arbitrary finite fitted locations and positive scale. The centering deviation can be audited by recording the location-equation residual; certifying the joint optimum also requires the scale KKT condition. These statements concern real arithmetic and do not certify a particular floating-point implementation. The primary numerical protocol requires a final safeguarded location solve at the fitted scale and residual checks before accepting a policy step. RQ2's recorded implementation and its more limited diagnostics are specified in Appendix F.2.

Uniformly shrinking every weight should weaken the group update rather than be removed by normalization. The group factor makes this behavior exact when the numerical floor is zero and further attenuates the update with a positive numerical floor.

Proposition D.4 (Common reliability scaling)**.** *Fix the reward arrays and all other groups, and replace one group's weights* w *by* cw *for* $c \in (0,1]$*, without changing the full admissible set. Use the same solution selection for the unchanged fitting objective. The relative weights, effective size, fitted parameters, and* φ *are unchanged. If* $\varepsilon_w = 0$*, then* $A^{\mathrm{STAR}}(cw) = cA^{\mathrm{STAR}}(w)$*. If* $\varepsilon_w > 0$*, every coordinate is multiplied by*

$$\frac{c^2(\nu_b+\varepsilon_w)}{c\nu_b+\varepsilon_w} \le c.$$

Proof. Both $p_i = w_i/W_b$ and $G_{\mathrm{eff}} = W_b^2/\sum_i w_i^2$ are invariant to a common positive scale, while $\bar{w}_b$ and ν_b each scale by c. Substitution in (43) proves the claim. □

For the same quality vector $r = (0,\dots,0,M)$, ordinary GRPO has one advantage $\sqrt{G-1}$ and $G-1$ advantages $-1/\sqrt{G-1}$, independently of $M > 0$. Multiplying only the outlier by a small reliability weight does not undo the negative clean signs already assigned by the contaminated normalization. Reliability-first fitting instead sends the outlier's mass to zero before solving the location equation.

Proof of Proposition 5.2. Write $n = G - 1$. The ordinary mean and population-form standard deviation of $(0, \ldots, 0, M)$ are M/G and $M\sqrt{n}/G$. Every clean standardized reward is therefore $-1/\sqrt{n}$, unchanged by post-hoc multiplication by its unit weight.

For the weighted comparator, define

$$\mu_w = \frac{\sum_i w_i r_i}{n + \epsilon} = \frac{\epsilon M}{n + \epsilon}, \qquad \sigma_w^2 = \frac{\sum_i w_i (r_i - \mu_w)^2}{n + \epsilon} = \frac{n\epsilon M^2}{(n + \epsilon)^2}.$$

Substituting into $\widetilde{A}_i = w_i(r_i - \mu_w)/\sigma_w$ gives $\widetilde{A}_{\text{clean}} = -\sqrt{\epsilon/n}$ and $\widetilde{A}_{\text{outlier}} = \sqrt{n\epsilon}$. Thus weighting before normalization attenuates every coordinate even with the ordinary weighted mean and variance.

For STAR, Theorem D.3 gives $|A_G^{\text{STAR}}| \le \epsilon$ without an exact solve. With the location equation satisfied, all n clean coordinates are equal and zero sum gives $nA_{\text{clean}}^{\text{STAR}} = -A_G^{\text{STAR}}$. Hence $|A_{\text{clean}}^{\text{STAR}}| \le \epsilon/n$. This is a finite-$\epsilon$ bound for every admitted group, including when its scale is shared with other groups.

For the location limit, $W = n + \epsilon$ and $G_{\text{eff}} = (n + \epsilon)^2/(n + \epsilon^2) \to n$, so the stated strict admission thresholds ensure eventual admission. The fitted location lies in $[0, M]$, and the scale lies in the fixed compact interval. Along any convergent subsequence of exact fits, the location equation is

$$n\varphi(-\widehat{\mu}_\epsilon, \widehat{v}_\epsilon; G_{\text{eff}}) + \epsilon\varphi(M - \widehat{\mu}_\epsilon, \widehat{v}_\epsilon; G_{\text{eff}}) = 0,$$

where $\varphi(u, v; n') = (z_A u/(\sqrt{n'}v))/\sqrt{1 + z_A^2 u^2/(n'v^2)}$. Continuity and $|\varphi| \le 1$ imply $\varphi(-\mu_\star, v_\star; n) = 0$ at the limit, hence $\mu_\star = 0$. Compactness then gives $\widehat{\mu}_\epsilon \to 0$ for the full sequence. The displayed advantage bounds already imply that all coordinates vanish. □

The comparison uses no standard-deviation stabilizer. Replacing σ_w by $\sigma_w + \delta$ with a fixed $\delta > 0$ gives $O(\epsilon)$ coordinates as $\epsilon \downarrow 0$ for fixed M; its constants can depend on M/δ. STAR's $|A_i| \le w_i$ is uniform over finite reward arrays and positive scales. The proposition is an algebraic design comparison, not an empirical ablation or a claim that every alternative robust estimator lacks such a bound.

D.1.4 Algebraic Consequences and Design Variants

Several short calculations clarify the role of numerical stabilization and the choice of denominator. The epsilon ε_w is common within a prompt, so it preserves zero sum when the weighted location equation holds:

$$\sum_i A_{b,i}^{\text{STAR}} = \frac{\bar{w}_b}{\nu_b + \varepsilon_w} \sum_i w_{b,i}\varphi_{b,i} = 0.$$

It only shrinks the second moment relative to the ideal RMS normalization.

If h is the identity and anchored quality rewards are transformed as $r'_{b,i} = ar_{b,i} + c_b$ with $a > 0$, the prompt locations transform exactly as $a\widehat{\mu}_b + c_b$ and the shared scale as $a\widehat{v}$ when both scale bounds are multiplied by a. The normalized score φ and reliability factors are invariant. The bounded nonlinear map h intentionally breaks exact affine equivariance to control optimizer gradients.

The one-outlier post-hoc calculation has already been established in the proof of Proposition 5.2. A denominator comparison isolates the role of the leading absolute factor $\bar{w}_b$. The mean-denominator ablation is

$$A_{b,i}^{\text{mean}} = \bar{w}_b \frac{w_{b,i}}{\bar{w}_b + \varepsilon_w}\varphi_{b,i},$$

whereas STAR uses the RMS denominator ν_b. Because $\nu_b \ge \bar{w}_b$, the RMS variant has no larger coordinate magnitude and yields the sharper energy bound in Theorem D.3. With one unit weight and the rest zero, the mean-denominator factor on the surviving score is one for $\varepsilon_w = 0$, while STAR's factor is $1/\sqrt{G}$. The leading $\bar{w}_b$, rather than the denominator choice alone, is what preserves common low reliability.

D.2 Initial Policy Direction and Its Scope

The advantage bounds below apply to the initial reward-side direction at the old policy. The clipped policy objective is stated first to identify this direction within the training update.

D.2.1 Policy Objective and Old-Policy Guarantee

For token t of rollout (b, i), let

$$\rho_{b,i,t}(\theta) = \frac{\pi_\theta(o_{b,i,t} \mid X_b, o_{b,i,<t})}{\pi_{\theta_{\text{old}}}(o_{b,i,t} \mid X_b, o_{b,i,<t})}, \quad L_{b,i} = |O_{b,i}|.$$

Define the clipped token term

$$\ell^{\text{clip}}_{b,i,t}(\theta) = \min\left\{\rho_{b,i,t}(\theta)A^{\text{STAR}}_{b,i}, \text{clip}(\rho_{b,i,t}(\theta), 1-\epsilon_{\text{clip}}, 1+\epsilon_{\text{clip}})A^{\text{STAR}}_{b,i}\right\}.$$

With $A^{\text{STAR}}_{b,i} = 0$ for skipped groups, STAR maximizes

$$J_{\text{STAR}}(\theta) = \mathbb{E}\left[\frac{1}{BG}\sum_{b=1}^{B}\sum_{i=1}^{G}\frac{1}{L_{b,i}}\sum_t \ell^{\text{clip}}_{b,i,t}(\theta) - \beta_{\text{KL}}D_{\text{KL}}(\pi_\theta\|\pi_{\text{ref}})\right]. \tag{44}$$

The old and reference policies are frozen for the inner optimization epoch, and gradients do not propagate through reward calls, canonicalization, reliability, fitted locations, or scale.

At θ_{old}, every likelihood ratio equals one. The following result bounds the resulting reward-side gradient; the KL term is handled separately by the policy objective.

Define

$$g^{\text{avg}}_{b,i} = \frac{1}{L_{b,i}}\sum_t \nabla_\theta \log \pi_{\theta_{\text{old}}}(o_{b,i,t} \mid X_b, o_{b,i,<t}), \quad U^{\text{STAR}}_b = \frac{1}{G}\sum_i A^{\text{STAR}}_{b,i} g^{\text{avg}}_{b,i}.$$

Theorem D.5. *If $\|g^{\text{avg}}_{b,i}\|_2 \le L_\pi$, every admissible group satisfies*

$$\|U^{\text{STAR}}_b\|_2 \le \bar{w}_b L_\pi.$$

Moreover, rollout i's contribution is bounded by

$$\left\|\frac{1}{G}A^{\text{STAR}}_{b,i} g^{\text{avg}}_{b,i}\right\|_2 \le \frac{L_\pi w_{b,i}}{G}.$$

Proof. Cauchy–Schwarz and Theorem D.3 give

$$\|U^{\text{STAR}}_b\|_2 \le \frac{1}{G}\left(\sum_i (A^{\text{STAR}}_{b,i})^2\right)^{1/2}\left(\sum_i \|g^{\text{avg}}_{b,i}\|_2^2\right)^{1/2} \le \bar{w}_b L_\pi.$$

The individual bound follows from $|A^{\text{STAR}}_{b,i}| \le w_{b,i}$. □

Corollary D.6 (Token-mean initial direction)**.** *Fix an aggregation batch $\mathcal{Q}$ of responses with positive valid-token counts L_i, total $T = \sum_{i\in\mathcal{Q}} L_i$, and reward loss reduced by this T. Include skipped responses in T and set their reward advantages to zero. Treat lengths, masks, and advantages as fixed for differentiation. At the old policy, suppose $\|g^{\text{avg}}_i\|_2 \le L_\pi$, where g^{avg}_i averages the score vectors of the same valid tokens. Then*

$$U_{\text{token}} = \frac{1}{T}\sum_{i\in\mathcal{Q}} L_i A^{\text{STAR}}_i g^{\text{avg}}_i, \qquad \|U_{\text{token}}\|_2 \le L_\pi \frac{\sum_{i\in\mathcal{Q}} L_i w_i}{T}.$$

An individual response contributes at most $L_\pi L_i w_i/T$. Exact fitting and split calibration are not required for these deterministic bounds.

Proof. At the old policy the likelihood ratios are one, so the reward-side derivative of each valid token term is A^{STAR}_i times its policy score. Summing within responses gives the displayed direction. The triangle inequality and $|A^{\text{STAR}}_i| \le w_i$ yield

$$\|U_{\text{token}}\|_2 \le \frac{1}{T}\sum_i L_i |A^{\text{STAR}}_i| \, \|g^{\text{avg}}_i\|_2 \le \frac{L_\pi}{T}\sum_i L_i w_i.$$

Skipped responses contribute zero and retain their token mass in the denominator. The individual bound is the same calculation for one summand. □

This corollary concerns the specified aggregation batch, not an assumed global reduction. If microbatch directions are combined with coefficients $a_m \geq 0$, the corresponding bound is $L_\pi \sum_m a_m \sum_{i \in \mathcal{Q}_m} L_i w_i / T_m$. It reduces to the global token-mean expression when $a_m = T_m / \sum_k T_k$; equal microbatch averaging generally uses a different weighting. The result controls the initial reward direction only, not the KL term, later PPO epochs, optimizer state, or parameter displacement.

The theorem concerns the initial reward-side direction at the old policy. STAR guarantees $|A_{b,i}^{\mathrm{STAR}}| \leq 1$, but the standard PPO minimum is not a two-sided hard clipping of the surrogate. For $A < 0$ and an unbounded likelihood ratio ρ, $\min\{\rho A, \mathrm{clip}(\rho) A\} = \rho A$ can be arbitrarily negative. Thus bounded advantages alone do not bound the multi-epoch surrogate or its parameter gradient. Theorem D.5 instead applies at θ_{old}, where $\rho = 1$; later-epoch bounds require an explicit ratio and policy-score condition.

The KL term complements reward-side reliability by constraining policy drift. The realized experiments use the same KL coefficient across methods. If attack outputs already dominate the policy, the primary STAR rule can abstain from unreliable groups, but a zero-sum relative estimator does not by itself create a force that makes every member of an entirely attacked group negative. In fully low-reliability groups, the admission rule converts the reward-side update into explicit abstention, which is the intended conservative behavior of STAR.

E Algorithms, Scope, and Reproducibility

The implementation protocols specify the computational sequence, numerical safeguards, interface metadata, and data separation needed to reproduce STAR.

E.1 Algorithms and Numerical Safeguards

The following protocols specify the primary method, from canonical rendering through discrepancy weighting, admissibility, robust fitting, and the policy loss. Both recorded STAR runs use token-mean reduction; RQ2 additionally uses a cross-evaluator pairing and different solver checks, as documented in Appendix F.

Protocol E.1 (Canonical evaluation)**. Input:** prompt X, policy token sequence O, policy decoder, canonical text normalizer, reward tokenizer, reward model R_ϕ. The procedure is

1. Decode O with the policy tokenizer using the exact generation configuration.
2. Apply the prespecified canonical transformations: Unicode normalization, removal of disallowed invisible characters, canonical whitespace, and the deployed chat template.
3. Retokenize the canonical text with the reward-model tokenizer.
4. Evaluate $R^{\mathrm{canon}} = R_\phi(X, C(O))$.
5. Independently evaluate the deployed representation to obtain $R^{\mathrm{obs}} = R_\phi(X, M_{\mathrm{obs}}(O))$.

Output: $(R^{\mathrm{obs}}, R^{\mathrm{canon}}, D = R^{\mathrm{obs}} - R^{\mathrm{canon}})$.

The canonicalizer specification, tokenizer hashes, special-token policy, maximum length, and truncation direction must be logged. A change in any of these items creates a new calibration context.

Protocol E.2 (Trusted discrepancy calibration)**. Input:** trusted benign outputs, fixed contexts, target level α, scale bounds, $z_{D,c}$, numerical tolerances. The procedure is

1. Randomly split trusted outputs into $\mathcal{I}_{\mathrm{fit}}^D$ and $\mathcal{I}_{\mathrm{cal}}^D$ before fitting.
2. For every sample, compute observed and canonical rewards and assign a context.
3. In every context, solve (16); record boundary hits and the residuals of (17)–(18).
4. Compute calibration scores by (24).

5. Set $k_\alpha = \lceil (n_{\mathrm{cal}} + 1)(1 - \alpha) \rceil$ and choose the conservative order statistic.

6. Freeze $(\widehat{m}_{D,c}, \widehat{s}_{D,c}, \widehat{q}_{1-\alpha})$ for the next policy-training epoch.

Output: context normalizers, conformal threshold, and context fallback map.

Protocol E.3 (Reliability-weight computation)**. Input:** paired rollout rewards, fitted discrepancy calibration, $\lambda, W_{\min}, G_{\min}$. The procedure is

1. Compute $S^D_{b,i}$ by (33) and $w_{b,i}$ by (34).

2. Compute $W_b, \bar{w}_b, p_{b,i}$, and $G_{\mathrm{eff},b}$.

3. If $W_b < W_{\min}$ or $G_{\mathrm{eff},b} < G_{\min}$, set the group's reward advantages to zero and exclude it from the shared-scale fit.

4. Otherwise pass $(w, p, \bar{w}, G_{\mathrm{eff}}, r)$ to the online estimator.

Output: a fixed admissible set and zero-valued reward advantages for skipped groups.

Protocol E.4 (Weighted minibatch fit)**. Input:** fixed admissible groups, z_A, $[v_{\min}, v_{\max}]$, tolerance, maximum iterations. The initialization uses

- $\mu_b^{(0)}$: weighted median or weighted mean of $r_{b,1:G}$;
- $v^{(0)}$: clipped pooled MAD or the previous-step scale.

The solver then alternates

1. Update each μ_b by Newton or safeguarded bisection on (40).

2. Update v by projected Newton or backtracking using (41).

3. Stop only when parameter change and KKT residuals meet the recorded tolerances; otherwise skip the policy step and log solver failure.

Output: $(\widehat{\mu}_{\mathcal{B}_{\mathrm{adm}}}, \widehat{v})$ and solver diagnostics.

Joint convexity makes every exact solution global. The KKT stopping rule controls numerical accuracy, and solver failure triggers a skipped policy step.

Protocol E.5 (STAR-GRPO training step)**. Input:** prompts, frozen old/reference policies, reward model, frozen discrepancy calibration, bounded map h, $G, \kappa, z_A, \lambda, W_{\min}, G_{\min}, v_{\min}, v_{\max}, \varepsilon_w$, solver tolerance/iterations, and clipping/KL parameters. The procedure is

1. Sample G rollouts per prompt and record complete generation metadata.

2. Compute $R^{\mathrm{obs}}, R^{\mathrm{canon}}, D$, and $r = h(R^{\mathrm{canon}} + \kappa D)$.

3. Compute $w, W_b, \bar{w}_b, G_{\mathrm{eff},b}$ and freeze the admissible set.

4. Fit prompt locations and the shared scale on admissible groups only.

5. Compute φ and the primary advantage (43); skipped groups retain zero reward advantage.

6. If every group is skipped or the solver fails, skip the complete optimizer step.

7. Otherwise stop reward-side gradients and optimize (44) with the exact $1/(BG)$ rollout reduction and per-response $1/L_{b,i}$ token average. Keep skipped entries in the loss tensor and do not recenter, whiten, or rescale A^{STAR}.

8. In distributed training, fit one shared scale over the global minibatch rather than one scale per device.

9. Log both rewards, discrepancy, context, reliability, absolute group reliability, solver/KKT status, abstention, $\|g^{\mathrm{avg}}_{b,i}\|$, the theorem-matched initial reward-gradient contribution, and the full PPO+KL+optimizer update separately.

Table 1: Distinct signals and their roles in STAR-GRPO.

Component	Input	Output	Role
Anchored quality	$R^{\mathrm{canon}}, R^{\mathrm{obs}}, \kappa$	$r = h(R_\kappa)$	Prespecified optimization signal
Canonical anchor	Policy output	R^{canon}	Reference representation
Discrepancy calibration	$R^{\mathrm{obs}} - R^{\mathrm{canon}}$	$w \in (0, 1]$	Rollout reliability
Weighted self-tuning	(w, r)	$(\widehat{\mu}, \widehat{v}, \varphi)$	Reliability-first relative score
Absolute group factor	$(\bar{w}, w, \varphi)$	A^{STAR}	Preserve collective distrust

Output: an updated policy or an explicitly recorded skipped step.

Additional counterfactual interfaces $M_1, \ldots, M_K$ may be evaluated offline. Define an audit score such as

$$Q_{\max} = \max_{k \in [K]} \{R_\phi(X, M_k(O)) - R^{\mathrm{canon}}\}.$$

The $K + 1$ rewards of one output are treated as one dependent vector. Conformal calibration applies across examples and does not require independence among views. Audit-only views must not downweight online samples unless they are reachable by the deployed system.

Observed and canonical evaluation requires approximately $2BG$ reward-model forwards per training batch. With K additional audit views, the cost is $(K + 2)BG$. The weighted fitting objective requires $O(BG)$ work per gradient or Newton pass and $O(B + BG)$ memory beyond policy training. Locations parallelize over prompts; the shared scale is one dimensional. Query overhead can be reduced through batching, shared-prefix caching, selective canonical evaluation, or distillation of the discrepancy detector, but all compute-normalized comparisons must report the resulting change in threat coverage.

E.2 Operational Scope and Reproducibility

E.2.1 Guarantee Scope and Design Conditions

STAR separates the optimization signal from its reliability: $r = h(R_\kappa)$ specifies the quality path, $p_{b,i}$ determines each rollout's contribution to the fitted baseline, and $\bar{w}_b$ controls absolute group update strength. Self-tuning adapts the discrepancy and reward scales, while α, λ, z_A, scale bounds, contexts, and admission thresholds remain explicit and reproducible configuration choices.

The guarantees are organized around the information supplied by the paired views:

1. **Paired evidence.** RQ1 pairs deployed and canonical representations of the same rollout; RQ2 pairs a rubric-conditioned proxy with a rubric-free semantic assessment. In both cases, the discrepancy is an observable measure of how strongly the optimized score is supported by an alternative view.
2. **Calibrated reliability.** Under clean exchangeability, split conformal calibration controls marginal clean downweighting at level α. Corollary C.7 gives the corresponding adaptive-round statement under its conditional total-variation premise.
3. **Robust group statistics.** Relative reliability $p_{b,i}$ enters the pseudo-Huber location–scale fit before normalization, reducing the influence of unsupported rewards on the baseline. The shared scale pools information across prompt groups, while the leading $\bar{w}_b$ restores absolute group reliability in the final update.
4. **Explicit abstention.** Groups with insufficient reliable mass receive zero reward advantage and are excluded from the robust fit, converting low support into a conservative optimization decision rather than a noisy group-relative update.
5. **Auditable computation.** Paired scoring requires two reward assessments per rollout, and the robust fitting pass is $O(BG)$ with parallel prompt locations and a one-dimensional shared scale. The resulting reward-side computation is straightforward to batch and log.

6. **Two empirical instantiations.** RQ1 uses split-calibrated token-interface discrepancies; RQ2 uses a fixed cross-evaluator discrepancy reference. Both share the same reliability-first advantage construction and the same separation between quality and learning influence.

These conditions make the theoretical statements and the two experimental instantiations directly auditable: the score pair, quality path, reliability map, admission rule, robust fit, and policy-loss reduction are all explicit parts of the method specification.

E.2.2 Reproducibility Record

A reproducible implementation begins with the exact interface specification. Release or log:

1. policy tokenizer name, version, and vocabulary hash;
2. reward tokenizer name, version, and vocabulary hash;
3. decode options and invalid-byte handling;
4. Unicode normalization form;
5. invisible-character policy;
6. whitespace normalization rules;
7. chat template and role markers;
8. maximum length, truncation direction, padding side, BOS/EOS policy;
9. special-token retention or removal;
10. exact observed-interface mapping.

Contexts are specified before conformal calibration. A default construction crosses reward model, tokenizer pair, task family, and logarithmic response-length bin. A deterministic hierarchy merges contexts with fewer than $n_{\min}$ fitting observations; this hierarchy is frozen before calibration scores are inspected.

The numerical specification includes the quality path h, κ, robustness parameters $z_{D,c}, z_A$, reliability parameters α, λ, scale bounds $v_{\min}, v_{\max}$, admission thresholds $W_{\min}, G_{\min}$, regularizers $\varepsilon_D, \varepsilon_w$, and PPO parameters $\epsilon_{\text{clip}}, \beta_{\text{KL}}$, together with solver tolerance, maximum iterations, and boundary-hit rates. If group size varies, the specification also states whether $W_{\min} = G\tau_W$. The primary objective uses the $1/(BG)$ rollout reduction, per-response length averaging, a global-minibatch shared scale, retained skipped entries, and no downstream advantage recentering or whitening. The archived STAR runs in both RQ1 and RQ2 retain the shared fit and skipped entries but use token-mean loss aggregation. The RQ1 canonical-GRPO control uses per-response length averaging. Appendix F distinguishes these recorded implementations from the reference specification.

The numerical record should make solver behavior auditable. For every batch log:

- objective decrease;
- maximum location-equation residual;
- scale KKT residual;
- number of iterations;
- shared-scale boundary indicator;
- minimum and median effective sample size;
- skipped-group count;
- maximum and RMS advantage;
- maximum $\|g_{b,i}^{\text{avg}}\|$ and theorem-matched initial reward-gradient mass;
- full PPO+KL+optimizer update norm.

Data provenance requires six immutable, disjoint identifier pools: (i) trusted discrepancy fitting, (ii) conformal calibration, (iii) policy training, (iv) attack and hyperparameter validation, (v) independent clean audit, and (vi) final attack plus semantic evaluation. Periodic recalibration consumes fresh trusted fitting and calibration identifiers. Reusing policy, validation, audit, or final examples in calibration invalidates the intended protocol even if no labels are used.

Table 2: RQ1 validation summary at step 855. STAR improves the canonical score while suppressing the runaway deployed-interface reward exploited by TOMPA-GRPO. Scores are raw reward-model outputs; length is in policy tokens.

Method	R_0^{canon}	R_{855}^{canon}	R_{855}^{obs}	Length_{855}
TOMPA-GRPO	–	–	9.641	2048.0
STAR-GRPO	3.121	4.902	-0.570	1872.1

Table 3: Common base settings for the three RQ1 runs. All reward-model evaluations use a 4,096-token input limit; loss aggregation is listed per method in Table 4. The calibration row applies only to STAR.

Component	Setting
Policy and reward model	Llama-3.2-1B-Instruct and Skywork-Reward-V2-Qwen3-8B
Data	10,000 WildChat training prompts; 100 curated NoveltyBench validation prompts
STAR calibration	1,000 disjoint WildChat prompts, used only for discrepancy calibration
Length limits	512 prompt tokens; 2,048 response tokens
Batching and groups	Training batch 64; validation batch 100; eight training and eight validation rollouts per prompt
Sampling	Sampling enabled; temperature 1.0; top-p 1.0; top-k -1
Optimizer	AdamW; learning rate 10^{-6}; weight decay 0.01; gradient clipping 1.0
PPO and KL	PPO mini-batch 64; micro-batch 8 per GPU; one PPO epoch; low-variance KL coefficient 0.001
Reward objective	KL excluded from the scalar reward; entropy coefficient 0
Hardware and precision	Actor: one node with four GPUs and tensor parallelism 2; reward model: one GPU and tensor parallelism 1; bfloat16
Schedule	Seed 42; validation before training and every five updates; configured `total_epochs=10`

F Realized Experimental Configurations

This section records the score pairings, quality paths, calibration choices, and optimization settings used in Sections 6.1 and 6.2. These details distinguish each empirical instantiation from the general protocol and its theoretical assumptions.

F.1 RQ1: Realized Token-Space Configuration

The following settings summarize the realized token-space runs used for Section 6.1.

The quality input to the STAR fit is $r = 10\tanh(R^{\text{canon}}/10)$. The observed/canonical reward curves and discrepancies use untransformed scores. The STAR-specific numerical parameters are summarized in Table 5.

For the observed interface, the policy response identifiers are mapped directly into the reward-model vocabulary with $\Phi(j) = j$; identifiers outside the reward vocabulary are clamped. The prompt is formatted for the reward model, but the response identifiers in this channel are not decoded and retokenized. For the canonical interface, STAR decodes the policy response, applies NFKC normalization, removes Unicode format and other invisible characters, normalizes whitespace, and retokenizes with the reward tokenizer. These operations are fixed before scoring and are applied to the same sampled policy output.

The STAR calibration run uses one sampled response per calibration prompt with seed 42 and reports 1,000 observations across six contexts. They are partitioned using `fit_fraction=0.7` before the conformal threshold is computed, yielding $q = 1.79285$. The original TOMPA logs contain only observed scores; STAR and canonical-GRPO record both reward views.

TOMPA-GRPO contains 1,064 training updates and validation through step 1,060; STAR contains 858 updates and validation through step 855; canonical-GRPO completes 855 updates. The primary TOMPA/STAR comparison in Table 2 uses the common step-855 endpoint. Figure 2 shows the corresponding validation trajectories, and Figure 3 records STAR training diagnostics through step

Table 4: Recorded RQ1 method settings. STAR and the canonical-only diagnostic control share the same bounded canonical quality path and 4,096-token reward-model input limit; STAR additionally uses discrepancy calibration and reliability-first robust advantages.

Parameter	TOMPA-GRPO	STAR-GRPO	Canonical-GRPO
Quality input	R^{obs}	$10\tanh(R^{\text{canon}}/10)$	$10\tanh(R^{\text{canon}}/10)$
Advantages	Group mean/std	Reliability-weighted robust fit	Group mean/std
Loss aggregation	`token-mean`	`token-mean`	`seq-mean-token-mean`
RM token limit	4,096	4,096; right truncation	4,096; right truncation
Online RM calls per rollout	One observed view	Observed and canonical	Observed and canonical; observed logged only
Discrepancy calibration	None	1,000 prompts; six contexts; $q = 1.79285$	None
Group fallback	None	`skip`	None
Logged training horizon	1,064 updates	858 updates	855 updates

Table 5: STAR-specific numerical parameters in the token-space run.

Parameter	Value
z_A, λ, α	2.0, 1.0, 0.05
`fit_fraction`, `min_context_size`, δ	0.7, 50, 0.05
`smin`, `smax`	0.001, 100.0
ε_D, ε_w	10^{-6}, 10^{-6}
`vmin`, `vmax`	0.001, 100.0
$W_{\min}$, $G_{\min}$	1.0, 2.0
`reward_transform`, `reward_bound`	`tanh`, 10.0
`calibration_rollout_n`, `calibration_seed`	1, 42
`calibration_do_sample`	`True`
`solver_tolerance`, `solver_max_iterations`	10^{-7}, 100

858. Over STAR's full recorded training horizon, canonical reward rises from -4.154 to 6.850, observed reward changes from -1.439 to -0.663, and mean discrepancy changes from 2.715 to -7.513. Mean training length rises from 785.3 to 1,843.9 tokens and the length-clip ratio from 0.236 to 0.895. The minimum individual reliability reaches 0.105, showing that the method can apply strong selective attenuation even while average group reliability remains high.

F.1.1 Canonical-GRPO Diagnostic Control

The diagnostic control uses `adv_estimator=grpo` with standard-deviation normalization and applies $10\tanh(R^{\text{canon}}/10)$ in the reward manager before advantage calculation; STAR applies the same bounded canonical map inside its reliability-first estimator. The control therefore isolates a canonical-only quality path from STAR's discrepancy calibration, robust group fit, absolute reliability factor, and admission rule. Both paired-score runs evaluate the observed and canonical views online.

All three RQ1 runs use a 4,096-token reward-model input limit. The paired-score runs record zero observed- or canonical-channel truncation at every validation checkpoint, so the reported validation comparisons are evaluated without reward-input clipping.

The control uses `seq-mean-token-mean`, averaging token losses within each response before averaging responses, whereas STAR and TOMPA-GRPO use `token-mean`. We therefore use this control as a quality-path diagnostic and keep the primary RQ1 trajectory comparison focused on TOMPA-GRPO versus STAR-GRPO. Both canonical runs use learning rate 10^{-6} and the same recorded software stack (PyTorch 2.8.0, vLLM 0.11.0, Transformers 4.57.6).

The control archive contains 855 consecutive training records, 172 validation records at steps $0, 5, \ldots, 855$, and a successful exit record. We extract raw `canonical_reward` and `observed_reward` means, checking them against the final summary and against their `mean@8`

Table 6: Additional RQ1 diagnostic control statistics. The late validation window averages the 11 checkpoints from step 805 to 855; training rows use the common step-855 endpoint.

Metric	STAR-GRPO	Canonical-GRPO
Canonical validation gain, steps 0–855	1.781	3.236
Late-window canonical reward	4.808	6.224
Late-window observed reward	-0.561	-1.865
Late-window response length (tokens)	1881.3	371.6
Training response length, step 855 (tokens)	1896.5	976.3
Training length-clip ratio, step 855	0.922	0.117

Table 7: RQ2 independent evaluation at the latest common eligible checkpoint. The Δ column reports STAR-GRPO minus GRPO, computed from the unrounded means. STAR improves all external-quality and reward-hacking diagnostics while optimizing the same scalar proxy reward.

Metric	GRPO	STAR-GRPO	Δ
Proxy score	0.5640	0.5492	-0.0148
Independent-judge score	0.2706	0.3174	$\mathbf{+0.0469}$
Proxy–judge gap	0.2935	0.2318	$\mathbf{-0.0617}$
Criterion pass rate	0.4752	0.5049	$\mathbf{+0.0297}$
Overclaim fraction	0.2464	0.2204	$\mathbf{-0.0260}$

counterparts. Its generic validation `reward/mean@8` equals 4.370 at step 855 because it averages the transformed rewards; the raw canonical mean is 6.123. STAR's generic reward log is untransformed, so comparing the two generic keys would mix scales. Training reward in the control is also transformed and must not be compared directly with STAR's raw training score.

The control attains a maximum validation canonical mean of 6.637 at step 785, but the main comparison retains the common step 855. At that endpoint its observed score is -1.862, giving $D = -7.985$. Its higher raw canonical reward and shorter outputs show that these behaviors can occur without STAR under the control configuration. They neither identify the causal effect of removing reliability nor establish superior human-perceived quality. Each method has one run, initial validation samples differ, and neither within-prompt sample deviations nor variation across checkpoints estimates uncertainty across training seeds.

F.2 RQ2: Realized Rubric-Proxy Configuration

The following tables give the realized configuration for the rubric-proxy comparison in Section 6.2. In this cross-evaluator instantiation, the proxy, semantic training anchor, and independent evaluation judge have separate roles. The training discrepancy does not measure a tokenizer or serialization change within a single reward model.

For rubric criteria with signed weights a_j and binary judge verdicts q_j, the proxy score is

$$R^{\text{proxy}} = \text{clip}\left(\frac{\sum_j a_j q_j}{\sum_j \max(a_j, 0)},\, 0,\, 1\right).$$

The implementation returns zero when the denominator is zero. Negative pitfall weights contribute to the numerator. The independent validation judge uses the same aggregation with Claude's verdicts. Gemini receives the request and decoded response without the fixed rubric and returns a semantic score in $[0, 1]$. The scalar training reward remains the proxy score; it is already bounded, so the STAR fit applies no additional $\tanh$ transform. This corresponds to $\kappa = 1$ and an identity quality map on the score range.

For a valid training pair, the fixed reference normalization gives

$$D^{\text{rubric}}_{b,i} = R^{\text{proxy}}_{b,i} - R^{\text{anchor}}_{b,i}, \qquad S_{b,i} = \frac{D^{\text{rubric}}_{b,i} - 0}{0.1}, \qquad w_{b,i} = \exp\{-2[S_{b,i} - 1.645]_+\}.$$

Table 8: Shared realized settings for the rubric-reward comparison.

Component	Setting
Policy model	Qwen3-4B
Training data	RubricHub-Medical
Evaluation data	HealthBench-Hard
Hardware	One node with four GPUs; rollout tensor parallelism 2
Rollouts per training prompt	16
Training batch size	64
PPO mini-batch size	64
PPO micro-batch size per GPU	2
Prompt/response limits	2,048 / 1,024 tokens
Optimizer	Adam-style optimizer; learning rate 10^{-6}
PPO schedule	One PPO epoch per update
Policy-loss aggregation	`token-mean`
PPO clipping	Ratio bounds 0.8 and 1.2; dual-clip constant 3.0
KL regularization	Enabled; coefficient 0.001; low-variance KL
Entropy coefficient	0
Validation	Batch size 32; one deterministic rollout; temperature 0; sampling disabled
Validation frequency	Every 20 training steps
Configured horizon	600 training steps; checkpoint resumption disabled

Table 9: Method-specific reward and advantage configuration. Both methods optimize the identical rubric-conditioned scalar proxy reward; STAR additionally uses a rubric-free semantic anchor exclusively to compute rollout reliability.

Parameter	GRPO	STAR-GRPO
Advantage estimator	`grpo`; standard deviation normalization	`star_grpo`; standard normalization disabled
Training reward	Fixed rubric-conditioned proxy score	Same fixed rubric-conditioned proxy score
Proxy judge	`openai/gpt-4o-mini`	`gpt-4o-mini`
Anchor judge	Not used	`gemini-2.5-flash-lite`; training only and rubric-free
Evaluation judge	`anthropic/claude-sonnet-4-6`	`claude-sonnet-4-6`; evaluation only
Evaluation usage	Not used in policy updates	Not used in policy updates
Rubric denominator	`positive`	`positive`
Training judge failure	Proxy failure yields zero scalar reward, retained in GRPO normalization	Invalid proxy/anchor pair receives zero reliability
Validation judge failure	Omit paired metrics if either judge fails	Same paired-metric omission

These constants remain unchanged during training. RQ2 uses the configured `star_conformal_threshold` as a fixed operational cutoff rather than fitting the split-conformal objects used in RQ1. The proxy and anchor share the same numerical range, and their discrepancy is used directly as the semantic support signal that drives STAR reliability.

Reliability enters the prompt-specific pseudo-Huber location fit and the minibatch-shared scale fit through $p_{b,i} = w_{b,i}/W_b$. The implementation retains the leading $\bar{w}_b$ in Equation (9), skips groups with $W_b < 2$ or $G_{\text{eff},b} < 2$, and performs no downstream advantage recentering or whitening. Locations are profiled by 40 bisection iterations for each candidate scale, and the scale is selected by 32 golden-section iterations on $[0.01, 1]$. Locations are recomputed at the selected scale, and finite-value validation of the fitted quantities remains satisfied throughout the recorded run.

Both methods use the trainer's `token-mean` reduction: valid token losses are summed and divided by the token count in the aggregation batch. Skipped STAR groups retain their loss entries with zero reward advantage, and the PPO implementation uses dual clipping with constant 3.0 for negative advantages. Corollary D.6 gives the corresponding length-weighted initial reward-direction control for this realized reduction.

Table 10: STAR-GRPO numerical parameters used for RQ2.

Parameter	Value
star_gap_center	0.0
star_gap_scale	0.1
star_conformal_threshold	1.645
star_reliability_decay	2.0
star_z_a	2.0
star_scale_min, star_scale_max	0.01, 1.0
star_min_weight_sum	2.0
star_min_effective_size	2.0
star_weight_epsilon	10^{-8}
Location / shared-scale search	40 bisection / 32 golden-section iterations
Solver validation	Finite-value checks for fitted locations and shared scale
Judge failure threshold for paired validation	5% maximum proxy/evaluation-judge failure rate

Training API validity is integrated into the same reliability mechanism. In GRPO, the configured fallback maps an invalid proxy call to zero scalar reward. In STAR, an invalid proxy or anchor call additionally receives zero reliability, excluding the unsupported pair from the robust reward fit. If reliable mass is insufficient for every group, STAR abstains from the optimizer step. This makes API validity and semantic disagreement follow one consistent influence rule.

Validation uses GPT-4o-mini and Claude on the same complete HealthBench-Hard rubric for each generated response; Gemini is not called. Proxy, independent-judge, gap, and criterion-level metrics are computed on responses for which both validation judges return valid outputs, using the same prompts, rubric aggregation, and validity rule for both methods. Criterion pass rate and overclaim are unweighted fractions over positive-weight criteria within each response, averaged across valid responses; the implementation falls back to all criteria when none has positive weight.

For the paired checkpoint comparison, both methods must have valid aggregate metrics and a maximum proxy/evaluation-judge failure rate of at most 5%. We use the latest checkpoint meeting this criterion for both methods. Table 7 reports means rounded to four decimals; the reported differences are computed from the unrounded means. The training proxy–anchor gap in Figure 5 is distinct from the validation proxy–judge gap.

The convexity result applies to fixed reward and weight arrays regardless of whether the paired scores come from one evaluator or two. The coordinate and second-moment bounds require only finite locations, a positive scale, and the stated advantage formula; exact zero sum additionally uses the location equation. RQ1 instantiates the split-calibrated reliability theorem, while RQ2 demonstrates the same advantage construction with a fixed cross-evaluator reference.

Evaluation protocol summary. The paired checkpoint comparison uses the latest checkpoint for which both methods have valid aggregate metrics and each validation judge satisfies the fixed 5% availability threshold. Table 7 reports means rounded to four decimals, with differences computed from the unrounded values. The training proxy–anchor gap in Figure 5 is distinct from the validation proxy–judge gap: the former drives STAR reliability during optimization, while the latter is an external diagnostic computed with Claude.

F.3 Summary of Assumptions and Guarantees

Table 11: Core STAR-GRPO design conditions and the guarantees they enable.

Design condition	Resulting guarantee	Role in STAR
Paired reward views	Observable discrepancy D	Measures support for the optimized score
Fixed h, κ	Auditable quality path $r = h(R_\kappa)$	Separates quality from reliability
Trusted disjoint fitting data	External discrepancy reference	Prevents current policy groups from defining their own baseline
Clean calibration exchangeability	Marginal clean downweighting $\leq \alpha$	Calibrates when attenuation begins
Self-tuned robust fit	Context location and scale control	Adapts to heterogeneous discrepancy scales
Attack-score separation	Exponential expected-weight attenuation	Converts discrepancy separation into influence reduction
$W_b, G_{\mathrm{eff,b}}$ admission	Well-defined fit or zero-advantage abstention	Protects low-support groups
Weighted location equation	Exact advantage zero sum	Preserves group-relative centering
Bounded score and $\bar{w}_b$ factor	Coordinate and second-moment bounds	Preserves absolute group reliability
Bounded policy score; specified reduction	Initial reward-direction bound	Links reliability to policy influence
View-dependent reward inflation	Discrepancy carries attack information	Target reward-hacking regime in RQ1/RQ2