APEX: ACCURACY PROJECTION FROM VERIFIER-LABELED EXPERIENCE FOR OFFLINE RLVR

Anonymous authors
Paper under double-blind review

ABSTRACT

Reinforcement learning with verifiable rewards (RLVR) is often presented as online KL-regularized policy improvement, while practical offline pipelines cache responses and reuse verifier labels. After logging, the learner cannot explore beyond reference support, so those labels can only justify a reference-supported policy. Reward-weighted and rejection-sampling SFT convert verifier labels into imitation weights or filters, but they do not isolate the policy target induced by binary-verifier RLVR. We identify that target as the accuracy-optimal reference projection, which is the zero-temperature limit of KL-regularized RLVR and, equivalently, the minimum-KL reference-supported policy with unit verifier accuracy. This projection formulates offline RLVR as a Bernoulli density-ratio problem in which correct responses support the projected policy and incorrect responses identify reference mass to remove. APEX (Accuracy Projection from Verifier-Labeled Experience) fits the KL reward-policy coordinate on logged reference rollouts with a binary correctness likelihood. Our analysis establishes population policy consistency, an exact accuracy-gap identity, a finite-sample bound scaling as $1/[Nq(1-q)]$ for group size N and correct-response fraction q, and a bias-variance explanation for weighted and rejection-sampling SFT baselines. Ratio-stratified GSM8K studies and BigMath fine-tuning support this analysis, with the largest APEX gains occurring in rollout groups that contain both correct-response support and incorrect-response labels. Beyond mathematical reasoning, APEX also outperforms offline baselines and online GRPO on Search-R1 multi-turn search and question answering.

1 INTRODUCTION

Reinforcement learning with verifiable rewards (RLVR) is a central post-training tool for reasoning models because many mathematical answers can be checked automatically. In this setting, a verifier converts each sampled solution into a correctness label, avoiding the ambiguity of learned preference rewards for final-answer tasks. Datasets such as GSM8K (Cobbe et al., 2021) and BigMath (Albalak et al., 2025) support this setting, and recent reasoning systems use correctness rewards to improve solution quality (Shao et al., 2024; Guo et al., 2025; Lambert et al., 2025). The standard population formulation is KL-regularized policy optimization around a reference model (Ziegler et al., 2019; Stiennon et al., 2020; Ouyang et al., 2022; Rafailov et al., 2023). The verifier provides an unambiguous reward signal, but the training loop remains computationally expensive. Online PPO-style or GRPO-style RLVR follows this formulation by repeatedly sampling from the current policy, verifying responses, and updating the model (Schulman et al., 2017; Shao et al., 2024; Guo et al., 2025). This procedure is effective, but it keeps generation, verification, and optimization coupled throughout training.

Offline RLVR decouples these stages by generating verifier-labeled responses once and reusing the logged rollouts for later optimization. This reuse is attractive when generation or verification is expensive, and it has led to reward-weighted supervised fine-tuning, filtering, ranking, and pairwise objectives on fixed response sets (Rafailov et al., 2023; Shu et al., 2026; Du et al., 2026; Mukherjee et al., 2025; Dong et al., 2023; Yuan et al., 2023; Baheti et al., 2024). We provide a discussion of related work on offline alignment and logged policy learning in Appx. A. This reuse also changes the recoverable target. The logging policy has already determined which completions enter the dataset, so

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.

an offline learner cannot rely on arbitrary on-policy exploration. The recoverable policy is therefore the reference-supported policy identified by verifier labels under this support and coverage constraint.

Binary verifiable rewards make that policy explicit. For the KL-regularized RLVR objective, every finite temperature gives a Boltzmann reweighting of the reference policy. As the temperature goes to zero, incorrect responses receive vanishing probability, and the limit becomes the reference policy conditioned on correct completions. This accuracy-optimal reference projection is also the minimum-KL reference-supported policy with unit verifier accuracy. It preserves the reference model's relative preferences among correct completions while removing incorrect-response mass. This gives offline RLVR a population target that is distinct from a generic weighted-imitation proxy (Sec. 2).

APEX learns this projection through its density ratio under reference sampling. A correct response supports the numerator of the ratio, while an incorrect response identifies reference-relative mass to reduce. To use both signals, APEX writes the current policy in the reference-relative KL coordinate used by DPO and BOLT and fits the policy-induced score with a Bernoulli likelihood for verifier labels $r \in \{-1, +1\}$. At the population level, the APEX policy risk is minimized exactly by the zero-temperature projection. Thus β scales the reward-policy coordinate without defining a finite-temperature weighted-SFT target. The same density-ratio view also explains why finite rollout composition matters. Mixed groups contain both support for the target and evidence about which reference mass the projection should remove. All-correct groups supply correct completions but little evidence about incorrect mass, while all-incorrect groups expose a support boundary because the observed batch has no correct completion for the projection. The mixed-correctness regime therefore provides the strongest signal for accuracy projection from verifier-labeled experience. The algorithm is given in Sec. 3, and Sec. 4 quantifies the difference between these regimes through the accuracy gap.

This perspective yields three contributions. ***First***, we identify the accuracy-optimal reference projection as a principled recoverable target for offline RLVR under the KL-regularized binary-verifier objective. APEX is policy-consistent for that projection, so each verifier-labeled response provides density-ratio evidence for preserving correct-response support and removing incorrect reference mass. Its role therefore extends beyond that of a weighted demonstration (Secs. 2 and 3). ***Second***, we analyze learning from finite logged data through the accuracy gap, moving beyond heuristic filtering rules. The exact gap identity and realizable mixed-correctness bound show that recovery improves with $Nq(1-q)$, where N is group size and q is the correct-response fraction, making the correct-incorrect ratio a statistical efficiency variable (Sec. 4). ***Third***, we explain weighted and rejection-sampling SFT baselines within the same logged-rollout setting. Finite weights leave target bias, while aggressive filtering increases weight variance when correct responses are rare. Ratio-stratified GSM8K experiments isolate the $q(1-q)$ mechanism, and BigMath fine-tuning shows that the same accuracy-projection learner outperforms offline baselines while remaining competitive with online GRPO on larger math-reasoning evaluations. Search-R1 experiments (Jin et al., 2025) extend the evaluation beyond mathematical reasoning to multi-turn search and question answering, where APEX outperforms both offline baselines and online GRPO on all seven QA benchmarks (Sec. 5).

2 From KL-Regularized RLVR to Accuracy Projection

Logged reference rollouts make support and coverage the first constraints in offline RLVR. A fixed batch can only reweight outputs generated by the logging policy, so the population target must lie within reference support before a sample-level loss can be justified. Within that support, the standard KL-regularized RLVR objective supplies the policy path, and binary-verifier evaluation selects its zero-temperature limit because pass@1 and accuracy count only correct responses. Finite temperature then measures the residual accuracy gap, while finite logged groups determine whether the projection has observable correct-response support and incorrect-response evidence. These objects define the density-ratio target estimated by APEX.

2.1 KL-Regularized RLVR Under Reference Support

The standard KL-regularized RLVR objective defines the reference-supported policy path. For logged rollouts, it rewards verified correctness while penalizing movement away from $\pi_{\text{ref}}(\cdot \mid x)$. With prompts $x \sim \mathcal{D}$, binary verifier reward $r(x, y) \in \{-1, +1\}$, and temperature $\beta > 0$, write the

objective as

$$\mathcal{J}_\beta(\pi) \triangleq \mathbb{E}_{x\sim\mathcal{D}}\left[\mathbb{E}_{y\sim\pi(\cdot|x)}[r(x,y)] - \beta D_{\mathrm{KL}}(\pi(\cdot \mid x)\|\pi_{\text{ref}}(\cdot \mid x))\right]. \tag{1}$$

The KL penalty restricts the policy path to reference support because mass outside that support has infinite cost. Logged-rollout training also has no samples there. Because Equation (1) separates over prompts, each prompt has the familiar Boltzmann reference tilt (proof in Appx. B.1).

Lemma 2.1 (Finite-Temperature KL-RLVR Path). *For any fixed prompt x and $\beta > 0$, the unique optimizer of $\max_\pi \mathcal{J}_\beta(\pi)$ on the reference support is*

$$Z_\beta(x) \triangleq \mathbb{E}_{y'\sim\pi_{\text{ref}}(\cdot|x)}[\exp(r(x,y')/\beta)], \qquad \pi_\beta{}^\star(y \mid x) = \frac{\pi_{\text{ref}}(y \mid x)\exp(r(x,y)/\beta)}{Z_\beta(x)}. \tag{2}$$

2.2 Accuracy-Optimal Reference Projection

The finite-temperature tilt defines the KL path, but binary-verifier evaluation depends only on the probability assigned to correct responses. Metrics such as pass@1 or accuracy estimate prompt-level verifier accuracy. For stochastic labels, the posterior $\eta(x,y) \triangleq \Pr(r = 1 \mid x, y)$ gives the same scale through $\mathbb{E}[r \mid x, y] = 2\eta(x,y) - 1$. In the deterministic binary setting, $\eta(x,y) = \mathbf{1}\{r(x,y) = 1\}$. For a candidate policy π, define the correctness sets and verifier accuracies as

$$\begin{aligned} &\mathcal{Y}_+(x) \triangleq \{y : r(x,y) = 1\}, \qquad \mathcal{Y}_-(x) \triangleq \{y : r(x,y) = -1\}, \\ &\pi^+(x) \triangleq \pi(\mathcal{Y}_+(x) \mid x), \qquad \pi_{\text{ref}}{}^+(x) \triangleq \pi_{\text{ref}}(\mathcal{Y}_+(x) \mid x). \end{aligned} \tag{3}$$

With binary rewards, $\mathbb{E}_{y\sim\pi(\cdot|x)}[r(x,y)] = 2\pi^+(x) - 1$, so the reward term in Equation (1) is verifier accuracy up to an affine scale. The KL term controls how much reference distortion is tolerated along this accuracy path. At finite β, every reference-supported incorrect response still receives positive probability under $\pi_\beta{}^\star$, contributing directly to the accuracy gap. When verifier accuracy dominates the KL tradeoff, the reference-supported limit removes this gap while staying on the same KL path. For $\pi_{\text{ref}}{}^+(x) > 0$, the limit is the reference policy conditioned on correct responses, which we denote by $\pi^\star$ (proof in Appx. B.1).

Theorem 2.2 (Zero-Temperature Accuracy Projection). *For $\pi_{\text{ref}}{}^+(x) > 0$, the finite-temperature path satisfies*

$$\lim_{\beta\downarrow 0} \pi_\beta{}^\star(y \mid x) = \pi^\star(y \mid x) \triangleq \frac{\pi_{\text{ref}}(y \mid x)}{\pi_{\text{ref}}{}^+(x)}\mathbf{1}\{r(x,y) = 1\}. \tag{4}$$

The zero-temperature limit removes incorrect responses while preserving reference preferences within $\mathcal{Y}_+(x)$. For any two correct completions, $\pi^\star$ preserves their probability ratio under $\pi_{\text{ref}}(\cdot \mid x)$. Thus the accuracy-optimal reference projection is a KL projection onto unit verifier accuracy. Equivalently, it minimizes KL divergence from the reference among reference-supported policies with unit verifier accuracy (proof in Appx. B.1), so

$$\pi^\star(\cdot \mid x) = \arg \min_{\pi \ll \pi_{\text{ref}},\ \pi(\mathcal{Y}_+(x)|x)=1} D_{\mathrm{KL}}(\pi(\cdot \mid x)\|\pi_{\text{ref}}(\cdot \mid x)). \tag{5}$$

2.3 Finite Temperature Leaves an Accuracy Gap

The accuracy-optimal reference projection is the binary-verifier benchmark. For any policy π, $1 - \pi^+(x)$ is the prompt-level accuracy gap under Equation (3). Because $(\pi^\star)^+(x) = 1$, the projection has zero gap. At finite temperature, the residual gap is the incorrect-response mass under $\pi_\beta{}^\star$. The finite-temperature target solves the KL-regularized objective exactly, but it can still have a large gap when correct responses are rare under the reference policy.

Proposition 2.3 (Finite-Temperature Accuracy Gap). *For any fixed prompt x with $\pi_{\text{ref}}{}^+(x) \in (0,1)$ and any $\beta > 0$, let $\pi_\beta{}^\star$ be the finite-temperature optimizer in Lem. 2.1. Define $(\pi_\beta{}^\star)^+(x) \triangleq \pi_\beta{}^\star(\mathcal{Y}_+(x) \mid x)$ and $(\pi^\star)^+(x) \triangleq \pi^\star(\mathcal{Y}_+(x) \mid x)$. Then the finite-temperature optimizer has a strictly larger accuracy gap than the accuracy-optimal reference projection:*

$$1 - (\pi_\beta{}^\star)^+(x) = \frac{1 - \pi_{\text{ref}}{}^+(x)}{\pi_{\text{ref}}{}^+(x)e^{2/\beta} + 1 - \pi_{\text{ref}}{}^+(x)} > 0 = 1 - (\pi^\star)^+(x). \tag{6}$$

The residual exposes the limitation of finite temperature. It vanishes as $\beta \downarrow 0$, but remains positive at every finite β whenever both correct and incorrect responses have reference support. Thus the finite-temperature KL optimizer can approximate the projection, but it cannot match its zero accuracy gap. The same denominator compares boosted correct-response reference mass with unboosted incorrect-response mass, giving the boost needed for any target accuracy.

Corollary 2.4 (Temperature Required for a Target Accuracy)**.** *Let $(\pi_{\beta^\star})^+(x)$ be the prompt-level verifier accuracy of the finite-temperature optimizer in Prop. 2.3. For any $\epsilon \in (0,1)$, $1-(\pi_{\beta^\star})^+(x) \le \epsilon$ holds exactly when*

$$\pi_{\text{ref}}{}^+(x)e^{2/\beta} \ge \frac{1-\epsilon}{\epsilon}\big(1-\pi_{\text{ref}}{}^+(x)\big). \tag{7}$$

The temperature requirement in Cor. 2.4 depends on the correct-incorrect mass ratio, so a temperature choice alone does not guarantee verifier accuracy. If $\pi_{\text{ref}}{}^+(x)e^{2/\beta} \le 1-\pi_{\text{ref}}{}^+(x)$, the exponential target still assigns at least half its mass to incorrect responses, so its binary-verifier value remains poor despite exact finite-temperature optimality. Cor. 2.4 equivalently requires $e^{2/\beta} \ge (1-\pi_{\text{ref}}{}^+(x))(1-\epsilon)/[\pi_{\text{ref}}{}^+(x)\epsilon]$, so driving the gap below ϵ for small $\pi_{\text{ref}}{}^+(x)$ requires an exponential correct-incorrect ratio of order $1/[\pi_{\text{ref}}{}^+(x)\epsilon]$. A larger ratio reduces target bias in $1-\pi^+$, but concentrates empirical fitting on rare correct responses. This is the finite-temperature weighted-SFT tradeoff analyzed in Prop. B.2. Correct-only rejection sampling is the infinite-ratio limit. It removes the population accuracy gap, but applying it to finite logged rollouts discards the incorrect-response labels that identify which logged mass to remove. The binary-verifier metric therefore changes the learning problem by specifying the zero-temperature projection as the target. Recovering that policy from logged data depends on the available correctness evidence.

2.4 Support and Evidence in Logged Rollouts

The accuracy-projection target fixes the population policy, while logged rollouts impose the support and coverage constraints of offline training. The projection only reweights correct completions within reference support. It preserves relative mass inside $\mathcal{Y}_+(x)$ and creates no mass outside the logging policy's support. If $\pi_{\text{ref}}{}^+(x) = 0$, no reference-supported policy can realize $\pi^\star$. A finite logged group imposes an additional sampling constraint because it may contain no correct response even when $\pi_{\text{ref}}{}^+(x) > 0$. Such an all-incorrect group reflects a limitation of offline support and does not establish that no correct completion exists. Its labels identify incorrect responses to suppress, but provide no correct completion whose reference-relative mass can estimate the projection numerator. This boundary is reflected in the accuracy-projection density ratio. Its numerator is estimable only from observed correct responses, while zeros on incorrect responses directly lower reference-relative mass.

3 APEX: Accuracy Projection from Verifier-Labeled Experience

With logged support fixed, a verifier label trains the policy in the KL reward-policy coordinate used by DPO and BOLT (Rafailov et al., 2023; Shu et al., 2026). For the finite-temperature optimizer in Lem. 2.1, rearranging the Boltzmann form writes the reward, up to a prompt-only shift, as $\beta\log(\pi_{\beta^\star}(y \mid x)/\pi_{\text{ref}}(y \mid x)) + \beta\log Z_\beta(x)$. This coordinate treats the policy/reference log ratio as a policy-implied reward estimate. APEX retains this coordinate but changes the observation model because a binary verifier reveals correctness and does not provide a calibrated finite-temperature reward value.

The accuracy-optimal reference projection determines the roles of correct and incorrect labels in density-ratio estimation. Under reference sampling, target matching is a density-ratio problem (Shu et al., 2026). For the projection, the ratio is

$$\rho^\star_{\text{acc}}(y \mid x) \triangleq \frac{\pi^\star(y \mid x)}{\pi_{\text{ref}}(y \mid x)} = \frac{\mathbf{1}\{r(x,y)=1\}}{\pi_{\text{ref}}{}^+(x)}. \tag{8}$$

This ratio turns each verifier label into a sample-level density-ratio signal. A correct response estimates the numerator and keeps reference-relative mass, scaled only by the prompt-level reference-accuracy normalizer. An incorrect response has zero target density, so it directly decreases reference-relative mass and is not used as a low-weight demonstration.

To make this signal trainable, APEX uses the current policy as an implicit reward estimator in the coordinate used by DPO and BOLT. The sampled normalizer $\hat{Z}(x)$ is the logged-rollout analogue of $Z_\beta(x)$ in the optimal-policy identity, and substituting π_θ for the unknown optimizer gives

$$f_\theta(x, y) = \beta \log \frac{\pi_\theta(y \mid x)}{\pi_{\text{ref}}(y \mid x)} + \beta \log \hat{Z}(x). \tag{9}$$

Here $\hat{Z}(x) = N^{-1} \sum_{i=1}^{N} \exp(r_i/\beta)$ is computed once from the logged reference group. Because $\hat{Z}(x)$ is prompt-only, it estimates the KL-coordinate intercept while leaving normalized within-prompt policy ratios unchanged. The response-dependent part of f_θ is the current policy's reference-relative log density ratio, scaled by β. Thus f_θ is the reward-scale estimate implied by the KL optimal-policy identity, not an arbitrary classifier score.

The projection ratio then determines how to fit this reward estimate. A verifier label $r \in \{-1, +1\}$ supplies the correct-versus-incorrect class, and we use the policy-implied reward as the Bernoulli logit in

$$\ell_{\text{APEX}}(r, f) = \log(1 + \exp[-rf]) \,. \tag{10}$$

This loss is the negative log-likelihood for classifying verifier-correct responses in the KL reward coordinate. Its Bayes logit is the posterior log-odds, and proper composite losses such as logistic loss are a standard approach to density-ratio estimation by classification (Gutmann & Hyvärinen, 2010; Reid & Williamson, 2010; Nguyen et al., 2010; Menon & Ong, 2016). Correct responses raise the implied reward and reference-relative density, while incorrect responses lower them. The Bernoulli likelihood keeps each sample update bounded.

Alg. 1 carries out this density-ratio update as a two-stage offline procedure. The generation stage fixes the evidence set by sampling from $\pi_{\text{ref}}(\cdot \mid x)$, computing verifier rewards, estimating $\hat{Z}(x)$, and storing $(x, y_i, r_i, \hat{Z}(x))$. The optimization stage evaluates Equation (9), fits Equation (10), and updates only π_θ. Rollouts fix the support, and learning changes only the reference-relative density assigned to cached responses.

Algorithm 1 APEX: Accuracy Projection from Verifier-Labeled Experience

Require: Prompts $\mathcal{X}_{\text{train}}$, reference policy π_{ref}, verifier $r(x, y) \in \{-1, +1\}$, rollout size N, minibatch size B, steps T, temperature β, step size α_{step}.
Ensure: Optimized policy π_θ.

Stage I: Reference rollout construction

Initialize rollout dataset $\mathcal{R} \leftarrow \emptyset$.
for $x \in \mathcal{X}_{\text{train}}$ **do**
  Sample $\{y_i\}_{i=1}^{N} \sim \pi_{\text{ref}}(\cdot \mid x)$.
  Evaluate $r_i \leftarrow r(x, y_i)$ for $i = 1, \ldots, N$.
  Estimate $\hat{Z}(x) \leftarrow N^{-1} \sum_{i=1}^{N} \exp(r_i/\beta)$.
  Store $\mathcal{R} \leftarrow \mathcal{R} \cup \{(x, y_i, r_i, \hat{Z}(x))\}_{i=1}^{N}$.
end for

Stage II: Accuracy-projection optimization

Initialize $\pi_\theta \leftarrow \pi_{\text{ref}}$.
for $t = 1, \ldots, T$ **do**
  Sample $\mathcal{B}_t = \{(x_j, y_j, r_j, \hat{Z}(x_j))\}_{j=1}^{B} \subset \mathcal{R}$.
  Compute $f_j \leftarrow f_\theta(x_j, y_j)$ using Equation (9).
  Set $\widehat{\mathcal{L}}_{\text{APEX}} \leftarrow B^{-1} \sum_{j=1}^{B} \ell_{\text{APEX}}(r_j, f_j)$.
  Take one optimizer step with $\nabla_\theta \widehat{\mathcal{L}}_{\text{APEX}}$ and α_{step}.
end for
return π_θ.

4 Accuracy-Gap Bounds and Ratio Efficiency

The accuracy-projection density ratio fixes the target, while Alg. 1 estimates it through the binary correctness likelihood. A classifier loss alone establishes label predictability, but the policy claim requires more. After the learned score is normalized into a policy, the relevant guarantee is that the induced policy matches the accuracy-optimal reference projection, not a finite-temperature imitation target. Policy consistency establishes this population link and leaves finite logged data as the remaining error source.

Theorem 4.1 (APEX Policy Consistency). *For any prompt x with $\pi_{\text{ref}}{}^{+}(x) > 0$, suppose θ is a population minimizer of the prompt-conditioned APEX loss obtained by averaging Equation* (10) *over reference rollouts. Then*

$$\pi_\theta(\cdot \mid x) = \pi^\star(\cdot \mid x). \tag{11}$$

Policy consistency gives the classifier-based update its policy meaning (proof in Appx. B.2). The learned score matters through the policy it normalizes into, and that policy is the KL path's zero-temperature endpoint. After prompt-only normalizers cancel in π_θ, the population minimizer removes density from $\mathcal{Y}_-(x)$ and preserves the reference distribution inside $\mathcal{Y}_+(x)$. This separates two roles that weighted-SFT objectives tend to conflate. The parameter β scales the reward-policy coordinate and bounds updates, while the normalized target remains the accuracy-optimal projection. A remaining performance gap therefore reflects finite-data error in recovering the projection from logged correctness evidence and does not indicate a different population target. The accuracy-gap identity measures that error on the same verifier-accuracy scale used in evaluation.

Theorem 4.2 (Accuracy-Gap Identity). *For any prompt x with ${\pi_{\text{ref}}}^+(x) > 0$ and any APEX policy $\pi_\theta \ll \pi_{\text{ref}}$, the accuracy gap under the binary verifier is exactly the reference-weighted density ratio that remains on incorrect responses:*

$$1 - {\pi_\theta}^+(x) = \mathbb{E}_{y\sim\pi_{\text{ref}}(\cdot|x)}\left[\frac{\pi_\theta(y \mid x)}{\pi_{\text{ref}}(y \mid x)}\mathbf{1}\{r(x,y) = -1\}\right]. \tag{12}$$

The identity turns the accuracy gap into measurable policy mass (proof in Appx. B.3). Since $(\pi^\star)^+(x) = 1$, every remaining error is reference-relative density left on $\mathcal{Y}_-(x)$. Incorrect rollouts therefore identify the exact mass that prevents the learned policy from matching the projection, giving them a role beyond that of discarded demonstrations. This identity places pass@1-style evaluation on the same scale as the density-ratio learner. Improving verifier accuracy requires reducing reference-relative density on incorrect completions, so increasing likelihood on known correct ones is not the only consideration. All-correct logged groups can imitate numerator support but cannot reveal which incorrect mass remains, while all-incorrect groups reveal mass to suppress but provide no correct anchor for the projected policy. Finite-sample recovery therefore depends on the amount of contrast in the logged group.

Theorem 4.3 (APEX Mixed-Correctness Accuracy-Gap Bound). *For any retained prompt group of size N with $k(x)$ correct responses and correct-response fraction $q = k(x)/N \in (0,1)$, assume the score class contains an accuracy-projection score whose induced policy is $\pi^\star$. Equivalently, this score represents the accuracy-projection density ratio up to a prompt-only shift. Assume also the standard local fast-rate and calibrated-surrogate conditions in Appx. B.3. If $\hat{\theta}$ exactly minimizes the empirical APEX loss over a bounded accuracy-projection score class with effective dimension d_{eff}, then, with probability at least $1-\delta$,*

$$1 - {\pi_{\hat{\theta}}}^+(x) \lesssim \frac{d_{\text{eff}} + \log(1/\delta)}{Nq(1-q)}. \tag{13}$$

The bound describes finite-data recovery in terms of the same accuracy gap. Under realizability and exact empirical optimization, there is no misspecification or optimization residual. The main statistical limitation is therefore explicit, with the remaining gap to $\pi^\star$ bounded at order $1/[Nq(1-q)]$. The term has the form of a case-control efficiency factor because accuracy projection needs both sides of the binary verifier label. Correct responses estimate the numerator of the projected density ratio, while incorrect responses estimate the denominator-side mass removed from the learned policy. If q is near one, the batch contains many examples to imitate but little contrast for suppressing incorrect completions. If q is near zero, the batch contains many failures but too little correct-response support to anchor the projected policy. This bound identifies mixed-correctness groups as the most informative regime and distinguishes the failure modes of the two boundary groups in ratio-stratified experiments. The controlled study in Appx. C.1 and Figure 2 shows smaller accuracy gaps near balanced allocations $q = 0.5$ and decreasing gaps as the sample size N increases, consistent with the statistical efficiency mechanism captured by $Nq(1-q)$.

Corollary 4.4 (Correctness-Ratio Efficiency). *Let q be the retained correct-response fraction in Thm. 4.3. For fixed N, d_{eff}, and δ, a balanced mixed-correctness group minimizes the upper bound in Thm. 4.3, giving*

$$\arg\min_{q\in(0,1)} \frac{d_{\text{eff}} + \log(1/\delta)}{Nq(1-q)} = \frac{1}{2}. \tag{14}$$

Minimizing the leading term converts the bound into the ratio rule (proof in Appx. B.3). The factor $q(1-q)$ measures the correctness contrast present in a retained rollout group. Either boundary is inefficient because $q \approx 1$ loses incorrect-response labels, while $q \approx 0$ loses correct-response support.

The leading term is minimized by a balanced mixed-correctness group, providing a direct empirical test of the theory. Figure 2 in Appx. C.1 provides a controlled empirical view of this ratio rule. Holding the sample size N fixed, the measured incorrect-response mass follows a U-shaped profile across q, with smaller gaps under balanced allocations and an empirical minimum near $q = 1/2$. In finite training, performance need not peak exactly at $q = 1/2$. When correct completions are scarce or model capacity is limited, preserving more correct-response support can shift the empirical optimum toward correct-rich groups. The more general implication is that changing the retained correct-incorrect ratio should affect APEX through the accuracy-projection contrast term, while weighted or rejection-sampling SFT follows its own bias-variance curve. Finite correctness weights leave accuracy-gap bias, while correct-only filtering removes that bias at the cost of $1/q$-type variance (formal statement in Prop. B.2). The empirical evaluation examines these two implications separately. Ratio-stratified GSM8K splits isolate the $q(1-q)$ mechanism, and BigMath fine-tuning tests whether respecting the logged-support boundary improves larger offline RLVR runs.

5 Experiments: Correctness Labels in Logged Rollouts

The accuracy-gap bound yields a testable prediction for offline RLVR with verifier-labeled rollouts. If APEX learns the accuracy-optimal reference projection, its advantage depends on two kinds of evidence in each logged group. Correct responses support the projection numerator, and incorrect responses identify reference mass to remove. Thm. 4.3 therefore predicts the clearest separation in mixed-correctness groups, and Cor. 4.4 identifies $q = 1/2$ as the optimum of the leading bound. Weighted and rejection-sampling SFT baselines follow a different curve because they continue to treat retained responses as demonstrations, incurring either finite-temperature target bias or high variance on rare correct samples (formalized in Prop. B.2). The main experiments consist of a ratio-stratified GSM8K study, a BigMath fine-tuning study, and a Search-R1 agentic study. The GSM8K study fixes the backbone, verifier, group size, and correctness composition (Sec. 5.1). The BigMath study uses the natural rollout composition and evaluates whether the same logged-support rule improves larger offline RLVR runs compared with offline weighted-SFT baselines and online GRPO (Sec. 5.2). The Search-R1 study extends the evaluation beyond mathematical reasoning to multi-turn search and question answering (Sec. 5.3).

5.1 Ratio-Stratified GSM8K Study

The GSM8K study varies the correct-response fraction q in Equation (13). The study uses `Qwen2.5-Math-1.5B` on GSM8K (Cobbe et al., 2021), which contains 7473 training problems, and samples $N = 8$ responses per prompt from the fixed reference policy. Prompt-group rejection sampling retains only groups with a prescribed empirical incorrect-response fraction, thereby changing q while keeping the backbone, verifier, group size, and rollout policy fixed. The comparison includes APEX, BOLT (Shu et al., 2026), VAR (Du et al., 2026), ReFit (Mukherjee et al., 2025), and PMPO (Abdolmaleki et al., 2025), all trained on the same reference-policy rollouts. This ratio stratification examines the mixed-correctness regime highlighted by the finite-sample bound, while all-correct and all-incorrect boundary groups reveal which part of the logged evidence is missing. More setup details are in Appx. D.

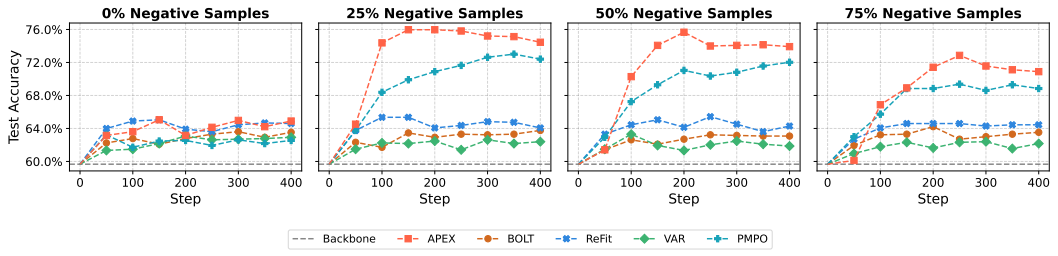


Figure 1: Ratio-stratified GSM8K rollouts show the effect of mixed correctness. Each panel reports test accuracy during training for retained prompt groups with the shown incorrect-response fraction, so q is the complementary correct-response fraction. All methods use the same reference-policy rollouts, and the dashed horizontal line is the `Qwen2.5-Math-1.5B` backbone.

The curves in Figure 1 show the largest APEX gains in the mixed-correctness regime. With 25%, 50%, and 75% incorrect responses, APEX achieves the highest accuracies at approximately 76.0%, 75.7%, and 72.9%, exceeding BOLT, VAR, ReFit, and PMPO in each split. These gains match the finite-sample scaling in Thm. 4.3, with retained groups being most useful when correct-response support and incorrect-response labels are both present. The best stratified result occurs at the correct-rich 25%-incorrect split, not the idealized 50% split. This suggests that correct-response support is the more limited resource in this finite-data experiment. The boundary groups further support this interpretation. With only correct responses, all methods improve over the backbone and the APEX advantage is modest because no incorrect labels identify reference mass to remove. With only incorrect responses, APEX and PMPO collapse sharply, and BOLT, ReFit, and VAR also fail to improve meaningfully. This collapse reflects an offline support boundary and does not imply that the prompts are intrinsically unsolvable. The logged group contains no correct-response anchor for $\pi^\star$, and Appx. C.2 examines this failure mode directly. The GSM8K study therefore supports both parts of the claim about correctness evidence. Mixed groups show the $Nq(1-q)$ efficiency advantage, while boundary groups show the limits of recovering the projection from logged rollouts.

5.2 BigMath Offline Fine-Tuning

Following the ratio-stratified study, the BigMath experiment tests whether the same evidence rule applies to naturally occurring rollout compositions without prescribing the ratio. The large-scale study uses `Qwen2.5-Math-7B` and `Qwen3-8B` as base models. For 50k BigMath training problems (Albalak et al., 2025), the fixed reference policy samples $N = 8$ responses per problem. Because all-incorrect groups contain no logged correct-response support for the projection numerator, training removes them and fine-tunes on the remaining data. Within each backbone, all offline methods use the same retained rollouts and verifier labels, so the comparison isolates the optimization objective from the data-retention rule. As in Sec. 5.1, the offline baselines are BOLT, ReFit (Mukherjee et al., 2025), VAR (Du et al., 2026), and PMPO (Abdolmaleki et al., 2025). Group-Relative Policy Optimization (GRPO (Guo et al., 2025)) serves as the online baseline. Evaluation uses six math reasoning benchmarks, including MATH-500 (Hendrycks et al., 2021), Minerva (Lewkowycz et al., 2022), Olympiad benchmarks (He et al., 2024), and AIME 2024–2026 (Zhang & Team, 2024; 2025; 2026). This transfer setting evaluates whether the accuracy-projection objective remains useful once the finite-support boundary is respected in larger offline fine-tuning. Experimental details are in Appx. D.

Table 1: **BigMath fine-tuning results.** APEX improves offline fine-tuning across math reasoning benchmarks. For AIME 24, 25, and 26, we report average accuracy over 32 responses per problem, while the other benchmarks use one response per problem. Parenthesized values are absolute percentage-point gains over the backbone in the same model block. The best scores are shown in bold, and the second-best scores are underlined.

Model	Method	Mathematical Reasoning			AIME avg@32			Summary
		MATH-500	Minerva	Olympiad	AIME 24	AIME 25	AIME 26	Avg.
Qwen2.5 -Math-7B	Backbone	57.2	18.4	25.8	7.4	2.5	2.4	19.0
	GRPO	**62.6** (+5.4)	20.6 (+2.2)	<u>32.0</u> (+6.2)	13.1 (+5.7)	<u>5.9</u> (+3.4)	<u>6.5</u> (+4.1)	<u>23.5</u> (+4.5)
	BOLT	59.6 (+2.4)	20.2 (+1.8)	28.8 (+3.0)	11.7 (+4.3)	4.4 (+1.9)	5.3 (+2.9)	21.7 (+2.7)
	VAR	61.4 (+4.2)	20.2 (+1.8)	28.8 (+3.0)	12.4 (+5.0)	5.5 (+3.0)	5.9 (+3.5)	22.4 (+3.4)
	ReFit	59.4 (+2.2)	19.5 (+1.1)	28.6 (+2.8)	<u>13.8</u> (+6.4)	4.9 (+2.4)	6.2 (+3.9)	22.1 (+3.1)
	PMPO	**62.6** (+5.4)	<u>21.0</u> (+2.6)	31.6 (+5.8)	11.1 (+3.8)	4.3 (+1.8)	4.1 (+1.7)	22.5 (+3.5)
	APEX	<u>62.4</u> (+5.2)	**22.1** (+3.7)	**32.8** (+7.0)	**14.3** (+6.9)	**7.2** (+4.7)	**6.8** (+4.4)	**24.3** (+5.3)
Qwen3-8B	Backbone	70.0	31.6	39.6	8.3	5.2	4.9	26.6
	GRPO	77.6 (+7.6)	<u>40.8</u> (+9.2)	**45.5** (+5.9)	11.5 (+3.1)	10.4 (+5.2)	<u>7.5</u> (+2.6)	<u>32.2</u> (+5.6)
	BOLT	71.6 (+1.6)	34.6 (+2.9)	42.7 (+3.1)	<u>11.6</u> (+3.2)	9.7 (+4.5)	7.1 (+2.2)	29.6 (+3.0)
	VAR	71.2 (+1.2)	33.8 (+2.2)	<u>43.5</u> (+3.9)	10.5 (+2.2)	<u>10.8</u> (+5.6)	6.6 (+1.7)	29.4 (+2.8)
	ReFit	72.6 (+2.6)	33.5 (+1.8)	42.7 (+3.1)	11.2 (+2.9)	9.6 (+4.4)	7.2 (+2.3)	29.5 (+2.9)
	PMPO	**78.8** (+8.8)	40.1 (+8.5)	**45.5** (+5.9)	11.2 (+2.9)	9.7 (+4.5)	<u>7.5</u> (+2.6)	32.1 (+5.5)
	APEX	<u>78.6</u> (+8.6)	**41.5** (+9.9)	**45.5** (+5.9)	**12.8** (+4.5)	**11.8** (+6.6)	**7.8** (+2.9)	**33.0** (+6.4)

The BigMath results in Table 1 extend the ratio evidence to larger offline fine-tuning. On `Qwen2.5-Math-7B`, APEX increases the average score from 19.0% to 24.3%, a $+5.3$-point gain over the backbone and a $+1.8$-point gain over the strongest offline baseline. It also achieves the best scores on Minerva, Olympiad, and all three AIME evaluations, while staying within 0.2 percentage points of online GRPO on MATH-500. On `Qwen3-8B`, APEX reaches the highest average score of 33.0%, compared with 26.6% for the backbone, 32.2% for GRPO, and 32.1% for PMPO. Because every offline method trains on the same retained rollouts, these gains point to differences in the learning signal and cannot be attributed to a data advantage. These results are consistent with the ratio-efficiency mechanism, under which accuracy-projection learning is most useful when sampled groups contain many incorrect responses but still retain enough correct responses to support the target policy. The BigMath gains do not come from a measurable loss of general language ability. Appx. C.3 shows that APEX remains comparable to the offline baselines on PIQA, HellaSwag, WinoGrande, and RACE after math fine-tuning.

5.3 Search Agentic Reinforcement Learning

We extend the evaluation beyond mathematical reasoning to Search-R1 (Jin et al., 2025), where a language model interleaves reasoning with search-tool calls that retrieve observations before producing a final answer. Starting from `Qwen2.5-3B-Instruct`, we train APEX, BOLT, VAR, ReFit, and PMPO on the same logged trajectories from Natural Questions (NQ) and HotpotQA (Kwiatkowski et al., 2019; Yang et al., 2018), with correctness labels determined by the final answer. GRPO serves as an online reference. Evaluation covers held-out NQ and HotpotQA questions and five additional QA datasets to assess transfer. Training and evaluation settings are given in Table 7.

Table 2: Accuracy evaluation for the Search-R1 task on `Qwen2.5-3B-Instruct`. Bold denotes the best score among offline methods in each column.

Method	NQ	TriviaQA	PopQA	HotpotQA	2Wiki	MuSiQue	Bamboogle	Avg.
Backbone	20.89	44.64	28.39	20.16	20.52	5.01	16.13	22.25
GRPO	34.82	56.25	38.14	30.43	30.08	8.65	21.77	31.45
BOLT	24.71	47.53	31.42	22.44	22.19	5.05	21.60	24.99
VAR	22.58	45.11	30.06	19.96	20.86	5.21	18.40	23.17
ReFit	30.06	50.10	33.67	24.88	25.67	5.79	18.40	26.94
PMPO	33.10	51.94	35.93	26.85	24.96	6.79	19.20	28.39
APEX	**38.17**	**57.09**	**40.25**	**32.68**	**33.25**	**9.72**	**24.80**	**33.71**

In Table 2, APEX reaches 33.71% average accuracy, compared with 28.39% for PMPO, the strongest offline baseline, a gain of 5.32 percentage points. It outperforms all four offline baselines on every benchmark and exceeds the online GRPO reference by 2.26 percentage points on average. These gains indicate that verifier-labeled trajectories remain effective for learning search policies from fixed offline data.

6 Conclusion

Offline RLVR involves more than implementing online RLVR at a lower cost. After rollouts are logged, the relevant policy is the accuracy-optimal reference projection supported by those rollouts. That policy is the zero-temperature limit of KL-regularized RLVR, and APEX is policy-consistent for it. With finite logged data, the accuracy gap decreases with $Nq(1-q)$, mixed-correctness groups supply the contrast needed for density-ratio estimation, and weighted or rejection-sampling SFT trades finite-temperature target bias against high weight variance. Empirically, ratio-stratified GSM8K rollouts show the mixed-correctness advantage and a correct-rich finite-data optimum, while BigMath fine-tuning shows transfer to larger offline RLVR runs. Search-R1 results extend these gains beyond mathematical reasoning to multi-turn search and question answering. The method remains restricted to fixed reference support. All-incorrect logged groups indicate that the observed batches lack correct-response support, not that a prompt is unsolvable.

AI USE STATEMENT

We used AI tools to polish writing and formatting of the manuscript. We did not use AI tools for retriving the idea, developing and proving the theoretical arguments, designing and implementing the experiments, and drafting the initial version of the manuscript. The authors take full responsibility for the final content of this work, including all text, claims and code.

ETHICS STATEMENT

This work involves no human-subject data collection, releases no new human-subject dataset, and introduces no user-facing system. The experiments use public research datasets and base models cited in the paper.

REPRODUCIBILITY STATEMENT

The main paper specifies the APEX objective, algorithm, theoretical assumptions, and evaluation protocol, while Appx. B provides complete proofs and Appx. D and D.1 to D.3 document data generation and filtering, baselines, training settings, and evaluation details. The datasets and base models are public research assets cited in the paper. The supplementary material includes code and configuration files for rollout generation, prompt-group filtering, offline and baseline training, answer extraction, and evaluation.

REFERENCES

Abbas Abdolmaleki, Bilal Piot, Bobak Shahriari, Jost Tobias Springenberg, Tim Hertweck, Michael Bloesch, Rishabh Joshi, Thomas Lampe, Junhyuk Oh, Nicolas Heess, Jonas Buchli, and Martin Riedmiller. Learning from negative feedback, or positive feedback or both. In *Proc. ICLR*, 2025.

Alon Albalak, Duy Phung, Nathan Lile, Rafael Rafailov, Kanishk Gandhi, Louis Castricato, Anikait Singh, Chase Blagden, Violet Xiang, Dakota Mahan, and Nick Haber. Big-Math: A large-scale, high-quality math dataset for reinforcement learning in language models. arXiv:2502.17387, arXiv, 2025.

Ashutosh Baheti, Ximing Lu, Faeze Brahman, Ronan Le Bras, Maarten Sap, and Mark Riedl. Leftover lunch: Advantage-based offline reinforcement learning for language models. In *Proc. ICLR*, 2024.

Peter L. Bartlett, Michael I. Jordan, and Jon D. McAuliffe. Convexity, classification, and risk bounds. *J. Amer. Statist. Assoc.*, 101(473):138–156, 2006.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems. arXiv:2110.14168, arXiv, 2021.

Hanze Dong, Wei Xiong, Deepanshu Goyal, Yihan Zhang, Winnie Chow, Rui Pan, Shizhe Diao, Jipeng Zhang, KaShun SHUM, and Tong Zhang. Raft: Reward ranked finetuning for generative foundation model alignment. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856.

Yuhao Du, Zhuo Li, Pengyu Cheng, Zhihong Chen, Yuejiao Xie, Xiang Wan, and Anningzhe Gao. RLHF in an SFT way: From optimal solution to reward-weighted alignment. *TMLR*, 2026.

Scott Fujimoto, David Meger, and Doina Precup. Off-policy deep reinforcement learning without exploration. In *Proc. ICML*, 2019.

Daya Guo et al. DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning. *Nature*, 645(8081):633–638, 2025.

Michael Gutmann and Aapo Hyvärinen. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *Proc. AISTATS*, 2010.

Chaoqun He, Renjie Luo, Yuzhuo Bai, Shengding Hu, Zhen Thai, Junhao Shen, Jinyi Hu, Xu Han, Yujie Huang, Yuxiang Zhang, Jie Liu, Lei Qi, Zhiyuan Liu, and Maosong Sun. OlympiadBench: A challenging benchmark for promoting AGI with olympiad-level bilingual multimodal scientific problems. In *Proc. ACL*, 2024.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the MATH dataset. In *Proc. NeurIPS Datasets and Benchmarks*, 2021.

Bowen Jin, Hansi Zeng, Zhenrui Yue, Jinsung Yoon, Sercan O Arik, Dong Wang, Hamed Zamani, and Jiawei Han. Search-r1: Training llms to reason and leverage search engines with reinforcement learning. In *Proc. COLM*, 2025.

Ilya Kostrikov, Ashvin Nair, and Sergey Levine. Offline reinforcement learning with implicit Q-learning. In *Proc. ICLR*, 2022.

Aviral Kumar, Aurick Zhou, George Tucker, and Sergey Levine. Conservative Q-learning for offline reinforcement learning. In *Proc. NeurIPS*, 2020.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466, 2019.

Nathan Lambert, Jacob Morrison, Valentina Pyatkin, Shengyi Huang, Hamish Ivison, Faeze Brahman, Lester James Validad Miranda, Alisa Liu, Nouha Dziri, Xinxi Lyu, Yuling Gu, Saumya Malik, Victoria Graf, Jena D. Hwang, Jiangjiang Yang, Ronan Le Bras, Oyvind Tafjord, Christopher Wilhelm, Luca Soldaini, Noah A. Smith, Yizhong Wang, Pradeep Dasigi, and Hannaneh Hajishirzi. Tülu 3: Pushing frontiers in open language model post-training. In *Proc. COLM*, 2025.

Sergey Levine, Aviral Kumar, George Tucker, and Justin Fu. Offline reinforcement learning: Tutorial, review, and perspectives on open problems. arXiv:2005.01643, arXiv, 2020.

Aitor Lewkowycz, Anders Andreassen, David Dohan, Ethan Dyer, Henryk Michalewski, Vinay Ramasesh, Ambrose Slone, Cem Anil, Imanol Schlag, Theo Gutman-Solo, Yuhuai Wu, Behnam Neyshabur, Guy Gur-Ari, and Vedant Misra. Solving quantitative reasoning problems with language models. arXiv:2206.14858, arXiv, 2022.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In *Proc. ICLR*, 2024.

Aditya Menon and Cheng Soon Ong. Linking losses for density ratio and class-probability estimation. In *Proc. ICML*, 2016.

Subhojyoti Mukherjee, Viet Dac Lai, Raghavendra Addanki, Ryan A. Rossi, Seunghyun Yoon, Trung Bui, Anup Rao, Jayakumar Subramanian, and Branislav Kveton. Offline RL by reward-weighted fine-tuning for conversation optimization. In *Proc. NeurIPS*, 2025.

XuanLong Nguyen, Martin J. Wainwright, and Michael I. Jordan. Estimating divergence functionals and the likelihood ratio by convex risk minimization. *IEEE Trans. Inf. Theory*, 56(11):5847–5861, 2010.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In *Proc. NeurIPS*, 2022.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D. Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In *Proc. NeurIPS*, 2023.

Mark D. Reid and Robert C. Williamson. Composite binary losses. *JMLR*, 11(83):2387–2422, 2010.

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

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

Yao Shu, Chenxing Wei, Hongbin Lin, Shuang Qiu, and Hui Xiong. Reference-sampled boltzmann projection for KL-regularized RLVR: Target-matched weighted SFT, finite one-shot gaps, and policy mirror descent. arXiv:2605.02469, arXiv, 2026.

Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel M. Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F. Christiano. Learning to summarize with human feedback. In *Proc. NeurIPS*, 2020.

Adith Swaminathan and Thorsten Joachims. Counterfactual risk minimization: Learning from logged bandit feedback. In *Proc. ICML*, 2015.

Jonathan Uesato, Nate Kushman, Ramana Kumar, Francis Song, Noah Siegel, Lisa Wang, Antonia Creswell, Geoffrey Irving, and Irina Higgins. Solving math word problems with process- and outcome-based feedback. arXiv:2211.14275, arXiv, 2022.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. In *Proc. EMNLP*, pp. 2369–2380, Brussels, Belgium, 2018. Association for Computational Linguistics.

Hongyi Yuan, Zheng Yuan, Chuanqi Tan, Wei Wang, Songfang Huang, and Fei Huang. Rrhf: Rank responses to align language models with human feedback. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Proc. NeurIPS*, volume 36, pp. 10935–10950. Curran Associates, Inc., 2023.

Yifan Zhang and Math-AI Team. American invitational mathematics examination (AIME) 2024. Hugging Face dataset, 2024.

Yifan Zhang and Math-AI Team. American invitational mathematics examination (AIME) 2025. Hugging Face dataset, 2025.

Yifan Zhang and Math-AI Team. American invitational mathematics examination (AIME) 2026. Hugging Face dataset, 2026.

Qingyue Zhao, Kaixuan Ji, Heyang Zhao, Tong Zhang, and Quanquan Gu. Towards a sharp analysis of offline policy learning for f-divergence-regularized contextual bandits. In *Proc. ICLR*, 2026.

Daniel M. Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B. Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. Fine-tuning language models from human preferences. arXiv:1909.08593, arXiv, 2019.

A Related Work

Online RLHF and RLVR. RLHF optimizes learned rewards with a KL penalty to a reference model, while DPO exposes the associated policy identity through reference-relative classification (Ziegler et al., 2019; Stiennon et al., 2020; Ouyang et al., 2022; Rafailov et al., 2023). RLVR systems apply correctness rewards in PPO- or GRPO-style online loops (Shao et al., 2024; Lambert et al., 2025; Guo et al., 2025). In contrast, fixed reference rollouts constrain the recoverable policy at logging time; under binary verifiers, we study the zero-temperature accuracy projection on that support rather than a generic weighted-imitation distribution.

Offline alignment from fixed responses. BOLT fits a finite-temperature Boltzmann target, VAR and ReFit use reward-weighted fine-tuning, and RAFT, RRHF, and A-LoL use selection, ranking, or sequence-level advantages (Shu et al., 2026; Du et al., 2026; Mukherjee et al., 2025; Dong et al., 2023; Yuan et al., 2023; Baheti et al., 2024). Whereas these methods use failures as low-weight, rejected, or ranking examples, APEX treats them as evidence about reference mass to remove; correct samples support the projection numerator, making mixed-correctness groups especially informative.

Logged policy learning and density-ratio estimation. Logged bandit and offline-RL theory emphasizes behavior-policy support, coverage, and conservative control of out-of-distribution actions (Swaminathan & Joachims, 2015; Fujimoto et al., 2019; Levine et al., 2020; Kumar et al., 2020; Kostrikov et al., 2022; Zhao et al., 2026). With the language-model reference as the known behavior policy, APEX targets verifier accuracy without value-function learning. Classifier-based density-ratio estimation and proper composite losses motivate fitting reference-relative odds with a logistic likelihood (Gutmann & Hyvärinen, 2010; Reid & Williamson, 2010; Bartlett et al., 2006; Nguyen et al., 2010; Menon & Ong, 2016). Unlike work on mathematical reasoning benchmarks, verifiable data, or process supervision (Cobbe et al., 2021; Albalak et al., 2025; Uesato et al., 2022; Lightman et al., 2024), we treat verifier labels as given and study the reference-supported policy update they identify.

B Proofs and Technical Details

Accuracy projection links the KL-regularized RLVR path, the APEX loss, and the observed ratio effects. The deferred derivations below make that link explicit. The first part derives the finite-temperature KL path and its zero-temperature accuracy projection. The second part proves that the Bernoulli reference-relative learner is policy-consistent for that projection and that finite mixed-correctness groups control the remaining accuracy gap. The remaining arguments address the weighted-SFT bias-variance comparison, the logged-support boundary, local update directions, and the finite-step optimization residuals that are not included in the main theorem's stated bound.

B.1 Accuracy Projection Along the KL-RLVR Path

Proof of Lem. 2.1. Fix a prompt x and write

$$Z_\beta(x) = \mathbb{E}_{y' \sim \pi_{\text{ref}}(\cdot|x)}[\exp(r(x, y')/\beta)]\,. \tag{15}$$

For any $\pi \ll \pi_{\text{ref}}$, the prompt-wise objective equals

$$-\beta D_{\text{KL}}\left(\pi(\cdot \mid x)\middle\| \frac{\pi_{\text{ref}}(\cdot \mid x)\exp(r(x,\cdot)/\beta)}{Z_\beta(x)}\right) + \beta \log Z_\beta(x). \tag{16}$$

The second term is independent of π, and the KL term is uniquely minimized at the normalized Boltzmann reference tilt. □

Proof of Thm. 2.2. For a binary verifier, $r(x, y) = 1$ on $\mathcal{Y}_+(x)$ and $r(x, y) = -1$ on $\mathcal{Y}_-(x)$. The normalizer in Equation (2) is

$$Z_\beta(x) = \pi_{\text{ref}}{}^+(x)e^{1/\beta} + (1 - \pi_{\text{ref}}{}^+(x))e^{-1/\beta}. \tag{17}$$

If $y \in \mathcal{Y}_+(x)$, then $\pi_\beta{}^\star(y \mid x) \to \pi_{\text{ref}}(y \mid x)/\pi_{\text{ref}}{}^+(x)$. If $y \in \mathcal{Y}_-(x)$, then $\pi_\beta{}^\star(y \mid x) \to 0$. Thus the KL path converges to the reference policy conditioned on correct responses. □

Proof of the minimum-KL characterization in Equation (5). Fix a policy $\pi \ll \pi_{\text{ref}}$ with $\pi(\mathcal{Y}_+(x) \mid x) = 1$. On $\mathcal{Y}_+(x)$, the reference policy conditioned on correct responses is

$$\pi_{\text{ref}}(y \mid x,\, y \in \mathcal{Y}_+(x)) = \frac{\pi_{\text{ref}}(y \mid x)}{{\pi_{\text{ref}}}^+(x)} \mathbf{1}\{r(x,y) = 1\}\,. \tag{18}$$

Because π places all mass on $\mathcal{Y}_+(x)$,

$$D_{\text{KL}}(\pi(\cdot \mid x) \| \pi_{\text{ref}}(\cdot \mid x)) = D_{\text{KL}}(\pi(\cdot \mid x) \| \pi_{\text{ref}}(\cdot \mid x,\, y \in \mathcal{Y}_+(x))) - \log {\pi_{\text{ref}}}^+(x). \tag{19}$$

The second term is independent of π, and the first term is uniquely minimized by the reference policy conditioned on $\mathcal{Y}_+(x)$. This policy is exactly $\pi^\star(\cdot \mid x)$. □

Proof of Prop. 2.3. For a binary verifier, the normalizer of the finite-temperature target is

$$Z_\beta(x) = {\pi_{\text{ref}}}^+(x) e^{1/\beta} + (1 - {\pi_{\text{ref}}}^+(x)) e^{-1/\beta}. \tag{20}$$

Therefore its prompt-level verifier accuracy is

$$(\pi_\beta{}^\star)^+(x) = \frac{{\pi_{\text{ref}}}^+(x) e^{1/\beta}}{{\pi_{\text{ref}}}^+(x) e^{1/\beta} + (1 - {\pi_{\text{ref}}}^+(x)) e^{-1/\beta}} = \frac{{\pi_{\text{ref}}}^+(x) e^{2/\beta}}{{\pi_{\text{ref}}}^+(x) e^{2/\beta} + 1 - {\pi_{\text{ref}}}^+(x)}. \tag{21}$$

Subtracting this quantity from one gives the displayed expression for $1 - (\pi_\beta{}^\star)^+(x)$. Because ${\pi_{\text{ref}}}^+(x) \in (0,1)$ and $\beta > 0$, the finite-temperature gap is strictly positive. By Equation (4), $\pi^\star$ is supported on $\mathcal{Y}_+(x)$, so $(\pi^\star)^+(x) = 1$ and $1 - (\pi^\star)^+(x) = 0$. □

Proof of Cor. 2.4. Substitute Equation (6) into the required accuracy-gap condition. Solving

$$\frac{1 - {\pi_{\text{ref}}}^+(x)}{{\pi_{\text{ref}}}^+(x) e^{2/\beta} + 1 - {\pi_{\text{ref}}}^+(x)} \le \epsilon \tag{22}$$

for the boosted positive mass gives

$${\pi_{\text{ref}}}^+(x) e^{2/\beta} \ge \frac{1-\epsilon}{\epsilon}(1 - {\pi_{\text{ref}}}^+(x)), \tag{23}$$

which proves the equivalence. □

B.2 APEX Policy Consistency

Proof of Thm. 4.1. Fix a prompt x with ${\pi_{\text{ref}}}^+(x) > 0$, and write $v(y) = \pi(y \mid x)/\pi_{\text{ref}}(y \mid x)$ for the reference-relative density ratio induced by a candidate policy. Feasibility requires $v(y) \ge 0$ and $\mathbb{E}_{y \sim \pi_{\text{ref}}(\cdot|x)}[v(y)] = 1$. Let $c(x) = \beta \log Z_\beta(x)$ denote the prompt-only KL-coordinate intercept, so the policy-implied score is $f_\pi(x,y) = \beta \log v(y) + c(x)$. The prompt-conditioned APEX risk separates over the positive and negative verifier regions as

$$\mathcal{L}_{\text{APEX}}(v \mid x) = \mathbb{E}_{y \sim \pi_{\text{ref}}(\cdot|x)}[\mathbf{1}\{r = 1\}\, h_+(v(y)) + \mathbf{1}\{r = -1\}\, h_-(v(y))]\,, \tag{24}$$

where

$$h_+(t) = \log\Big(1 + e^{-c(x)} t^{-\beta}\Big), \qquad h_-(t) = \log\Big(1 + e^{c(x)} t^{\beta}\Big). \tag{25}$$

The function $h_-(t)$ is increasing, while $h_+(t)$ is decreasing and strictly convex on $t > 0$. If a feasible v assigns positive density to $\mathcal{Y}_-(x)$, shifting an infinitesimal amount of that density to $\mathcal{Y}_+(x)$ decreases the negative-region loss and also lowers the positive-region loss because h_+ is decreasing. Every minimizer therefore places zero density on $\mathcal{Y}_-(x)$. The feasibility constraint then reduces to

$$\mathbb{E}_{y \sim \pi_{\text{ref}}(\cdot|x)}[v(y) \mathbf{1}\{r(x,y) = 1\}] = 1. \tag{26}$$

Equivalently, under the reference policy conditioned on $\mathcal{Y}_+(x)$, the conditional mean of v is $1/{\pi_{\text{ref}}}^+(x)$. Within the correct region, Jensen's inequality and the strict convexity of h_+ give

$$\mathbb{E}_{y \sim \pi_{\text{ref}}(\cdot|x,\, y \in \mathcal{Y}_+(x))}[h_+(v(y))] \ge h_+\big(\mathbb{E}_{y \sim \pi_{\text{ref}}(\cdot|x,\, y \in \mathcal{Y}_+(x))}[v(y)]\big) = h_+(1/{\pi_{\text{ref}}}^+(x)), \tag{27}$$

with equality only when $v(y) = 1/{\pi_{\text{ref}}}^+(x)$ almost surely on $\mathcal{Y}_+(x)$. The unique minimizing density ratio is therefore

$$v^\star(y) = \frac{\mathbf{1}\{r(x,y) = 1\}}{{\pi_{\text{ref}}}^+(x)}, \qquad \pi(y \mid x) = \frac{\pi_{\text{ref}}(y \mid x)}{{\pi_{\text{ref}}}^+(x)} \mathbf{1}\{r(x,y) = 1\}\,. \tag{28}$$

This induced policy is exactly the accuracy-optimal reference projection $\pi^\star(\cdot \mid x)$. □

B.3 Finite-Sample Accuracy Gap and Ratio Efficiency

Proof of Thm. 4.2. By the definition of ${\pi_\theta}^+(x)$ and $\mathcal{Y}_-(x) = \{y : r(x, y) = -1\}$,

$$1 - {\pi_\theta}^+(x) = \pi_\theta(\mathcal{Y}_-(x) \mid x). \tag{29}$$

Since $\pi_\theta \ll \pi_{\text{ref}}$, the same quantity can be written under reference sampling as

$$\mathbb{E}_{y\sim\pi_{\text{ref}}(\cdot|x)}\left[\frac{\pi_\theta(y \mid x)}{\pi_{\text{ref}}(y \mid x)}\mathbf{1}\{r(x, y) = -1\}\right], \tag{30}$$

which proves the identity. □

Lemma B.1 (Mixed-Correctness Case-Control Reduction)**.** *For a prompt x with $0 < {\pi_{\text{ref}}}^+(x) < 1$, the positive and negative conditional reference laws are*

$$d^+(y \mid x) = \frac{\pi_{\text{ref}}(y \mid x)\mathbf{1}\{r(x, y) = 1\}}{{\pi_{\text{ref}}}^+(x)}, \qquad d^-(y \mid x) = \frac{\pi_{\text{ref}}(y \mid x)\mathbf{1}\{r(x, y) = -1\}}{1 - {\pi_{\text{ref}}}^+(x)}. \tag{31}$$

For a retained group with correct-response fraction $q \in (0, 1)$, the prompt-conditioned population limit of the mixed binary correctness likelihood is

$$\mathcal{L}_q(f \mid x) = q\mathbb{E}_{y\sim d^+(\cdot|x)}[\ell_{\text{APEX}}(+1, f(x, y))] + (1 - q)\mathbb{E}_{y\sim d^-(\cdot|x)}[\ell_{\text{APEX}}(-1, f(x, y))]. \tag{32}$$

The Bayes score for this risk differs from the posterior-odds score separating $d^+(\cdot \mid x)$ and $d^-(\cdot \mid x)$ only by a prompt-only intercept.

Proof of Lem. B.1. Conditioning the retained rollout group on its observed label composition gives the positive law $d^+(\cdot \mid x)$ and the negative law $d^-(\cdot \mid x)$. The empirical class fraction q changes only the class prior of this binary classification problem. For logistic loss, that prior changes the Bayes logit by the prompt-constant shift $\log(q/(1 - q))$. The shift cancels when the reference-relative score is normalized into a policy. □

Proof of Thm. 4.3. Fix a prompt x and suppress it from the notation. Let $\mathcal{L}_q$ be the case-control population risk in Lem. B.1. By the lemma, changing q changes only the prompt-level intercept of the Bayes score, so the normalized policy target remains the accuracy-optimal reference projection. Let $\hat{\theta}$ be an exact empirical minimizer over the bounded accuracy-projection score class. Under the stated local curvature and effective-dimension assumptions, the standard fast-rate bound for locally strongly curved logistic empirical risk minimization gives, with probability at least $1 - \delta$,

$$\mathcal{L}_q(f_{\hat{\theta}}) - \inf_{f\in\mathcal{F}} \mathcal{L}_q(f) \le C\frac{d_{\text{eff}} + \log(1/\delta)}{Nq(1 - q)}. \tag{33}$$

The factor $q(1 - q)$ is the binary case-control curvature supplied by observing both verifier classes. Under realizability, the best class member induces $\pi^\star$ up to a prompt-only shift, so no approximation term appears. The calibrated surrogate-transfer condition in the theorem statement converts logistic excess risk into the remaining negative density in Equation (12), which gives

$$1 - {\pi_{\hat{\theta}}}^+(x) \lesssim \frac{d_{\text{eff}} + \log(1/\delta)}{Nq(1 - q)}. \tag{34}$$

□

The surrogate-transfer step can be stated directly on a bounded density-ratio class. Let $v(y) = \pi_\theta(y \mid x)/\pi_{\text{ref}}(y \mid x)$ and assume $0 \le v(y) \le e^S$. If $v^\star(y) = \mathbf{1}\{r(x, y) = 1\}/{\pi_{\text{ref}}}^+(x)$, then the prompt-wise logistic excess risk bounds the remaining negative mass through the inequality

$$\mathcal{L}_{\text{APEX}}(v \mid x) - \mathcal{L}_{\text{APEX}}(v^\star \mid x) \ge c_{\beta,x,S}\mathbb{E}_{y\sim\pi_{\text{ref}}(\cdot|x)}[v(y)\mathbf{1}\{r(x, y) = -1\}], \tag{35}$$

for a positive constant $c_{\beta,x,S}$. The inequality follows because the positive-region part is minimized by the Jensen-optimal constant $v = 1/{\pi_{\text{ref}}}^+(x)$, while on the negative region

$$\log\left(1 + e^{c(x)}v(y)^\beta\right) \ge c_{\beta,x,S}v(y) \tag{36}$$

on the bounded range. The right-hand side is exactly the accuracy gap by Equation (12).

The main theorem isolates the statistical accuracy-gap term. In the appendix setting, two standard residuals can be added. Relaxing realizability adds a calibrated misspecification term ϵ_{app}. Replacing exact empirical optimization with a T-step optimizer satisfying a local PL inequality with curvature proportional to $q(1-q)$ adds an optimization residual of the form

$$\epsilon_{\text{opt}}(T, q) \lesssim (1 - \alpha_{\text{sgd}}\mu_0 q(1-q))^T \Delta_0 + \frac{\alpha_{\text{sgd}}\beta^2 G^2}{B}. \tag{37}$$

These residuals enter additively and do not change the $Nq(1-q)$ evidence term emphasized in the main text.

Proof of Cor. 4.4. For fixed N, d_{eff}, and δ, the only q-dependent term in Equation (13) is $1/[q(1-q)]$. The product $q(1-q)$ is maximized over $q \in (0,1)$ at $q = 1/2$, which proves the stated minimizer. □

A variance-refined version gives the usual Neyman allocation when the two conditional classes have unequal leading variances. If the leading class-conditional term is $\mathcal{V}_+/q + \mathcal{V}_-/(1-q)$, minimizing it gives

$$q = \frac{\sqrt{\mathcal{V}_+}}{\sqrt{\mathcal{V}_+} + \sqrt{\mathcal{V}_-}}. \tag{38}$$

Additional terms that account for correct-response support, such as a term proportional to $1/(Nq)$, shift the empirical optimum toward mixed-correctness groups with more correct responses.

B.4 Weighted-SFT Baseline Theory

Weighted and rejection-sampling SFT baselines provide the comparison family used in the experiments. They fit nonnegative reweighted imitation distributions, whereas the Bernoulli density-ratio objective uses signed label information. For this comparison family, the population policy has the reference-relative density ratio

$$\rho_w(x, y) = \frac{w(x, y)}{\mathbb{E}_{y' \sim \pi_{\text{ref}}(\cdot|x)}[w(x, y')]}, \qquad w(x, y) \geq 0. \tag{39}$$

This formula is specific to nonnegative weighted or filtered imitation objectives and does not describe offline RLVR in general.

Proposition B.2 (Weighted-SFT Bias-Variance Tradeoff). *For the exact two-level weighted-SFT empirical minimizer with correct-response fraction q and correct-to-incorrect weight ratio R, the learned policy $\hat{\pi}_R$ satisfies, with probability at least $1-\delta$,*

$$1 - \hat{\pi}_R{}^+(x) \leq \frac{1-q}{qR+1-q} + C_w \frac{qR^2+1-q}{(qR+1-q)^2} \frac{d_{\text{eff}} + \log(1/\delta)}{N}. \tag{40}$$

Proof. For a retained group with correct-response fraction q, a two-level weighted-SFT target assigns total normalized prompt-level verifier accuracy

$$Q_R(q) = \frac{qR}{qR+1-q}. \tag{41}$$

Its population target accuracy gap is therefore $\Delta_R(q) = (1-q)/(qR+1-q)$. The normalized sample weight is $R/(qR+1-q)$ on correct responses and $1/(qR+1-q)$ on incorrect responses, so the normalized second moment is

$$M_2(q, R) = \frac{qR^2+1-q}{(qR+1-q)^2}. \tag{42}$$

Standard weighted empirical-risk concentration multiplies the unweighted d_{eff}/N statistical term by this second moment, proving Equation (40). As $R \to \infty$, the target bias vanishes and $M_2(q, R) \to 1/q$, giving the correct-only rejection-sampling limit. □

Theorem B.3 (Stable Nonnegative Imitation Barrier). *For a binary verifier with correct-response set $\mathcal{Y}_+(x)$ and reference verifier accuracy ${\pi_{\text{ref}}}^+(x)$, every nonnegative weighted-imitation policy with $\mathbb{E}_{y\sim\pi_{\text{ref}}(\cdot|x)}[\rho_w(x,y)^2] \le M$ satisfies*

$$ {\pi_w}^+(x) \le \sqrt{M{\pi_{\text{ref}}}^+(x)}, \qquad {\pi_w}^+(x) \ge 1-\varepsilon \implies M \ge \frac{(1-\varepsilon)^2}{{\pi_{\text{ref}}}^+(x)}. \tag{43} $$

Proof. For a binary verifier,

$$ {\pi_w}^+(x) = \mathbb{E}_{y\sim\pi_{\text{ref}}(\cdot|x)}[\rho_w(x,y)\mathbf{1}\{y\in\mathcal{Y}_+(x)\}]\,. \tag{44} $$

Cauchy's inequality gives

$$ {\pi_w}^+(x) \le \sqrt{\mathbb{E}_{y\sim\pi_{\text{ref}}(\cdot|x)}[\rho_w(x,y)^2]\pi_{\text{ref}}(\mathcal{Y}_+(x)\mid x)} \le \sqrt{M{\pi_{\text{ref}}}^+(x)}. \tag{45} $$

The lower bound on M follows by rearranging this inequality under ${\pi_w}^+(x) \ge 1-\varepsilon$. □

For binary verifier rewards, BOLT is a representative finite-temperature member of this weighted-SFT family. Its empirical prompt-wise weights reduce to

$$ \hat{w}_+(k) = \frac{Ne^{2/\beta}}{ke^{2/\beta}+N-k}, \qquad \hat{w}_-(k) = \frac{N}{ke^{2/\beta}+N-k}, \tag{46} $$

where k is the number of correct responses in a group of size N. When $k=0$, every incorrect response has $\hat{w}_-(0)=1$, so the empirical objective on that group becomes ordinary SFT on mistaken completions. When $k>0$ and $\beta\downarrow 0$, the objective approaches correct-only rejection sampling and inherits the $1/q$-type variance described in Prop. B.2.

B.5 Logged-Support Boundaries and Local Dynamics

The support boundary follows directly from the accuracy-projection form and explains why the experiments treat all-incorrect groups as a separate regime. For $p = {\pi_{\text{ref}}}^+(x) > 0$, the least reference-distorting policy under the unit-accuracy constraint is

$$ \pi^\star(y\mid x) = \pi_{\text{ref}}(y\mid x,\, y\in\mathcal{Y}_+(x)) = \frac{\pi_{\text{ref}}(y\mid x)}{p}\mathbf{1}\{r(x,y)=1\}\,. \tag{47} $$

If $p=0$, no policy absolutely continuous with respect to $\pi_{\text{ref}}(\cdot\mid x)$ can place probability one on $\mathcal{Y}_+(x)$. For N independent reference rollouts, the logged correct-response count is defined by

$$ k(x) = \sum_{i=1}^{N}\mathbf{1}\{y_i\in\mathcal{Y}_+(x)\} \tag{48} $$

and satisfies $\Pr[k(x)=0] = (1-{\pi_{\text{ref}}}^+(x))^N$. On the event $k(x)=0$, the observed batch has no correct-response anchor for the numerator of the accuracy-projection density ratio. This limitation can be formalized through an indistinguishability argument. Two verifier worlds can agree on all observed incorrect labels while assigning correctness to two disjoint unobserved regions of reference support. Any offline-only algorithm outputs the same policy in both worlds, so it cannot guarantee unit accuracy in both. Thus an all-incorrect logged group marks a support and coverage boundary and does not establish that no correct completion exists.

Proposition B.4 (Bounded Correctness Evidence). *Let $u_\theta(x,y) = \log(\pi_\theta(y\mid x)/\pi_{\text{ref}}(y\mid x))$ and $s(x,y)=r(x,y)\in\{-1,+1\}$. For each sample, the APEX loss satisfies*

$$ -\frac{\partial\ell_{\text{APEX}}}{\partial u_\theta} = s(x,y)\beta\sigma(-s(x,y)f_\theta(x,y)), \qquad \left|\frac{\partial\ell_{\text{APEX}}}{\partial u_\theta}\right| \le \beta. \tag{49} $$

For a mixed binary prompt with ${\pi_{\text{ref}}}^+(x)\in(0,1)$, correct responses have a negative loss derivative and incorrect responses have a positive loss derivative at reference initialization $u_\theta=0$. Gradient descent therefore raises the reference-relative density of correct responses and lowers that of incorrect responses, with bounded per-sample magnitude.

Proof. Since $f_\theta(x, y) = \beta u_\theta(x, y) + \beta \log Z_\beta(x)$, direct differentiation gives

$$\frac{\partial \ell_{\text{APEX}}}{\partial u_\theta} = -\beta s(x, y)\sigma(-s(x, y)f_\theta(x, y)). \tag{50}$$

The magnitude bound follows from $\sigma(\cdot) \in [0, 1]$. At reference initialization, $u_\theta = 0$ and $f_\theta(x, y) = \beta \log Z_\beta(x)$. For a mixed binary prompt, $\beta \log Z_\beta(x) \in (-1, 1)$. If $r = 1$, the derivative is $-\beta\sigma(-\beta \log Z_\beta(x)) < 0$. If $r = -1$, the derivative is $\beta\sigma(\beta \log Z_\beta(x)) > 0$. The signs give the stated gradient-descent directions. □

Proposition B.5 (Direct Incorrect-Response Dynamics). *Consider one small gradient step from θ_0 on a fixed rollout group, with the common factor $1/N$ absorbed into the step size α_{step}. Let $\phi_i = \nabla_\theta \log \pi_{\theta_0}(y_i \mid x)$, $K_{ij} = \langle \phi_i, \phi_j \rangle$, and $s_i = r(x, y_i)$. For a nonnegative weighted-SFT objective,*

$$\Delta_w \log \pi(y_j \mid x) = \alpha_{\text{step}} \sum_i \hat{w}_i K_{ji}. \tag{51}$$

For APEX,

$$\Delta_{\text{APEX}} \log \pi(y_j \mid x) = \alpha_{\text{step}} \sum_i s_i \omega_i K_{ji}, \qquad \omega_i = \beta\sigma(-s_i f_{\theta_0}(x, y_i)) > 0. \tag{52}$$

For an incorrect sample j, the weighted-SFT self-term is $+\alpha_{\text{step}}\hat{w}_j K_{jj} \geq 0$, whereas the APEX self-term is $-\alpha_{\text{step}}\omega_j K_{jj} < 0$ whenever $K_{jj} > 0$.

Proof. The weighted-SFT loss has first-order update direction $\sum_i \hat{w}_i \nabla_\theta \log \pi_{\theta_0}(y_i \mid x)$, which gives the displayed kernel expansion after Taylor expanding $\log \pi_\theta(y_j \mid x)$. For APEX, Prop. B.4 gives the update coefficient $s_i\omega_i$ on the same log-probability feature ϕ_i. If $s_j = -1$, the self-term changes sign under APEX but remains nonnegative under weighted SFT. □

B.6 One-Label Prompt Boundary Calculations

When ${\pi_{\text{ref}}}^+(x) = 0$ or ${\pi_{\text{ref}}}^+(x) = 1$, the verifier reward is constant over the reference support, so the KL-regularized finite-temperature optimizer is the reference policy. The sample-level APEX derivatives are nevertheless nonzero. All-incorrect samples are pushed downward in log-density ratio, and all-correct samples are pushed upward. This local pressure is well defined in mixed groups because normalization and opposite labels identify the density-ratio direction, but a finite group containing only one label does not supply the same contrast.

The all-incorrect population objective illustrates the boundary. Let $v(y) = \pi(y \mid x)/\pi_{\text{ref}}(y \mid x)$ with $v \geq 0$ and $\mathbb{E}_{\pi_{\text{ref}}}[v] = 1$. If ${\pi_{\text{ref}}}^+(x) = 0$, then $r = -1$ and the prompt-wise loss is

$$\mathcal{L}_{\text{APEX}}(v \mid x) = \mathbb{E}_{y\sim\pi_{\text{ref}}(\cdot|x)}\left[\log\left(1 + e^{-1}v(y)^\beta\right)\right]. \tag{53}$$

If the reference distribution admits sets E_ϵ with probability ϵ, the feasible density $v_\epsilon(y) = \epsilon^{-1}\mathbf{1}\{y \in E_\epsilon\}$ has loss

$$\mathcal{L}_{\text{APEX}}(v_\epsilon \mid x) = \epsilon \log\left(1 + e^{-1}\epsilon^{-\beta}\right) \to 0 \qquad (\epsilon \downarrow 0). \tag{54}$$

Thus the all-incorrect objective can assign a lower loss to concentration on arbitrarily small reference-supported regions than to the reference density ratio $v \equiv 1$. This calculation justifies filtering all-incorrect groups in the large-scale offline experiments because the logged group provides incorrect-side evidence but no observed correct-response support for the accuracy-projection numerator.

All-correct groups have a different boundary. When ${\pi_{\text{ref}}}^+(x) = 1$, the loss is

$$\mathcal{L}_{\text{APEX}}(v \mid x) = \mathbb{E}_{y\sim\pi_{\text{ref}}(\cdot|x)}\left[\log\left(1 + e^{-1}v(y)^{-\beta}\right)\right], \qquad \mathbb{E}_{\pi_{\text{ref}}}[v] = 1. \tag{55}$$

The integrand is strictly convex in $v > 0$, so Jensen's inequality makes $v \equiv 1$ the population minimizer. Finite samples still create upward pressure on observed correct completions, which explains why all-correct groups can act as ordinary imitation data but provide little incorrect-response contrast.

B.7 SGD Stationarity Bound

The main accuracy-gap theorem assumes exact empirical optimization to isolate the statistical role of mixed-correctness evidence in its stated rate. The following stationarity bound records the standard finite-step term for the empirical APEX objective.

Assumption B.6 (Bounded Log-Density-Ratio Gradient)**.** For each prompt x and response y, the gradient of $u_\theta(x, y) = \log(\pi_\theta(y \mid x)/\pi_{\text{ref}}(y \mid x))$ satisfies the bound $\|\nabla_\theta u_\theta(x, y)\| \le G$.

Assumption B.7 (Lipschitz Smoothness)**.** The population loss $\mathcal{L}_{\text{APEX}}$ is L-smooth in θ.

Theorem B.8 (SGD Stationarity Bound)**.** *Let B be the batch size and $\alpha_{\text{sgd}} \le 1/L$ be the SGD step size for the empirical APEX objective in Alg. 1. Under Assumps. B.6 and B.7,*

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\big[\|\nabla\mathcal{L}_{\text{APEX}}(\theta_t)\|^2\big] \le \frac{2\Delta_0}{\alpha_{\text{sgd}}T} + \frac{L\alpha_{\text{sgd}}\beta^2G^2}{B}, \tag{56}$$

where $\Delta_0 = \mathcal{L}_{\text{APEX}}(\theta_0) - \inf_\theta \mathcal{L}_{\text{APEX}}(\theta)$.

Proof. For one sample, $\ell_{\text{APEX}}(r, f) = \log(1 + \exp[-rf])$ satisfies

$$\left|\frac{\partial\ell_{\text{APEX}}}{\partial f}\right| = |r\sigma(-rf)| \le 1. \tag{57}$$

Since $f_\theta(x, y) = \beta u_\theta(x, y) + \beta \log Z_\beta(x)$, the per-sample gradient norm is at most βG. The minibatch gradient is unbiased for $\nabla\mathcal{L}_{\text{APEX}}(\theta_t)$ and has conditional variance at most β^2G^2/B. Smoothness gives

$$\mathbb{E}_t[\mathcal{L}_{\text{APEX}}(\theta_{t+1})] \le \mathcal{L}_{\text{APEX}}(\theta_t) - \alpha_{\text{sgd}}\|\nabla\mathcal{L}_{\text{APEX}}(\theta_t)\|^2 + \frac{L\alpha_{\text{sgd}}^2}{2}\mathbb{E}_t[\|\hat{g}_t\|^2]. \tag{58}$$

Combining this bound with $\alpha_{\text{sgd}} \le 1/L$ and summing over $t = 0, \dots, T-1$ proves the claim. □

C Additional Experimental Evidence

C.1 Controlled Study of Accuracy-Gap Scaling

We conduct a controlled synthetic experiment to study the effects of sample size N and correct-response fraction q on the accuracy gap, as captured by the $1/[Nq(1-q)]$ term in Thm. 4.3. For a fixed prompt, the response space consists of 4096 equally spaced points in $[0, 1)$, each with reference probability $1/4096$. A deterministic verifier labels responses in $[0.2, 0.7)$ as correct and all others as incorrect, so each class carries half the reference probability mass. We train APEX with $\beta = 0.1$ using a Fourier policy $\pi_\theta(y) \propto \pi_{\text{ref}}(y)\exp\big(\theta^\top\phi_d(y)\big)$, where ϕ_d contains $d \in \{32, 64\}$ sine and cosine features and $\theta \in [-20, 20]^d$. Each group contains Nq correct and $N(1-q)$ incorrect responses sampled from the corresponding reference conditionals. We keep the population objective fixed by weighting the two class-conditional losses equally and setting the score intercept using the fixed prior $q_0 = 0.5$.

We first fix $N = 12800$ and vary q over 11 values between 0.1 and 0.9. In each trial, we optimize the policy and compute its accuracy gap exactly as $G(\theta) = 1 - \pi_\theta^+ = \sum_{y:r(y)=-1}\pi_\theta(y)$. Figure 2 (left) shows a broad U-shaped dependence on q, with the smallest gaps near $q = 0.5$.

We then fix $q = 0.5$ and vary N over $\{3200, 6400, 12800, 25600\}$. As shown in Figure 2 (right), the accuracy gap decreases monotonically with N, approaching the population benchmark for each policy dimension. Together, these trends support the benefits of balanced label composition and larger sample sizes predicted by the $1/[Nq(1-q)]$ term.

C.2 All-Incorrect GSM8K Boundary Study

The all-incorrect split isolates the offline support boundary and is not intended to represent a typical training regime. Every response in the retained prompt group is incorrect, so the logged data provide incorrect-side evidence without correct-response support. The boundary calculation in Appx. B.6

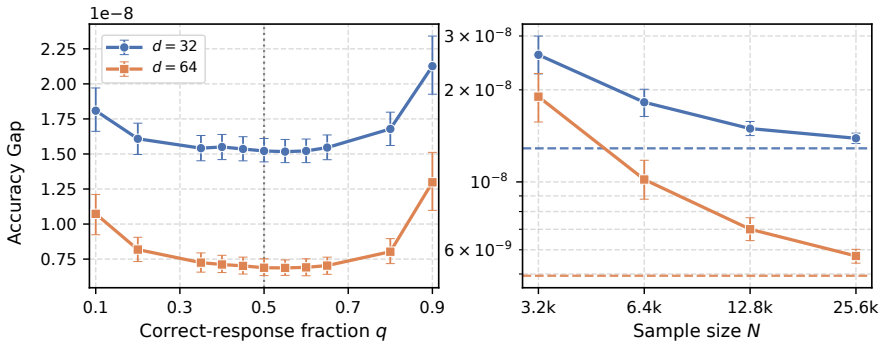


Figure 2: Accuracy gap versus correct-response fraction (left, $N = 12800$) and sample size (right, $q = 0.5$). Each condition is evaluated over 500 trials. Points show trial means with pointwise 95% bootstrap confidence intervals. The dashed horizontal lines in the right panel show the accuracy gaps of policies fitted to the full population objective.

gives the same explanation at the population-objective level. All-incorrect APEX training can favor concentration on part of the support when no observed correct response determines the numerator. Filtering all-incorrect groups in the large-scale experiment therefore enforces a support-and-coverage restriction and is not simply a data-cleaning heuristic.

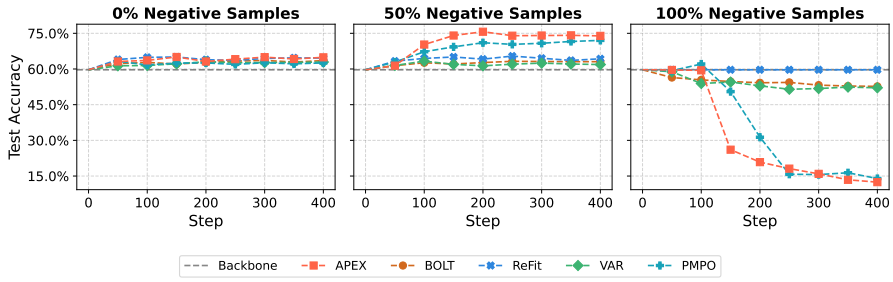


Figure 3: All-incorrect groups expose the offline support boundary. The panels report test accuracy during training for APEX, BOLT (Shu et al., 2026), ReFit (Mukherjee et al., 2025), VAR (Du et al., 2026), and PMPO (Abdolmaleki et al., 2025) under different correct-incorrect ratios, including the all-incorrect split. The dashed horizontal line is the `Qwen2.5-Math-1.5B` backbone.

C.3 General-Capability Retention Study

The incorrect-label signal in APEX could improve math accuracy by over-specializing the model to verifier-labeled reasoning data. The retention study tests this alternative explanation by evaluating the BigMath-tuned checkpoints from Sec. 5.2 on PIQA, HellaSwag, WinoGrande, and RACE, four general language understanding benchmarks.

Table 3: BigMath fine-tuning shows no distinctive forgetting penalty for APEX. Scores report accuracy as a percentage on general language understanding benchmarks, and parentheses give percentage-point changes relative to the corresponding backbone.

Model	Method	PIQA	HellaSwag	WinoGrande	RACE
Qwen2.5 -Math-7B	Backbone	74.2	65.4	65.0	36.8
	GRPO	74.0 (-0.2)	65.3 (-0.1)	64.1 (-0.9)	36.6 (-0.3)
	BOLT	73.9 (-0.3)	65.2 (-0.2)	63.9 (-1.1)	36.6 (-0.3)
	VAR	73.7 (-0.5)	65.2 (-0.2)	63.9 (-1.1)	36.6 (-0.3)
	ReFit	73.7 (-0.5)	65.2 (-0.2)	63.3 (-1.7)	36.7 (-0.1)
	PMPO	73.9 (-0.3)	65.4 (+0.0)	64.2 (-0.8)	36.3 (-0.6)
	APEX	74.2 (-0.1)	65.3 (-0.1)	64.1 (-0.9)	36.1 (-0.8)
Qwen3-8B	Backbone	79.5	78.7	71.3	42.3
	GRPO	79.1 (-0.4)	78.5 (-0.2)	72.1 (+0.8)	41.6 (-0.7)
	BOLT	79.2 (-0.3)	78.6 (-0.1)	71.3 (+0.1)	41.4 (-0.9)
	VAR	79.2 (-0.3)	78.5 (-0.2)	71.9 (+0.6)	41.5 (-0.8)
	ReFit	79.2 (-0.3)	78.5 (-0.1)	71.7 (+0.5)	41.1 (-1.2)
	PMPO	79.3 (-0.2)	78.6 (-0.1)	71.5 (+0.2)	41.4 (-0.9)
	APEX	79.2 (-0.3)	78.5 (-0.2)	72.0 (+0.7)	41.5 (-0.8)

The results in Table 3 show that APEX does not increase general-capability degradation relative to the offline baselines. For both base models, the changes on general benchmarks remain small and comparable to those of BOLT, VAR, ReFit, PMPO, and GRPO. The comparison supports the interpretation that the math gains in Table 1 arise from using the logged rollout groups for accuracy projection, with no broad degradation of non-math capability.

D Experimental Setup

This section describes the setup for the GSM8K, BigMath, and Search-R1 experiments.

D.1 Reference-Policy Rollout Data Generation

Both math studies use the same two-stage generation-and-optimization protocol, matching Alg. 1. For each training prompt x, the corresponding backbone model initializes the reference policy π_{ref}, which samples a group of $N = 8$ responses $\{y_i\}_{i=1}^{N}$. Each response is then scored by the accuracy reward function and assigned a binary verifier reward $r_i = r(x, y_i) \in \{+1, -1\}$, where $+1$ denotes a correct final answer and -1 denotes an incorrect final answer. Boltzmann-weighted baselines use the same binary utility $s(x, y) = r(x, y)$ and $s_i = s(x, y_i)$. Thus APEX and all offline weighted-SFT baselines train on the same fixed response data, which keeps the comparison focused on the training objective without introducing differences in exploration.

For the ratio-stratified GSM8K study, `Qwen2.5-Math-1.5B` serves as both the reference policy and the initial trainable policy. Responses are sampled for the GSM8K training prompts and then filtered by rejection sampling at the prompt-group level. For a prescribed incorrect-response fraction, a prompt group is retained only when the empirical fraction of incorrect responses in $\{y_i\}_{i=1}^{N}$ matches that target. This protocol creates ratio-stratified splits with different correct-incorrect ratios while holding the backbone model, reward function, and group size fixed.

For the large-scale study, 50k prompts from BigMath are paired with two stronger backbones, `Qwen2.5-Math-7B` and `Qwen3-8B`. Unlike the ratio-stratified GSM8K study, the BigMath study does not enforce a fixed correct-incorrect ratio. The protocol keeps the naturally generated response groups and removes all-incorrect groups before offline training. This filtering follows the logged-support boundary in Sec. 2.4, under which the accuracy-projection numerator is observable only when incorrect-response labels are paired with at least one correct response. After filtering, the same remaining prompt groups are used for APEX, BOLT, VAR, ReFit, and PMPO.

D.2 BASELINES

The math experiments compare APEX with four offline weighted-SFT baselines and one online RLVR baseline. The offline methods train on the fixed response sets described in Appx. D.1. Within the math experiments, the online baseline appears only in the large-scale study, where it uses the same training prompt split but generates new responses during policy optimization.

BOLT (Shu et al., 2026) uses reward-weighted supervised fine-tuning with the empirical prompt-wise importance weight

$$w_{\text{BOLT}}(x, y) = \frac{\exp(s(x, y)/\beta)}{\frac{1}{N}\sum_{n=1}^{N} \exp(s(x, y_n)/\beta)}. \tag{59}$$

VAR (Du et al., 2026) is the reward-weighted alignment baseline derived from the KL-regularized target policy, with weight

$$w_{\text{VAR}}(x, y) = \frac{\pi_{\text{ref}}(y \mid x) \exp(s(x, y)/\beta)}{\sum_{y'} \pi_{\text{ref}}(y' \mid x) \exp(s(x, y')/\beta)}. \tag{60}$$

ReFit (Mukherjee et al., 2025) sets the sample-level training coefficient to the verifier reward, so

$$w_{\text{ReFit}}(x, y) = r(x, y). \tag{61}$$

For each logged prompt group, PMPO (Abdolmaleki et al., 2025) uses the accept-and-reject objective

$$\begin{aligned} \widehat{\mathcal{L}}_{\text{PMPO}}(\theta \mid x) = & -\alpha \,\text{Mean}_{i:r_i=+1} \log \pi_\theta(y_i \mid x) \\ & + (1-\alpha)\,\text{Mean}_{i:r_i=-1} \log \pi_\theta(y_i \mid x) + \beta_{\text{PMPO}}\,\text{Mean}_i\, K_i. \end{aligned} \tag{62}$$

Here K_i sums the categorical $\mathrm{KL}(\pi_{\text{ref}} \| \pi_\theta)$ over response-token positions using the cached response prefixes and the full vocabulary. We fix $\alpha = 0.5$ and $\beta_{\text{PMPO}} = 0.5$ for all experiments.

GRPO (Guo et al., 2025) is an online group-relative policy optimization baseline. For a response group sampled from the current policy, it forms group-normalized advantages,

$$\hat{A}_i = \frac{r_i - \text{Mean}(r)}{\text{Std}(r) + \epsilon}, \tag{63}$$

and optimizes the clipped policy-ratio objective

$$\hat{\mathcal{L}}_{\text{GRPO}}(\theta \mid x) = -\frac{1}{N}\sum_{i=1}^{N} \min\left\{ \hat{A}_i \frac{\pi_\theta(y_i \mid x)}{\pi_{\text{ref}}(y_i \mid x)}, \hat{A}_i \,\text{clip}\left(\frac{\pi_\theta(y_i \mid x)}{\pi_{\text{ref}}(y_i \mid x)}, 1-\epsilon, 1+\epsilon \right) \right\}. \tag{64}$$

Table 4: Methods compared in the experiments.

Method	Learning Type	Objective
GRPO	Online RLVR	Clipped policy objective (63) and (64).
BOLT	Offline weighted SFT	Empirical Boltzmann weights.
VAR	Offline weighted SFT	Reference-normalized reward weights.
ReFit	Offline weighted SFT	Reward-weighted fine-tuning.
PMPO	Offline binary feedback	Pos. / Neg. likelihood with reference KL (62).
APEX	Offline accuracy projection	Bernoulli density-ratio objective.

D.3 TRAINING HYPERPARAMETERS

Within each experiment, all offline methods use a shared training configuration. Using the same optimization budget for all offline objectives makes it easier to attribute the reported differences to the learning signal. Tables 5 to 7 report the settings for the GSM8K, BigMath, and Search-R1 experiments.

Table 5: Hyperparameters for the ratio-stratified GSM8K experiment.

Hyperparameter	Value
Base model	`Qwen2.5-Math-1.5B`
Training dataset	`GSM8K` (Cobbe et al., 2021)
Rollout group size N	8
Reward temperature β	0.1
Learning rate	$5e-6$
Batch size	128
Number of epochs	1

Table 6: Hyperparameters for the large-scale experiment.

Hyperparameter	Value
Base models	`Qwen2.5-Math-7B`, `Qwen3-8B`
Training dataset	`BigMath` (Albalak et al., 2025)
Rollout group size N	8
Reward temperature β	0.1
Learning rate	$5e-6$
Batch size	256
Number of epochs	1

Table 7: Training and evaluation settings for Search-R1 experiments.

Setting	Value
Base model	`Qwen2.5-3B-Instruct`
Training dataset	25600 NQ and 25600 HotpotQA
Rollout group size N	5
Retriever	E5-base-v2
Max search turns	4
Reward temperature β	0.1
Learning rate	10^{-6}
Batch size	512
Number of epochs	1