TOKEN UTILITY FOR ON-POLICY DISTILLATION VIA LOCAL TRUST-REGION OPTIMIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Token-selection methods for on-policy distillation (OPD) prioritize student-visited positions using diverse signals, including uncertainty, teacher–student disagreement, and mismatch direction. We instead derive token importance from the distillation objective itself: how much can each position's loss decrease under an equal local KL budget on the student's predictive distribution? The answer is the utility $s_t = g_t^\top F_t^\dagger g_t$, where g_t is the logit gradient and F_t is the categorical Fisher matrix. For reverse-KL it equals the student-weighted variance of the student–teacher log-probability ratio, and for forward-KL it equals the Pearson χ^2 divergence, so the appropriate utility follows from the choice of objective. On shared training checkpoints, several independently motivated criteria, including student entropy, KL disagreement, and TIP, concentrate their importance on nearly the same positions as the derived utility at a 50% selection fraction, while diverging at stricter fractions and for direction-specific criteria. We instantiate the utility as Fisher Utility Weighting (FUW), a continuous, mean-normalized token weighting. Across seven mathematical reasoning benchmarks and three Qwen3 student–teacher pairs, FUW improves over matched uniform OPD under both reverse-KL (1.35–4.0 macro points) and forward-KL (1.08–2.71 macro points), and is competitive with existing selective methods using a single closed-form score derived from the training objective.

1 INTRODUCTION

On-policy distillation (OPD) transfers capabilities from a stronger teacher while retaining the student's state distribution (Agarwal et al., 2024; Gu et al., 2024). The student generates the training trajectories, and the teacher supplies dense distributional supervision at the prefixes that the student actually visits. This avoids the state-distribution mismatch of offline distillation and has become an effective recipe for improving reasoning models. Yet dense supervision does not imply that every supervised position is equally useful. Some student and teacher distributions are already closely aligned; elsewhere they disagree, but the correction may require a large change in the student's predictive distribution. Uniform OPD, which weighs all token positions the same, does not distinguish between these cases.

Recent work addresses this heterogeneity using properties of the visited positions. Entropy-based methods emphasize uncertain decisions (Jin et al., 2026). TIP (Xu et al., 2026) combines student uncertainty with teacher–student divergence; DEAR (Xiao et al., 2026) separates decision tokens from supporting evidence tokens; Token Teachability (Wang et al., 2026) asks whether teacher corrections lie within compatible student support; and TIDE (Yu et al., 2026) distinguishes student-excess from student-deficit probability mass. These approaches operationalize token importance through properties such as uncertainty, disagreement, reasoning role, support compatibility, or mismatch direction. We take a different approach, defining importance through the optimization problem itself: how much can each position's expected distillation objective improve under an equal local budget on predictive-distribution change?

Suppose that every position is allowed the same infinitesimal change in the student's next-token distribution. Which position permits the largest decrease in its expected distillation loss? This equal-budget comparison separates the amount of disagreement from how effectively a local update can resolve it. We formulate an independent local trust-region problem at each student-visited prefix. Using the Fisher geometry induced by KL divergence, the maximum first-order improvement under a

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.

common radius ε is $\sqrt{2\varepsilon s_t}$, where

$$s_t = g_t^\top F_t^\dagger g_t.$$

Here, g_t is the gradient of the token-level distillation objective with respect to the student logits and F_t is the categorical Fisher matrix. For reverse-KL OPD, the general score admits the closed form

$$s_t = \mathrm{Var}_{a\sim p_{S,t}}\left[\log \frac{p_{S,t}(a)}{p_{T,t}(a)}\right].$$

Thus, the variance of the log-probability ratio is not introduced as a disagreement heuristic. It is the optimization utility induced by an equal local distributional budget constraint. We convert this score into Fisher Utility Weighting (FUW), which applies damped, mean-normalized weights to the token-level OPD objective. Unlike selection schemes such as TIP and TIDE, which decide whether a position contributes, FUW uses the utility's magnitude to set how much each position contributes, with weights proportional to the improvement scale $\sqrt{s_t}$. Since the utility is derived from the objective rather than chosen, it also provides a reference against which existing criteria can be compared. We find that several of them concentrate importance on largely the same positions as the derived utility, and that the framework accounts for where they diverge. Our contributions are threefold:

Main Contributions.

- We develop a common optimization framework for token utility in OPD. Comparing the maximum first-order loss reduction under an equal local KL budget yields $s_t = g_t^\top F_t^\dagger g_t$. This utility reduces exactly to student-weighted log-ratio variance for reverse-KL and to $\chi^2(p_T\|p_S)$ for forward-KL. (§2, 3, 4, 5.4)
- On shared checkpoints, student entropy, KL disagreement, gradient norm, TIP Soft-OR, and TIDE-deficit retain nearly all of the derived utility's top-half importance mass, and conversely. This suggests that these independently motivated criteria often identify similar high-importance support despite their different constructions. They diverge at stricter selection fractions and for the direction-specific TIDE-excess criterion. (§5.2)
- We turn the utility into Fisher Utility Weighting (FUW) and evaluate it across seven mathematical reasoning benchmarks and three student–teacher pairs. FUW improves macro mean accuracy over uniform OPD on all three pairs under the same candidate-set construction and loss implementation. Controlled studies examine the roles of score-to-position alignment, weight concentration, continuous weighting, and the distillation objective. (§5.1, 5.3, 5.4)

2 Theory

The KL–Fisher expansion and the natural-gradient solution to a Fisher-constrained steepest-descent problem are classical results (Amari & Nagaoka, 2000; Kakade, 2001). Our contribution is to view token importance in on-policy distillation through the lens of local optimization: comparing positions under a common distributional budget yields objective-specific utilities in closed form. The trust region serves only to derive these per-position utilities; it is not imposed as a constraint during training.

2.1 Problem Formulation

On-policy distillation (OPD) trains the student on trajectories sampled from its own policy while using a teacher model to provide token-level supervision. For a student-generated trajectory $x = (x_1, \ldots, x_T) \sim p_S$, define the sampled token-level reverse-KL log-ratio at position t as

$$\ell_t = \log \frac{p_S(x_t \mid x_{<t})}{p_T(x_t \mid x_{<t})}.$$

The standard reverse-KL OPD objective is

$$\mathcal{L}_{\mathrm{OPD}} = \mathbb{E}_{x\sim p_S}\left[\sum_{t=1}^{T} \ell_t\right].$$

Our goal is to determine how the optimization effort should be allocated across the token positions in this loss. To assess the optimization opportunity at a particular student-visited prefix $h_t = x_{<t}$, define

$$p_{S,t}(\cdot) = p_S(\cdot \mid h_t), \qquad p_{T,t}(\cdot) = p_T(\cdot \mid h_t).$$

Rather than assessing a position using only the particular token x_t sampled on the current rollout, we consider the expected token-level loss over the student's next-token distribution. For $A_t \sim p_{S,t}$,

$$L_t := \mathbb{E}_{A_t \sim p_{S,t}} \left[\log \frac{p_{S,t}(A_t)}{p_{T,t}(A_t)} \right].$$

This expectation is exactly the full-vocabulary reverse-KL:

$$L_t = D_{\mathrm{KL}} \left(p_{S,t} \| p_{T,t} \right).$$

Thus, L_t is the conditional expectation of the sampled token-level OPD log-ratio,

$$L_t = \mathbb{E}[\ell_t \mid h_t].$$

L_t marginalizes the next-token sampling randomness while keeping the student-visited prefix fixed. We use the full-distribution loss $\{L_t\}_{t=1}^T$ at each student visited position to determine how optimization effort should be allocated across these positions.

Let $z_t \in \mathbb{R}^V$ denote the student logits at position t, and define

$$g_t = \nabla_{z_t} L_t.$$

A natural measure of token importance would be the Euclidean gradient magnitude $\|g_t\|_2$. However, this measures sensitivity per unit movement in logit coordinates. Equal Euclidean perturbations of the logits can induce very different changes in the corresponding predictive distributions.

We instead compare positions under an equal local change in the student's predictive distribution. Specifically, for each position t, we ask for the largest first-order reduction in the expected token-level loss that can be achieved while restricting the local change in the student distribution:

$$\max_{\delta_t} \quad -g_t^\top \delta_t \tag{1}$$

$$\text{subject to} \quad D_{\mathrm{KL}} \left(p_{S,t}(\cdot; z_t) \| p_{S,t}(\cdot; z_t + \delta_t) \right) \le \varepsilon. \tag{2}$$

The variable δ_t is an infinitesimal logit perturbation. The constraint compares positions by the induced change in their predictive distributions, rather than by an arbitrary Euclidean logit displacement.

2.2 Maximum Improvement Within the Local KL Ball

Using the local Fisher approximation, the token-specific optimization problem becomes

$$\Delta_t^\star = \max_{\delta_t} -g_t^\top \delta_t \qquad \text{subject to} \qquad \delta_t^\top F_t \delta_t \le 2\varepsilon.$$

This is the classical Fisher-constrained steepest-descent problem (Kakade, 2001). Its solution points in the natural-gradient direction:

$$\delta_t^\star = -\sqrt{\frac{2\varepsilon}{g_t^\top F_t^\dagger g_t}} F_t^\dagger g_t,$$

where $F_t^\dagger$ is the Moore–Penrose pseudoinverse. The pseudoinverse is required because the categorical Fisher is singular along the constant-logit-shift direction. Since L_t depends on the logits only through softmax, g_t lies in the range of F_t. When $g_t = 0$, we take $\delta_t^\star = 0$.

The corresponding maximum first-order improvement is

$$\Delta_t^\star = -g_t^\top \delta_t^\star = \sqrt{2\varepsilon \, g_t^\top F_t^\dagger g_t}.$$

Because every token position is evaluated under the same maximum radius ε, the ordering induced by $\Delta_t^\star$ is identical to the ordering induced by the normalized score

$$s_t := \frac{(\Delta_t^\star)^2}{2\varepsilon} = g_t^\top F_t^\dagger g_t.$$

We call s_t the *local token optimization utility*. It measures the maximum first-order improvement available at token position t after normalizing out the common maximum distributional budget. The full Lagrangian derivation appears in Appendix A.2.

Importantly, this local geometry is determined entirely by the trust-region constraint on the student distribution. It does not depend on the direction of the teacher–student distillation divergence. The choice of distillation objective enters through the gradient g, while the local metric is determined by F.

3 Closed-Form Utility for Reverse-KL OPD

We now specialize the general token utility to the reverse-KL objective used in OPD. The same framework applies to other distillation objectives through their objective-specific gradient g; we derive the corresponding forward-KL utility in Appendix A.4.

Define the vocabulary-wide log-probability ratio

$$r(v) = \log \frac{p_S(v)}{p_T(v)}.$$

Theorem 1 (Reverse-KL Token Optimization Utility)**.** *For the token-level reverse-KL objective*

$$L = D_{\mathrm{KL}}(p_S \| p_T),$$

the local token optimization utility satisfies

$$s = g^\top F^\dagger g = \mathrm{Var}_{v \sim p_S}\left[\log \frac{p_S(v)}{p_T(v)}\right].$$

Proof sketch. Let $r \in \mathbb{R}^V$ denote the vector with entries $r_v = \log(p_S(v)/p_T(v))$. Differentiating the reverse-KL with respect to the student logits gives

$$g = Fr,$$

where

$$F = \mathrm{diag}(p_S) - p_S p_S^\top.$$

Therefore,

$$g^\top F^\dagger g = r^\top F F^\dagger F r = r^\top F r. \tag{3}$$

For the categorical Fisher,

$$r^\top F r = \sum_v p_S(v) r_v^2 - \left(\sum_v p_S(v) r_v\right)^2 \tag{4}$$

$$= \mathrm{Var}_{v \sim p_S}[r_v]. \tag{5}$$

Substituting the definition of r_v gives the result. A complete derivation is provided in Appendix A.3.

Practical computation. The utility can be computed directly in $O(V)$ without constructing or inverting the Fisher matrix:

$$s = \sum_v p_S(v) \left(r_v - \sum_u p_S(u) r_u\right)^2.$$

Since student and teacher logits are already available during distillation, this requires only elementwise operations and reductions.

Importantly, variance is not posited as a heuristic measure of token importance; it arises from the Fisher-constrained optimization problem. Utility is high when student-supported actions have sufficiently different teacher–student log-ratios to permit substantial improvement through local probability redistribution, and low when those log-ratios are nearly constant.

Scope. The analysis is local to each student-visited predictive distribution and does not model coupling through shared transformer parameters. Thus, the derived utility characterizes local optimization opportunity rather than a globally optimal weighting rule; our end-to-end experiments test whether this signal remains useful under parameter coupling.

4 TOKEN IMPORTANCE UNDER REVERSE-KL OPD

Relationship to existing criteria. Existing criteria capture different notions of token importance. TIP develops an optimization-motivated account but operationalizes its practical Soft-OR criterion through student uncertainty and teacher–student disagreement, while TIDE distinguishes student-excess from student-deficit mismatch. FUW instead derives a continuous weighting signal directly from the distillation objective through $g_t^\top F_t^\dagger g_t$, which changes with the objective. Table 4 in Appendix B.2 summarizes the existing criteria and Section 5 compares their effective support empirically.

Fisher Utility Weighting (FUW) We now turn the reverse-KL utility derived in Section 3 into a practical training rule. Theorem 1 defines Fisher utility on the full vocabulary. For training, we use the candidate set

$$\mathcal{C}_t = \mathrm{Top}_{16}(p_{S,t}) \cup \mathrm{Top}_{16}(p_{T,t}), \qquad |\mathcal{C}_t| \le 32, \tag{6}$$

and renormalize each distribution on this set. For $m \in \{S, T\}$ and $v \in \mathcal{C}_t$, let

$$\widehat{p}_{m,t}(v) = \frac{p_{m,t}(v)}{\sum_{u \in \mathcal{C}_t} p_{m,t}(u)}. \tag{7}$$

We compute the restricted log-ratio, clip it, and evaluate its student-weighted variance:

$$r_t^{\mathcal{C}}(v) = \log \frac{\widehat{p}_{S,t}(v)}{\widehat{p}_{T,t}(v)}, \qquad r_t^{\mathrm{clip}}(v) = \mathrm{clip}\big(r_t^{\mathcal{C}}(v), -c, c\big), \qquad \widehat{s}_t = \mathrm{Var}_{v \sim \widehat{p}_{S,t}}\Big[r_t^{\mathrm{clip}}(v)\Big]. \tag{8}$$

We set $c = 5$, so $0 \le \widehat{s}_t \le c^2$. Without clipping, $\widehat{s}_t$ is the Fisher utility of the reverse-KL loss on $\mathcal{C}_t$, with the candidate set held fixed. Clipping stabilizes this score and makes it an approximation to that utility.

The per-position training loss is the reverse-KL on the same renormalized candidate set:

$$\widehat{L}_t = D_{\mathrm{KL}}(\widehat{p}_{S,t} \| \widehat{p}_{T,t}). \tag{9}$$

For the N valid positions in a batch, we set $u_t = \sqrt{\widehat{s}_t}$ and use mean-one weights

$$w_t = \begin{cases} \dfrac{N u_t}{\sum_{j=1}^N u_j}, & \sum_{j=1}^N u_j > 0, \\ 1, & \sum_{j=1}^N u_j = 0. \end{cases} \tag{10}$$

The resulting objective is

$$\mathcal{L}_{\mathrm{FUW}} = \frac{1}{N} \sum_{t=1}^N \mathrm{stopgrad}(w_t) \widehat{L}_t. \tag{11}$$

Uniform OPD uses the same candidate set, renormalization, and per-position loss, with $w_t = 1$ for every position. Thus, its comparison with FUW changes the weighting rule while holding this loss implementation fixed.

The square-root transform follows from the equal-radius improvement scale, $\Delta_t^\star \propto \sqrt{s_t}$, and also dampens score concentration. We therefore use $u_t = \sqrt{\widehat{s}_t}$ as a theoretically motivated weighting signal. We normalize these utilities to have unit mean and treat the resulting weights as stop-gradient during optimization. Unless otherwise specified, Fisher Utility Weighting (FUW) refers to this reverse-KL instantiation; Section 5.4 applies the same construction to forward KL using its corresponding objective-specific utility.

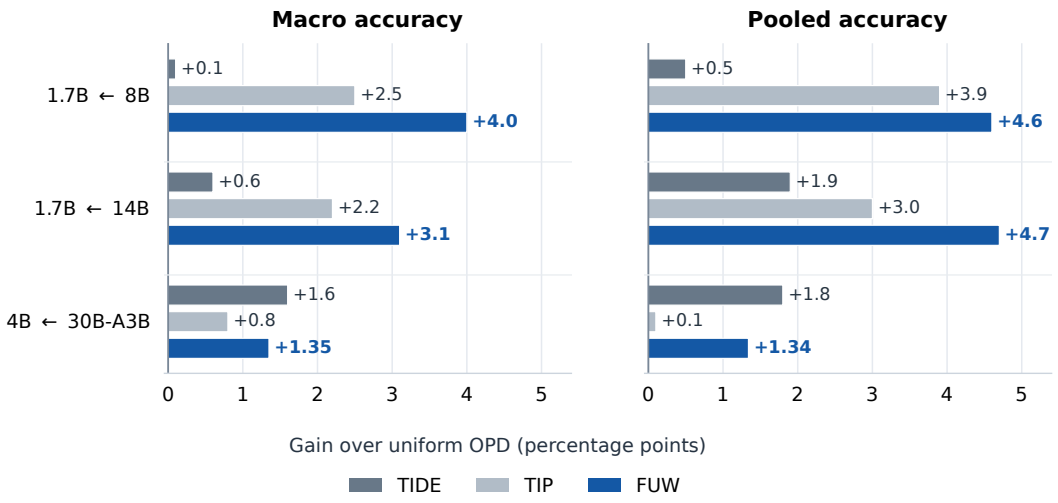

Figure 1: Macro and pooled average gains over uniform OPD across three Qwen3 student–teacher pairs. Bars show percentage-point differences in the reported point estimates. Macro weights the seven benchmarks equally; pooled weights all 1,620 problems equally. Uniform OPD and FUW use the same candidate-set construction and restricted reverse-KL loss, differing in their token weights. All methods share the same training setup. Uniform OPD, TIP, and FUW use the same candidate-set construction; TIDE uses its native support (Appendix B.1).

5 Experiments

Our experiments evaluate the derived token utility from these complementary perspectives: end-to-end distillation performance, its relationship to existing token-importance criteria, and controlled variations of the scoring and weighting procedure. We evaluate both reverse- and forward-KL OPD across three student–teacher pairs and seven mathematical-reasoning benchmarks, using the objective-specific utility induced by each distillation objective. Additional concentration, robustness, and diagnostic analyses are reported in Appendix B.

5.1 RQ1: Does FUW improve over uniform reverse-KL OPD across model pairs?

We evaluate reverse-KL OPD on three Qwen3 student–teacher pairs: Qwen3-1.7B-Base ← Qwen3-8B, Qwen3-1.7B-Base ← Qwen3-14B, and Qwen3-4B-Base ← Qwen3-30B-A3B-Instruct-2507. For all model pairs, training uses DAPO-Math-17k (Yu et al., 2025) for 279 steps (one epoch), 256 sequences per step, and an 8,192 tokens maximum training context with maximum response length set to 7,168 tokens and maximum prompt length set to 1,024 tokens. FUW uses $w_t \propto \sqrt{\widehat{s}_t}$. Figure 1 summarizes aggregate performance; per-benchmark results along with details of the evaluation setup and baselines are in Appendix B.1.

Benchmarks. We evaluate on 1,620 problems from seven mathematical-reasoning benchmarks: AIME24 (30), AIME25 (30), AMC23 (83), BRUMO25 (30), MATH-500 (500), Minerva (272), and OlympiadBench (675). AIME24/25, AMC23, and BRUMO25 form the four smaller benchmarks. Evaluations use Macro Avg@8, which weights benchmarks equally, and pooled Avg@8, which weights problems equally.

Results. FUW improves on uniform OPD on all three pairs. Its Macro Avg@8 gains are 4.0, 3.1, and 1.35 percentage points for the 1.7B ← 8B, 1.7B ← 14B, and 4B ← 30B-A3B pairs, respectively. The corresponding pooled gains are 4.6, 4.7, and 1.34 points. Thus, the gain over uniform OPD persists across the evaluated model pairs, although its size varies.

On the 1.7B ← 8B pair, FUW achieves the highest aggregate scores among the trained students: 26.8 Macro Avg@8 and 39.2 pooled Avg@8. It exceeds TIP Soft-OR by 1.5 and 0.7 points, respectively, and has the highest point estimate on five of the seven individual benchmarks among the trained

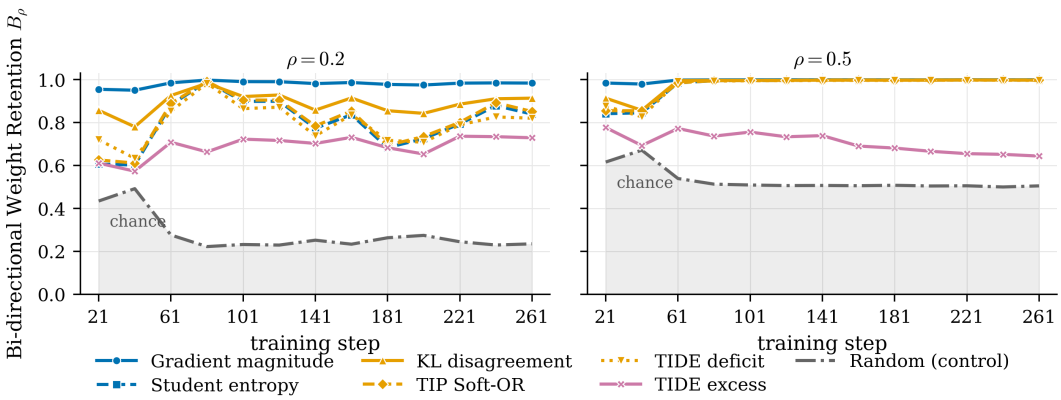

Figure 2: Effective-support agreement B_ρ with FUW during uniform reverse-KL OPD of Qwen3-1.7B-Base ← Qwen3-14B. At $\rho = 0.2$, agreement is criterion- and checkpoint-dependent; at $\rho = 0.5$, most criteria approach near-complete retention, while TIDE-excess remains distinct. The random-selection control accounts for agreement induced by the concentration of FUW's weights.

methods. The pooled FUW–TIP difference is small: its paired-bootstrap 95% confidence interval includes zero.

FUW also has the highest aggregate point estimates among trained methods on the 1.7B ← 14B pair. On the 4B ← 30B-A3B pair, TIDE is slightly ahead: its Macro and pooled Avg@8 are 46.4 and 58.0, compared with FUW's 46.15 and 57.54. Thus, FUW consistently improves over matched-implementation uniform OPD, while its differences from selective baselines are smaller and less uniform.

5.2 RQ2: Do different token-importance criteria prioritize the same positions?

We compare FUW with Euclidean gradient magnitude, reverse-KL disagreement, student entropy, TIP Soft-OR, and the excess and deficit components of TIDE; their definitions and relationships to FUW are summarized in Table 4 in Appendix B.2. We compute all these token importance criteria on the checkpoints obtained during uniformly weighted reverse-KL OPD of Qwen3-1.7B-Base ← Qwen3-14B. Since we use uniformly weighted OPD the token-importance criteria do not determine the model states being compared.

Different importance criteria may disagree on the exact ordering of positions while still concentrating their weight on similar positions of the trajectory. We therefore measure agreement in both directions. Let S_X^ρ denote the top-ρ fraction of positions selected by criterion X, and let $w_{X,i}$ denote its weight on position i. We define

$$\mathrm{REV}_\rho(X) = \frac{\sum_{i \in S_X^\rho} w_{\mathrm{FUW},i}}{\sum_{i \in S_{\mathrm{FUW}}^\rho} w_{\mathrm{FUW},i}}, \qquad \mathrm{CONT}_\rho(X) = \frac{\sum_{i \in S_{\mathrm{FUW}}^\rho} w_{X,i}}{\sum_{i \in S_X^\rho} w_{X,i}}.$$

REV asks how much of FUW's attainable top-ρ weight is retained when positions are selected by X, while CONT asks the converse. We summarize the two directions using

$$B_\rho(X) = \min\{\mathrm{REV}_\rho(X), \mathrm{CONT}_\rho(X)\},$$

which we call bidirectional weight retention. Importantly, high B_ρ measures agreement in effective high-importance support, not equality of the underlying weight vectors. Two criteria can attain $B_\rho \approx 1$ while assigning different relative weights to individual positions within that shared support.

We evaluate $\rho = 0.2$, close to TIDE's realized active fraction (0.264), and $\rho = 0.5$, matching TIP's 50% Soft-OR gate. As shown in Figure 2, at $\rho = 0.2$, agreement is substantial but criterion- and checkpoint-dependent, with gradient magnitude remaining particularly close to FUW. At $\rho = 0.5$, gradient magnitude, student entropy, KL disagreement, TIP Soft-OR, and TIDE-deficit achieve $B_{0.5} \geq 0.985$ from step 61 onward, whereas TIDE-excess remains distinct. Thus, for the Qwen3-1.7B-Base ← Qwen3-14B pair, several independently motivated criteria identify nearly the same

effective high-importance support at the broader operating fraction, although they may assign different relative weights within that support. Appendix B.2 further analyzes $B_\rho(X)$ into its directional components $\mathrm{REV}_\rho(X)$ and $\mathrm{CONT}_\rho(X)$ and defines the random-control, showing that the near-unit $B_{0.5}$ values for most criteria reflect nearly symmetric agreement, whereas TIDE-excess exhibits asymmetric containment. It also reports a cross-pair robustness analysis on Qwen3-1.7B-Base ← Qwen3-8B along an FUW-trained trajectory; although these checkpoints are not independent of the reference criterion, the same qualitative effective-support structure appears at both $\rho = 0.2$ and 0.5.

5.3 RQ3: Which components of FUW matter?

These ablations use reverse-KL distillation on the Qwen3-1.7B-Base ← Qwen3-8B pair with the same training setup as RQ1. We separate the roles of the metric, position-specific utility information, and weighting form. Replacing $g_t^\top F_t^\dagger g_t$ with the Euclidean score $\|g_t\|_2^2$ gives comparable downstream accuracy, although its weights become substantially more concentrated during training: effective retention falls to 0.104 and the maximum token weight reaches 96. FUW avoids this late-stage concentration, although the two metrics are not separated by downstream accuracy at the present training budget. Permuting FUW weights within each batch, while preserving their marginal distribution, lowers pooled Avg@8 by 1.02 points ($p = 0.028$; Table 6), showing that position-specific utility information matters. In contrast, a hard top-half FUW mask reaches 38.4 pooled Avg@8, compared with 39.2 for continuous FUW and 38.4 for TIP Soft-OR at the same 50% retention fraction. This is consistent with RQ2, where FUW and TIP identify nearly the same effective support at $\rho = 0.5$. The complete paired contrasts appear in Appendix B.3.

5.4 RQ4: How does the token-utility framework extend to forward-KL OPD?

The token-utility framework extends directly to forward-KL distillation: applying the same local trust-region formulation yields the Pearson $\chi^2(p_T\|p_S)$ divergence as the induced utility, rather than the log-ratio variance obtained for reverse-KL. A complete derivation of the forward-KL utility is provided in Appendix A.4. For practical weighting, we upper-clip the log-ratio $\tilde{r} = \log(p_T/p_S)$ at $\tau = 3$, i.e., $p_T/p_S \le e^\tau \approx 20$, when computing the detached utility weights. The utility is computed on the same candidate set C_t, and weights are formed as in Eqs. (10)–(11). The forward-KL loss itself is unchanged.

We test whether weighting forward-KL OPD using its corresponding utility improves over uniform forward-KL distillation. We use the same experimental setup as in RQ1, changing only the objective-specific components: the forward-KL loss and its corresponding Pearson $\chi^2(p_T\|p_S)$ utility. Uniform forward-KL uses the same forward-KL loss with uniform token weights. Objective-specific utility weighting improves both Macro and pooled Avg@8 on all three model pairs. Averaged equally across the three model pairs, utility weighting improves Macro Avg@8 by 1.65 points and pooled Avg@8 by 1.95 points. Detailed results are shown in Table 7 in Appendix B.4.

Thus, the same local trust-region principle developed in Section 2 yields a usable utility for a different objective: the Pearson $\chi^2(p_T\|p_S)$ divergence for forward-KL, in place of the log-ratio variance for reverse-KL. Weighting by the corresponding objective-specific utility improves aggregate Avg@8 over uniform weighting under both reverse- and forward-KL distillation.

6 Related Work

On-policy distillation. On-policy distillation (OPD) trains a student model on trajectories sampled from its own policy while receiving token-level supervision from a stronger teacher (Agarwal et al., 2024; Gu et al., 2024). Recent work has demonstrated that OPD substantially improves reasoning performance over offline distillation, while simultaneously exposing new optimization challenges unique to student-generated trajectories. Luo et al. (2026) attribute instability to repetitive-token advantage amplification, while Li et al. (2026) characterize failure modes arising from signal degradation on drifted prefixes. Other work has proposed stabilizing modifications including reward extrapolation (Yang et al., 2026), relaxed clipping strategies (Ko et al., 2026), and entropy-aware training recipes (Jin et al., 2026). Collectively, these studies suggest that the distillation signal is highly non-uniform across token positions and motivate selective supervision strategies.

Token importance and selective supervision. Entropy-based methods prioritize uncertain positions (Wang et al., 2025; Jin et al., 2026). TIP (Xu et al., 2026) combines uncertainty and teacher–student divergence; DEAR (Xiao et al., 2026) distinguishes decision and supporting evidence tokens; Token Teachability (Wang et al., 2026) measures whether disagreement lies within compatible student support; and TIDE (Yu et al., 2026) separates student-excess from student-deficit mass. REOPOLD (Ko et al., 2026) instead treats token log-ratios as rewards and combines clipping with entropy-guided sampling. FUW is complementary: it derives a score from achievable loss reduction under a common distributional budget rather than prescribing a token property.

Information geometry and optimization-based token utility. Our work is based on the information-geometric view of optimization. The local Fisher geometry induced by KL divergence and the resulting natural-gradient formulation have a long history in machine learning (Amari & Nagaoka, 2000; Kakade, 2001). Trust-region methods such as TRPO similarly constrain policy updates using KL divergence (Schulman et al., 2015), while PPO provides a practical approximation to trust-region policy optimization (Schulman et al., 2017). Rather than applying these ideas to a shared parameter-space update, we use the local KL geometry as an analytical lens for characterizing individual student-visited positions. Specifically, we define token utility as the maximum first-order reduction in the token-level distillation objective under a common local budget on predictive-distribution change. This yields the Fisher-normalized quantity $g_t^\top F_t^\dagger g_t$, which reduces to student-weighted log-ratio variance for reverse KL and Pearson $\chi^2(p_T\|p_S)$ divergence for forward KL. This provides a complementary perspective to recent notions based on uncertainty, disagreement, reasoning role, support compatibility, or mismatch type: rather than specifying a property that makes a token informative or learnable, we define the local optimization problem and let utility emerge from its solution.

7 Discussion and Conclusion

We derived token utility for on-policy distillation (OPD) from the objective being optimized. The local trust-region formulation defines utility through the optimization opportunity induced by the distillation objective, yielding the general quantity $g_t^\top F_t^\dagger g_t$. Under reverse-KL this reduces to the student-weighted variance of the log-probability ratio, while under forward-KL it becomes the Pearson $\chi^2(p_T\|p_S)$ divergence. Across three student–teacher pairs, weighting by the corresponding utility improves aggregate performance over uniform weighting under both objectives.

Distinct importance criteria can nevertheless identify similar effective high-importance support. Under reverse-KL OPD, several criteria with different motivations exhibit near-complete bidirectional retention with FUW at the broader operating fraction, while agreement is more criterion- and checkpoint-dependent at the more selective fraction and TIDE-excess remains distinct. Importantly, shared effective support does not imply identical weighting behavior, as the criteria can distribute importance differently within that support.

Our analysis is local and first-order: it evaluates predictive distributions independently at fixed student-visited prefixes, whereas shared transformer parameters couple positions during training. Accordingly, the derived utility characterizes local optimization opportunity rather than a globally optimal shared-parameter weighting rule. Our experiments are restricted to Qwen3 model pairs and mathematical reasoning, leaving broader architectures and agentic tasks for future study.

Overall, the framework provides an optimization-based construction for token utility in OPD: specify the distillation objective, derive its local utility under a common predictive-distribution budget, and use that utility to characterize or prioritize student-visited positions.

AI USE STATEMENT

The authors used LLM assistants to edit prose, check consistency between tables and text, and implement plotting functions for figures. The authors designed the experiments, verified every reported number against the run artifacts, and take full responsibility for the content.

ETHICS STATEMENT

This work studies token weighting for on-policy distillation of language models on publicly available mathematical-reasoning data and benchmarks. It involves no human subjects or personal data. Distillation transfers capabilities from a larger teacher to a smaller student; the student may therefore inherit the teacher's errors and biases, and more efficient distillation lowers the cost of deploying capable models. We evaluate only mathematical reasoning and make no claims about safety-relevant behavior of the distilled students.

REPRODUCIBILITY STATEMENT

Section 4 specifies the Fisher Utility Weighting (FUW) score, candidate-set construction, clipping, and weighting. Section 5.1 and Appendix B.1 give the model pairs, training data, training budget, sampling settings, and evaluation protocol. The Appendix A contains complete derivations for all theoretical results. Each configuration is trained once; reported uncertainty reflects paired bootstrap resampling over evaluation problems. For every run we retain the configuration, intermediate checkpoints, and raw evaluation outputs. We will release code and configurations upon publication.

REFERENCES

Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stanczyk, Sabela Ramos, Matthieu Geist, and Olivier Bachem. On-policy distillation of language models: Learning from self-generated mistakes. In *International Conference on Learning Representations (ICLR)*, 2024. URL https://arxiv.org/abs/2306.13649.

S. Amari and H. Nagaoka. *Methods of Information Geometry*. Fields Institute Communications. American Mathematical Society, 2000. ISBN 9780821805312. URL https://books.google.com/books?id=qGzCwAEACAAJ.

Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. Minillm: On-policy distillation of large language models. In *International Conference on Learning Representations (ICLR)*, 2024.

Woogyeol Jin, Taywon Min, Yongjin Yang, Dennis Wei, Yi Zhou, Swanand Ravindra Kadhe, Nathalie Baracaldo, and Kimin Lee. Entropy-aware on-policy distillation of language models. In *International Conference on Machine Learning (ICML)*, 2026.

Sham Kakade. A natural policy gradient. In *Proceedings of the 15th International Conference on Neural Information Processing Systems: Natural and Synthetic*, NIPS'01, pp. 1531–1538, Cambridge, MA, USA, 2001. MIT Press.

Jongwoo Ko, Sara Abdali, Young Jin Kim, Tianyi Chen, and Pashmina Cameron. Scaling reasoning efficiently via relaxed on-policy distillation. *arXiv:2603.11137*, 2026. URL https://arxiv.org/abs/2603.11137.

Yaxuan Li, Yuxin Zuo, Bingxiang He, Jinqian Zhang, Chaojun Xiao, Cheng Qian, Tianyu Yu, Huanang Gao, Wenkai Yang, Zhiyuan Liu, and Ning Ding. Rethinking on-policy distillation of large language models: Phenomenology, mechanism, and recipe. *arXiv preprint arXiv:2604.13016*, 2026.

Feng Luo, Yu-Neng Chuang, Guanchu Wang, Zicheng Xu, Xiaotian Han, Tianyi Zhang, and Vladimir Braverman. Demystifying opd: Length inflation and stabilization strategies for large language models. *arXiv preprint arXiv:2604.08527*, 2026.

John Schulman, Sergey Levine, Philipp Moritz, Michael Jordan, and Pieter Abbeel. Trust region policy optimization. In *International Conference on Machine Learning*, 2015.

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

Shenzhi Wang, Le Yu, Chang Gao, Chujie Zheng, Shixuan Liu, Rui Lu, Kai Dang, Xionghui Chen, Jianxin Yang, Zhenru Zhang, Yuqiong Liu, An Yang, Andrew Zhao, Yang Yue, Shiji Song, Bowen Yu, Gao Huang, and Junyang Lin. Beyond the 80/20 rule: High-entropy minority tokens drive effective reinforcement learning for llm reasoning, 2025. URL https://arxiv.org/abs/2506.01939.

Yuanyi Wang, Su Lu, Yanggan Gu, Pengkai Wang, Yifan Yang, Zhaoyi Yan, Congkai Xie, Jianmin Wu, and Hongxia Yang. Not all disagreement is learnable: Token teachability in on-policy distillation. *arXiv: 2605.26844*, 2026.

Jinwei Xiao, Zhuowen Han, Yueqing Sun, Zhengxi Lu, Yuxin Liu, Zhiyuan Yao, Wentao Chen, Qi Gu, and Xunliang Cai. Finding the evidence: Discovering decision-supporting tokens for on-policy reasoning distillation. *arXiv:2606.22830*, 2026. URL https://arxiv.org/abs/2606.22830.

Yuanda Xu, Hejian Sang, Zhengze Zhou, Ran He, Zhipeng Wang, and Alborz Geramifard. Tip: Token importance in on-policy distillation. *arXiv preprint arXiv:2604.14084*, 2026. URL https://arxiv.org/abs/2604.14084.

Wenkai Yang, Weijie Liu, Ruobing Xie, Kai Yang, Saiyong Yang, and Yankai Lin. Learning beyond teacher: Generalized on-policy distillation with reward extrapolation. *arXiv preprint arXiv:2602.12125*, 2026.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, Xin Liu, Haibin Lin, Zhiqi Lin, Bole Ma, Guangming Sheng, Yuxuan Tong, Chi Zhang, Mofan Zhang, Wang Zhang, Hang Zhu, Jinhua Zhu, Jiaze Chen, Jiangjie Chen, Chengyi Wang, Hongli Yu, Yuxuan Song, Xiangpeng Wei, Hao Zhou, Jingjing Liu, Wei-Ying Ma, Ya-Qin Zhang, Lin Yan, Mu Qiao, Yonghui Wu, and Mingxuan Wang. Dapo: An open-source llm reinforcement learning system at scale, 2025. URL https://arxiv.org/abs/2503.14476.

Zichao Yu, Chengzhi Yu, Shengze Xu, Yujin Han, Bingqing Jiang, Xu Wang, and Difan Zou. Mismatch matters: On-policy distillation beyond token agreement. *arXiv:2608.09836*, 2026. URL https://arxiv.org/abs/2608.09836.

Appendices

A Additional Theoretical Derivations

A.1 Derivation of the Fisher-Metric Constraint

In the main text we constrain the amount by which the student's predictive distribution changes,

$$D_{\mathrm{KL}}(p_S(\cdot; z) \,\|\, p_S(\cdot; z + \delta)) \leq \varepsilon.$$

This appendix derives the corresponding local quadratic approximation.

Define

$$C(\delta) = D_{\mathrm{KL}}(p_S(\cdot; z) \,\|\, p_S(\cdot; z + \delta)) \,.$$

Using the definition of KL divergence,

$$C(\delta) = \sum_i p_S(i; z) \log \frac{p_S(i; z)}{p_S(i; z + \delta)}.$$

Since the first term is independent of δ,

$$C(\delta) = -\sum_i p_S(i; z) \log p_S(i; z + \delta) + \text{const}.$$

Taking the derivative with respect to δ_j,

$$\frac{\partial C}{\partial \delta_j} = -\sum_i p_S(i) \frac{\partial}{\partial \delta_j} \log p_S(i; z + \delta).$$

Using the softmax identity

$$\frac{\partial \log p(i)}{\partial z_j} = \mathbf{1}[i = j] - p(j),$$

gives

$$\frac{\partial C}{\partial \delta_j} = p_S(j; z + \delta) - p_S(j; z).$$

Evaluating at the expansion point,

$$\delta = 0,$$

yields

$$\nabla C(0) = 0.$$

Similarly,

$$C(0) = 0.$$

Therefore,

$$C(\delta) = \frac{1}{2}\delta^\top H\delta + O(\|\delta\|^3),$$

where

$$H = \nabla_\delta^2 C(0).$$

Differentiating once more,

$$H_{ij} = \frac{\partial p_S(j)}{\partial z_i}.$$

Using the softmax derivative,

$$\frac{\partial p(j)}{\partial z_i} = p(j)(\mathbf{1}[i=j] - p(i)),$$

gives

$$H = \mathrm{diag}(p_S) - p_S p_S^\top.$$

Recognizing this as the Fisher information matrix (Amari & Nagaoka, 2000),

$$F = \mathrm{diag}(p_S) - p_S p_S^\top,$$

we obtain

$$D_{\mathrm{KL}}(p_S(\cdot; z) \,\|\, p_S(\cdot; z+\delta)) = \frac{1}{2}\delta^\top F\delta + O(\|\delta\|^3),$$

which is the local Fisher-metric constraint used in the main text.

A.2 Solution of the Fisher-Constrained Optimization

We derive the solution to

$$\max_\delta \; -g^\top\delta \qquad \text{subject to} \qquad \delta^\top F\delta \le 2\varepsilon.$$

Here F is the Fisher information matrix of the categorical student distribution. The derivation follows the classical Fisher-constrained steepest-descent argument of natural gradient methods (Kakade, 2001). Since the softmax is invariant to constant logit shifts, F is positive semi-definite rather than positive definite. Its null space contains the all-ones vector $\mathbf{1}$. We therefore restrict the optimization to the simplex tangent space

$$\mathcal{T} = \left\{\delta \in \mathbb{R}^V : \mathbf{1}^\top\delta = 0\right\},$$

on which the Fisher is positive definite when the student distribution has full support. Equivalently, its inverse action in the ambient logit space is represented by the Moore–Penrose pseudoinverse $F^\dagger$.

Because the objective is linear and the feasible set is convex, a nonzero optimum lies on the boundary:

$$\delta^\top F\delta = 2\varepsilon.$$

Introduce a Lagrange multiplier λ and define

$$\mathcal{J}(\delta, \lambda) = -g^\top\delta - \lambda\left(\delta^\top F\delta - 2\varepsilon\right).$$

The sign convention for λ is arbitrary; we choose it so that the resulting optimizer points in the descent direction. The first-order stationarity condition with respect to δ is

$$\nabla_\delta \mathcal{J} = -g - 2\lambda F\delta = 0.$$

Hence,

$$F\delta = -\frac{1}{2\lambda} g.$$

Since the gradient of the categorical KL satisfies $\mathbf{1}^\top g = 0$, it lies in the range of F. Applying the pseudoinverse gives

$$\delta = -\frac{1}{2\lambda} F^\dagger g + u,$$

where u lies in the null space of F. A null-space component changes all logits by a common shift and therefore changes neither the softmax distribution nor the objective. We select the minimum-norm representative by taking $u = 0$:

$$\delta = -\frac{1}{2\lambda} F^\dagger g. \tag{12}$$

Substituting Equation (12) into the active constraint gives

$$\begin{aligned} 2\varepsilon &= \delta^\top F \delta \\ &= \frac{1}{4\lambda^2} g^\top F^\dagger F F^\dagger g. \end{aligned}$$

Because g lies in the range of F,

$$F F^\dagger g = g,$$

and therefore

$$2\varepsilon = \frac{1}{4\lambda^2} g^\top F^\dagger g.$$

Solving for the positive multiplier magnitude,

$$\lambda = \frac{1}{2}\sqrt{\frac{g^\top F^\dagger g}{2\varepsilon}}.$$

Substituting back into Equation (12) yields

$$\boxed{\delta^\star = -\sqrt{\frac{2\varepsilon}{g^\top F^\dagger g}} F^\dagger g}.$$

Thus, the optimal perturbation points in the natural-gradient direction $-F^\dagger g$, with its scale chosen so that it exhausts the permitted local distributional budget.

Finally, its first-order improvement is

$$\begin{aligned} -g^\top \delta^\star &= \sqrt{\frac{2\varepsilon}{g^\top F^\dagger g}}\, g^\top F^\dagger g \\ &= \boxed{\sqrt{2\varepsilon\, g^\top F^\dagger g}}. \end{aligned}$$

For a common budget ε across token positions, the factor 2ε and the square root are common monotone transformations. Consequently, ranking tokens by maximum local first-order improvement is equivalent to ranking them by

$$g^\top F^\dagger g.$$

A.3 Closed Form for Reverse-KL Token Utility

For a fixed student-visited prefix, let

$$p_i = p_S(i), \qquad q_i = p_T(i),$$

and define

$$r_i = \log \frac{p_i}{q_i}.$$

The reverse-KL objective is

$$L = D_{\mathrm{KL}}(p\|q) = \sum_i p_i \log \frac{p_i}{q_i} = \sum_i p_i r_i.$$

We first derive its gradient with respect to the student logits z. For the softmax distribution,

$$\frac{\partial p_i}{\partial z_j} = p_i(\mathbf{1}[i=j] - p_j).$$

Moreover,

$$\frac{\partial}{\partial p_i}\left(p_i \log \frac{p_i}{q_i}\right) = \log \frac{p_i}{q_i} + 1 = r_i + 1.$$

Applying the chain rule,

$$\begin{aligned} g_j = \frac{\partial L}{\partial z_j} &= \sum_i (r_i+1)\frac{\partial p_i}{\partial z_j} \\ &= \sum_i (r_i+1)p_i(\mathbf{1}[i=j] - p_j) \\ &= p_j(r_j+1) - p_j \sum_i p_i(r_i+1). \end{aligned}$$

Since $\sum_i p_i = 1$,

$$\sum_i p_i(r_i+1) = \sum_i p_i r_i + 1.$$

Therefore,

$$\begin{aligned} g_j &= p_j\left(r_j - \sum_i p_i r_i\right) \\ &= p_j\left(r_j - \mathbb{E}_{i\sim p}[r_i]\right). \end{aligned}$$

Now recall that the categorical Fisher information matrix is

$$F = \mathrm{diag}(p) - pp^\top.$$

Multiplying F by r gives

$$(Fr)_j = p_j r_j - p_j \sum_i p_i r_i,$$

which is exactly the expression above for g_j. Hence,

$$\boxed{g = Fr.}$$

The Fisher-constrained token utility derived in the main text is

$$s = g^\top F^\dagger g.$$

Substituting $g = Fr$,

$$\begin{aligned} s &= (Fr)^\top F^\dagger (Fr) \\ &= r^\top F F^\dagger F r \\ &= r^\top F r, \end{aligned}$$

where we use the Moore–Penrose identity

$$FF^\dagger F = F.$$

Expanding the categorical Fisher,

$$\begin{aligned} r^\top F r &= r^\top \left(\mathrm{diag}(p) - pp^\top\right) r \\ &= \sum_i p_i r_i^2 - \left(\sum_i p_i r_i\right)^2 \\ &= \mathbb{E}_{i\sim p}[r_i^2] - \mathbb{E}_{i\sim p}[r_i]^2 \\ &= \mathrm{Var}_{i\sim p}[r_i]. \end{aligned}$$

Substituting $r_i = \log(p_S(i)/p_T(i))$ therefore yields

$$\boxed{s = \mathrm{Var}_{i\sim p_S}\left[\log \frac{p_S(i)}{p_T(i)}\right].}$$

A.4 Closed Form for Forward-KL Token Utility

The local Fisher geometry derived in Appendix A.1 depends only on the constraint on changes to the student distribution. Changing the distillation objective therefore changes the utility through its gradient g, while the Fisher metric remains the same.

Let $p = p_S$ and $q = p_T$. For the forward-KL objective,

$$L_{\mathrm{FKL}} = D_{\mathrm{KL}}(q\|p) = \sum_i q_i \log \frac{q_i}{p_i}.$$

Differentiating with respect to the student logits gives

$$g = p - q.$$

Since p and q are probability distributions,

$$\mathbf{1}^\top g = 0.$$

From Appendix A.2, the local token utility is

$$s_{\mathrm{FKL}} = g^\top F^\dagger g, \qquad F = \mathrm{diag}(p) - pp^\top.$$

Let $D = \mathrm{diag}(p)$. Since $p^\top D^{-1} = \mathbf{1}^\top$,

$$FD^{-1}g = (D - pp^\top)D^{-1}g = g - p(\mathbf{1}^\top g) = g.$$

Thus $D^{-1}g$ is a solution to the Fisher system $Fv = g$. Any two solutions differ only by a null-space component of F. As established in Appendix A.2, this component is a constant-logit shift and is orthogonal to g. Therefore,

$$g^\top F^\dagger g = g^\top D^{-1} g = \sum_i \frac{g_i^2}{p_i}.$$

Substituting $g_i = p_i - q_i$ gives

$$\boxed{s_{\mathrm{FKL}} = \sum_i \frac{(q_i - p_i)^2}{p_i} = \chi^2(q\|p)}.$$

Thus, under the same local predictive-distribution geometry, reverse-KL and forward-KL induce different token utilities: reverse-KL yields the student-weighted variance of the log-probability ratio, whereas forward KL yields the Pearson χ^2 divergence.

A.5 Exact Finite-Radius Utility for Reverse-KL

Relation to the local Fisher utility. The preceding derivations define token utility through a local quadratic approximation to the predictive-distribution KL constraint. We now ask whether this quantity also corresponds to an exact distribution-space optimization problem before taking the local

approximation. For reverse-KL, we solve the finite-radius problem exactly and show that the Fisher utility is recovered as its normalized infinitesimal limit.

At one student-visited prefix, write

$$p_i = p_S(i; z), \qquad q_i = p_T(i), \qquad x_i(\delta) = p_S(i; z + \delta),$$

and assume $p_i, q_i > 0$. Since softmax can represent every distribution in the interior of the simplex, optimizing over δ up to a common logit shift is equivalent to

$$\min_x \quad D_{\mathrm{KL}}(x\|q) \qquad \text{subject to} \qquad D_{\mathrm{KL}}(p\|x) \le \varepsilon, \quad \sum_i x_i = 1.$$

The objective and feasible set are convex in x, so the optimizer is unique. If $D_{\mathrm{KL}}(p\|q) \le \varepsilon$, then $x^\star = q$. Otherwise, the constraint is active.

For the active case, introduce multipliers $\lambda > 0$ and ν:

$$\mathcal{J}(x, \lambda, \nu) = \sum_i x_i \log \frac{x_i}{q_i} + \lambda \left(\sum_i p_i \log \frac{p_i}{x_i} - \varepsilon \right) + \nu \left(\sum_i x_i - 1 \right).$$

Stationarity with respect to x_i gives

$$\log \frac{x_i}{q_i} + 1 + \nu - \lambda \frac{p_i}{x_i} = 0. \tag{13}$$

Let $u_i = \lambda p_i / x_i$ and $\alpha = \lambda e^{1+\nu}$. Equation (13) then implies

$$u_i e^{u_i} = \alpha \frac{p_i}{q_i}.$$

Using the principal Lambert function W_0, defined by $W_0(y)e^{W_0(y)} = y$, yields

$$u_i = W_0\left(\alpha \frac{p_i}{q_i} \right).$$

Normalizing x therefore gives

$$\boxed{x_i^\star(\alpha) = \frac{p_i / W_0(\alpha p_i / q_i)}{\sum_j p_j / W_0(\alpha p_j / q_j)}}. \tag{14}$$

The scalar $\alpha > 0$ is chosen so that

$$D_{\mathrm{KL}}\left(p\|x^\star(\alpha)\right) = \varepsilon. \tag{15}$$

As $\alpha \downarrow 0$, $x^\star(\alpha) \to q$, while as $\alpha \uparrow \infty$, $x^\star(\alpha) \to p$. Thus Equation (15) is a one-dimensional root problem for $0 < \varepsilon < D_{\mathrm{KL}}(p\|q)$.

Because $x_i \propto p_i e^{\delta_i}$, an exact optimal logit perturbation is

$$\boxed{\delta_i^\star = -\log W_0\left(\alpha \frac{p_i}{q_i} \right) + c},$$

where the common shift c does not affect the softmax distribution.

The exact finite-radius improvement is

$$U(\varepsilon) := D_{\mathrm{KL}}(p\|q) - D_{\mathrm{KL}}(x^\star\|q).$$

To connect it to the local utility, let $r_i = \log(p_i/q_i)$. As $\varepsilon \downarrow 0$, the exact optimizer satisfies $x^\star = p + O(\sqrt{\varepsilon})$. The objective and constraint expansions in logit space give

$$D_{\mathrm{KL}}(p\|x(\delta)) = \frac{1}{2}\delta^\top F \delta + O(\|\delta\|^3),$$
$$D_{\mathrm{KL}}(x(\delta)\|q) = D_{\mathrm{KL}}(p\|q) + g^\top \delta + O(\|\delta\|^2).$$

Applying the Fisher-constrained solution and using $g^\top F^\dagger g = \mathrm{Var}_{i\sim p}[r_i]$ yields

$$U(\varepsilon) = \sqrt{2\varepsilon \ \mathrm{Var}_{i\sim p}\left[\log \frac{p_i}{q_i}\right]} + O(\varepsilon).$$

Consequently,

$$\boxed{s = \lim_{\varepsilon\downarrow 0} \frac{U(\varepsilon)^2}{2\varepsilon} = \mathrm{Var}_{i\sim p}\left[\log \frac{p_i}{q_i}\right]}.$$

Thus, s is not merely a quantity produced by the quadratic Fisher approximation. It is the radius-independent coefficient governing the leading-order improvement of the exact constrained problem. At finite ϵ, however, the exact utility $U(\epsilon)$ depends on the chosen radius and need not preserve the ranking induced by s.

B Additional Experimental Studies

This appendix collects the controlled studies that complement the main end-to-end comparison. Unless stated otherwise, the student, teacher, data, training budget, generation settings, and paired bootstrap procedure match Section 5.

B.1 Reverse-KL Evaluation Results Across Model Pairs

Evaluation. We sample at temperature 0.7 and top-$p = 0.95$. For each problem, we generate $K_b = 16$ responses on AIME24, AIME25, AMC23, and BRUMO25, and $K_b = 8$ responses on MATH-500, Minerva, and OlympiadBench. The accuracy of problem i in benchmark b is the fraction of its K_b responses that are correct, denoted $a_{b,i}$. We report

$$\text{Macro} = \frac{1}{7}\sum_{b=1}^{7}\frac{1}{n_b}\sum_{i=1}^{n_b} a_{b,i}, \qquad \text{Pooled} = \frac{1}{1620}\sum_{b=1}^{7}\sum_{i=1}^{n_b} a_{b,i}, \tag{16}$$

where n_b is the number of problems in benchmark b. Thus, Macro weights benchmarks equally, while Pooled weights problems equally regardless of their number of sampled responses. We estimate aggregate uncertainty and paired contrasts using 20,000 paired bootstrap resamples over problems.

Baselines. Uniform OPD uses the same candidate-set construction, renormalized reverse-KL loss, and training setup as FUW, with $w_t = 1$ at every position. Each candidate set is the union of the student's top-16 and teacher's top-16 tokens. On average we observe the union consists of about 19 tokens. We also compare against TIDE using its native support and TIP Soft-OR at 50% retention using FUW's candidate-set construction.

End-to-end mathematical reasoning results for the different model pairs are shown in Tables 1, 2 and 3. Macro averages benchmarks equally; Pooled averages all 1,620 problems. Best trained-student result in each column is **bold**.

Table 1: Qwen3-1.7B-Base distilled from Qwen3-8B

Method	AIME24	AIME25	AMC23	MATH-500	Minerva	Olymp.	BRUMO25	Avg@8 Macro	Avg@8 Pooled
Student (Base)	2.3±1.6	0.4±0.4	7.4±2.1	17.5±1.1	4.7±0.9	7.6±0.6	2.9±1.8	6.1±0.5	9.9±0.5
Teacher (8B)	30.6±7.4	20.6±7.1	62.3±4.5	83.9±1.4	28.4±2.5	47.7±1.7	25.8±7.3	42.8±1.9	55.1±1.1
OPD	8.8±4.4	5.4±3.2	32.0±3.9	61.6±1.7	16.8±1.7	25.6±1.4	9.4±4.8	22.8±1.2	34.5±1.0
TIDE	11.0±4.5	2.3±1.5	33.9±3.9	62.7±1.7	16.5±1.8	25.9±1.4	8.3±4.4	22.9±1.1	35.0±1.0
TIP (Soft-OR, 50%)	10.6±4.8	5.0±3.4	34.9±4.3	66.9±1.7	19.5±2.0	29.2±1.5	11.0±5.3	25.3±1.4	38.4±1.0
FUW (Ours)	**12.7±4.9**	**8.5±4.1**	**37.0±4.4**	67.5±1.7	**20.3±2.1**	**29.7±1.5**	11.9±5.1	**26.8±1.4**	**39.2±1.1**

B.2 Token-Importance Agreement Analysis

Token-importance criteria. Table 4 summarizes the criteria considered in our agreement analysis and their relationship to the reverse-KL Fisher utility. KL disagreement and the squared Euclidean

Table 2: Qwen3-4B-Base distilled from Qwen3-30B-A3B-Instruct-2507

Method	AIME24	AIME25	AMC23	MATH-500	Minerva	Olympiad	BRUMO25	Avg@8 Macro	Avg@8 Pooled
Student 4B-Base	4.4±2.6	0.0±0.0	17.4±3.2	21.1±1.0	5.8±0.9	15.1±0.9	8.8±3.6	10.4±0.8	14.9±0.6
Teacher 30B-A3B	75.2±6.7	65.4±7.4	91.6±2.7	95.8±0.8	35.2±2.7	67.6±1.7	66.0±6.6	71.0±1.8	72.2±1.0
OPD	29.8±7.5	28.7±7.6	60.5±4.7	84.4±1.3	33.3±2.5	47.6±1.7	29.6±7.4	44.8±2.0	56.2±1.1
TIDE	31.5±8.0	25.8±7.5	59.0±4.8	**86.6±1.3**	33.2±2.5	**50.3±1.7**	**38.8±7.4**	**46.4±2.0**	**58.0±1.1**
TIP	**32.3±7.7**	23.5±7.2	61.9±4.6	85.2±1.3	32.3±2.6	47.3±1.7	36.5±7.4	45.6±2.0	56.3±1.1
FUW (Ours)	26.7±7.2	**32.3±7.6**	**63.4±4.5**	85.6±1.4	**33.8±2.6**	49.2±1.8	32.1±7.5	46.15±1.98	57.54±1.01

Table 3: Qwen3-1.7B-Base distilled from Qwen3-14B

Method	AIME24	AIME25	AMC23	MATH-500	Minerva	Olympiad	BRUMO25	Avg@8 Macro	Avg@8 Pooled
Student (Qwen3-1.7B-Base)	2.3±1.6	0.4±0.4	7.4±2.1	17.5±1.1	4.7±0.9	7.6±0.6	2.9±1.8	6.1±0.5	9.9±0.5
Teacher (Qwen3-14B)	31.2±7.2	28.1±7.6	63.6±4.4	86.2±1.3	32.7±2.7	51.7±1.7	35.0±7.9	46.9±2.0	58.6±1.0
OPD	10.6±4.7	**7.3±3.8**	31.6±3.9	62.2±1.8	18.7±1.8	26.4±1.4	10.2±4.9	23.9±1.3	35.5±0.9
TIDE	10.0±4.8	5.0±3.4	34.4±4.1	65.6±1.7	18.8±1.9	28.3±1.4	9.4±4.4	24.5±1.3	37.4±0.9
TIP	10.4±4.8	6.7±3.2	34.1±4.2	66.6±1.7	20.1±2.0	29.3±1.5	**15.6±6.2**	26.1±1.4	38.5±0.9
FUW (Ours)	**11.9±5.1**	7.1±3.9	**36.3±4.3**	**69.4±1.7**	**22.2±2.0**	**30.2±1.5**	11.9±4.7	**27.0±1.3**	**40.2±0.9**

gradient norm are analytically related to the same log-ratio underlying FUW, whereas student entropy, TIP Soft-OR, and TIDE characterize different properties of the position.

Criterion	Score or quantity	Relationship to Fisher utility
FUW (ours)	$\mathrm{Var}_p[r]$	Maximum first-order reverse-KL reduction, normalized for an equal infinitesimal change in the predictive distribution.
KL disagreement	$D_{\mathrm{KL}}(p\|q) = \mathbb{E}_p[r]$	Mean of the log-ratio whose centered variance defines Fisher utility. Similar KL values need not imply identical Fisher utilities.
Squared Euclidean gradient norm	$\|g\|_2^2 = \sum_i p_i^2 (r_i - \bar{r})^2$	Uses the same centered log-ratio deviations, with an additional factor of p_i that relatively favors high-probability student outcomes.
Student entropy	$H(p) = -\mathbb{E}_p[\log p_i]$	Measures student uncertainty without using the teacher distribution.
TIP Soft-OR	$\widehat{H} + \widehat{D} - \widehat{H}\widehat{D}$	Combines normalized uncertainty and disagreement scores.
TIDE	Excess- and deficit-mismatch scores	Distinguishes the two directions of student–teacher probability mismatch.

Table 4: Token-importance criteria used in the agreement analysis. Here p and q denote the student and teacher next-token distributions at a fixed prefix, $r_i = \log(p_i/q_i)$, and $\bar{r} = \mathbb{E}_{i\sim p}[r_i]$. The first three rows have the stated analytical relationships; the remaining criteria characterize different properties of the position.

Agreement analysis. We compute all criteria on checkpoints obtained during uniformly weighted reverse-KL OPD of Qwen3-1.7B-Base ← Qwen3-14B. Since none of the importance criteria determines this training trajectory, it provides a controlled setting for analyzing criterion agreement.

The main-text analysis summarizes the two directions using bidirectional weight retention, $B_\rho(X) = \min\{\mathrm{REV}_\rho(X), \mathrm{CONT}_\rho(X)\}$. Here, we decompose B_ρ into its directional components to determine whether high effective-support agreement is symmetric or instead reflects one-sided containment. Importantly, high bidirectional weight retention measures agreement in effective high-importance support, not equality of the underlying weight vectors. Two criteria can attain $B_\rho \approx 1$ while assigning different relative weights to individual positions within that shared support. This distinction is reflected in the substantial differences in score concentration across criteria shown in Table 5.

Let S_X^ρ denote the top-ρ fraction of positions according to criterion X, and let $w_{X,i}$ denote the corresponding importance weight assigned to position i. As defined in Section 5.2

$$\mathrm{REV}_\rho(X) = \frac{\sum_{i\in S_X^\rho} w_{\mathrm{FUW},i}}{\sum_{i\in S_{\mathrm{FUW}}^\rho} w_{\mathrm{FUW},i}},$$

Table 5: Fraction of positions needed to account for a given fraction of score mass.

Score	50% mass	90% mass	95% mass
FUW	0.077	0.328	0.462
Euclidean $\|g\|_2^2$	0.035	0.121	0.168
Reverse-KL divergence	0.038	0.231	0.345
TIP Soft-OR	0.055	0.317	0.474
Uniform	0.500	0.900	0.950

measures how much of FUW's attainable top-ρ weight is retained when positions are instead selected by X, and

$$\mathrm{CONT}_\rho(X) = \frac{\sum_{i \in S^\rho_{\mathrm{FUW}}} w_{X,i}}{\sum_{i \in S^\rho_X} w_{X,i}},$$

measures the converse. Their difference,

$$A_\rho(X) = \mathrm{CONT}_\rho(X) - \mathrm{REV}_\rho(X),$$

characterizes directional asymmetry. Positive values indicate that FUW's selected set captures more of X's attainable weight than X's selected set captures of FUW's; negative values indicate the reverse. Values near zero indicate approximately symmetric retention.

Random selection controls. To distinguish criterion agreement from retention expected under random selection, we use a permutation control that preserves the weight multiset while destroying the correspondence between positions and scores. Let

$$D_\rho = \frac{\sum_{i \in S^\rho_{\mathrm{FUW}}} w_{\mathrm{FUW},i}}{\sum_i w_{\mathrm{FUW},i}}$$

denote the fraction of FUW's total weight contained in its top-ρ positions. Since each position is included in a uniformly random top-ρ set with probability ρ,

$$\mathbb{E}[\mathrm{REV}_{\rho,\mathrm{random}}] = \frac{\rho}{D_\rho}.$$

Thus, the control accounts for retention induced by the selection fraction and the concentration of FUW's weight distribution. The measured permutation control agrees with this analytic expectation to within 0.003 at all 13 checkpoints.

Figures 3 and 4 report the decomposition of B_ρ at $\rho = 0.2$ and $\rho = 0.5$, corresponding to the 20% and 50% retention fractions. The former is close to TIDE's realized active fraction in our implementation, while the latter matches the 50% Soft-OR configuration recommended by TIP.

At $\rho = 0.2$, the two directions can differ materially for student entropy and TIP Soft-OR, with asymmetry reaching approximately 0.18 over the latter half of training. These differences coincide with changes in the relative concentration of the criteria's weights, so we do not interpret them as evidence of directional containment.

At $\rho = 0.5$, Euclidean gradient magnitude, student entropy, KL disagreement, TIP Soft-OR, and TIDE-deficit exhibit nearly symmetric agreement with FUW: their two directional retention scores remain close throughout the later checkpoints. Their near-unit $B_{0.5}$ values therefore reflect high retention in both directions. TIDE-excess exhibits a qualitatively different pattern. Its asymmetry is large and positive, indicating that FUW's top-half set captures substantially more of TIDE-excess's attainable weight than the TIDE-excess set captures of FUW's. By the later checkpoints, FUW's selected set captures nearly all of the high-weight support identified by TIDE-excess, while the converse retention is substantially lower. Thus, unlike the near-symmetric agreement observed for the other criteria, the lower $B_{0.5}$ of TIDE-excess reflects asymmetric effective-support containment. This behavior is consistent with the distinct role of TIDE-excess, which explicitly isolates one direction of student–teacher mismatch rather than measuring scalar disagreement, uncertainty, or local optimization opportunity.

Overall, for the Qwen3-1.7B-Base ← Qwen3-14B pair studied under uniform OPD, several independently motivated criteria concentrate importance on largely the same effective support as the

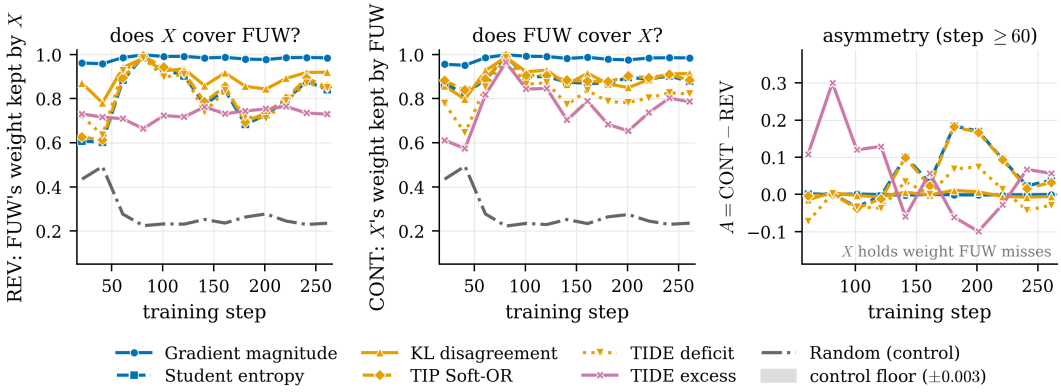

Figure 3: **Directional decomposition at $\rho = 0.2$.** REV measures the fraction of FUW's attainable top-ρ weight retained when positions are selected by criterion X, while CONT measures the converse. Directional agreement is more sensitive to differences in weight concentration at this operating fraction.

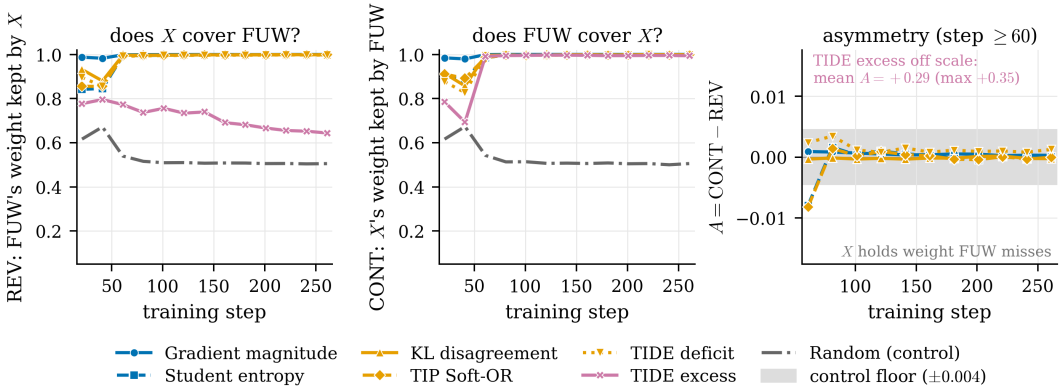

Figure 4: **Directional decomposition at $\rho = 0.5$.** Most criteria exhibit nearly symmetric, near-complete retention with FUW, whereas TIDE-excess exhibits asymmetric effective-support containment.

Fisher-derived utility at $\rho = 0.5$. For gradient magnitude, student entropy, KL disagreement, TIP Soft-OR, and TIDE-deficit, this relationship is nearly symmetric, indicating shared effective high-importance support rather than one-sided containment. TIDE-excess is the notable exception: its high-importance support is largely contained within FUW's broader support, while the converse does not hold. These results concern effective support rather than point-wise equivalence of the weighting rules.

Cross-pair robustness. We additionally evaluate bidirectional weight retention on a second student–teacher pair using checkpoints generated during FUW training. Unlike the uniformly weighted trajectory above, these checkpoints are not independent of the reference criterion, since FUW itself determines the training trajectory. We therefore treat this experiment as a cross-pair robustness check rather than an independent test of criterion agreement.

The similar qualitative structure observed earlier is preserved across both operating fractions as shown in Figure 5. At $\rho = 0.5$, gradient magnitude, student entropy, KL disagreement, TIP Soft-OR, and TIDE-deficit exhibit near-complete bidirectional retention with FUW throughout most of training, whereas TIDE-excess remains substantially lower. At $\rho = 0.2$, the same separation emerges after the earliest checkpoints: the scalar criteria approach high effective-support agreement with FUW, while TIDE-excess remains distinctly separated. In both cases, agreement is substantially above the

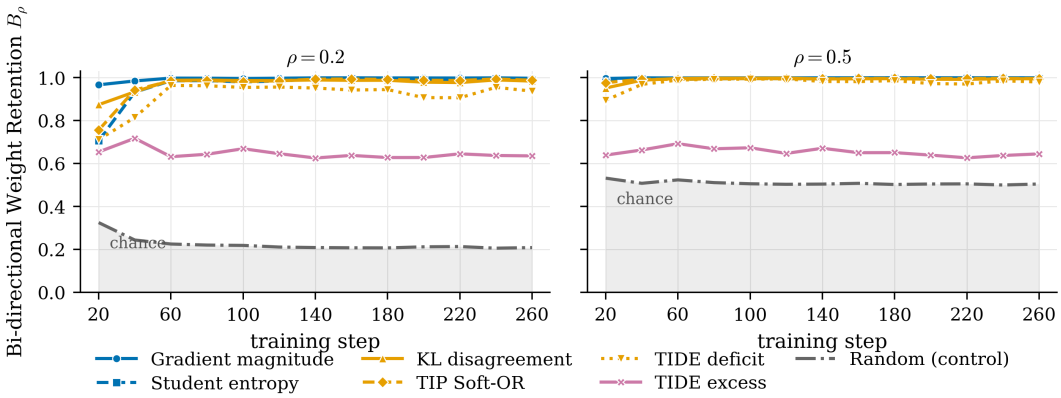

Figure 5: Cross-pair robustness of effective-support agreement. Bidirectional weight retention B_ρ with FUW on Qwen3-1.7B-Base ← Qwen3-8B along an FUW-trained trajectory, at $\rho = 0.2$ (left) and $\rho = 0.5$ (right). Beyond the early training steps, most criteria show near-complete effective-support agreement with FUW at both retention fractions, while TIDE-excess remains distinctly lower.

random-selection control. Thus, the effective-support structure observed under uniform OPD for the primary model pair is qualitatively reproduced on a different model pair and training trajectory.

B.3 Weighting, metric, and ranking controls

All controls in this subsection use the same re-normalized union of the student's and teacher's top-16 candidates obtained during the reverse-KL distillation of the Qwen3-1.7B-Base ← Qwen3-8B pair. They therefore isolate the scoring or weighting rule without changing candidate support. Table 6 reports paired pooled-score differences relative to FUW. A negative value favors FUW.

Table 6: Paired controls on the union top-16 support for reverse-KL OPD of Qwen3-1.7B-Base ← Qwen3-8B pair. The Euclidean variant replaces the Fisher utility by $\|g_t\|_2^2$; the permutation control randomly reassigns FUW weights within each batch. Negative Δ favors FUW.

Method	**Pooled Avg@8**	**Δ vs. FUW**	**95% CI**	***p***
FUW	39.17	–	–	–
Euclidean utility	38.9	-0.27	$[-1.09,\ 0.56]$	0.527
Permuted FUW	38.15	-1.02	$[-1.92,\ -0.11]$	**0.028**
Hard top-half FUW	38.41	-0.76	$[-1.62,\ 0.11]$	0.087

The Euclidean variant reaches pooled Avg@8 of 38.9, close to FUW's 39.2, with no significant paired difference ($p = 0.527$). However, its weighting dynamics differ substantially from FUW. Effective retention decreases from 0.327 early in training to 0.104 by step 118, while the maximum token weight rises from 29 to 96. The additional factor of p_S implicit in the Euclidean score increasingly concentrates weight on high-probability student outcomes. Fisher normalization avoids this late-stage concentration, although the two metrics are not separated by downstream accuracy at the present training budget.

The permutation control preserves the complete multiset of FUW weights and changes only their assignment to positions. Its significant reduction (-1.02, $p = 0.028$) shows that position-specific utility information contributes beyond the marginal weight distribution: the positions FUW identifies are not interchangeable. Hard selection does the opposite, preserving which positions FUW identifies as high utility while discarding their relative weights. Used as a top-half mask, the FUW score reaches 38.4 pooled Avg@8, matching TIP Soft-OR at the same 50% retention fraction (38.4). This matches the effective-support analysis in Section 5.2: the two criteria identify nearly the same support at $\rho = 0.5$. Continuous FUW reaches 39.2; on this pair, its difference from the hard mask is not significant ($p = 0.087$).

B.4 FORWARD-KL EVALUATION RESULTS ACROSS MODEL PAIRS

Table 7 reports the per-benchmark results for the forward-KL experiments summarized in Section 5.4.

Method	AIME24	AIME25	AMC23	MATH	Minerva	Olymp.	BRUMO	Avg@k Macro	Avg@k Pooled
Qwen3-1.7B-Base ← Qwen3-8B									
FKL (Uniform)	4.6±3.2	2.1±1.5	25.7±3.7	56.9±1.7	19.4±1.9	22.4±1.3	**9.4±3.1**	20.07±0.94	31.78±0.83
FKL + Utility (Ours)	**12.1±4.5**	**4.4±3.0**	**28.3±4.0**	**63.6±1.8**	**20.4±2.1**	**24.7±1.3**	6.0±3.0	**22.78±1.14**	**35.21±0.89**
Qwen3-1.7B-Base ← Qwen3-14B									
FKL (Uniform)	3.8±1.6	0.5±0.3	**29.4±3.8**	57.7±1.7	**20.0±1.9**	22.4±1.3	**9.6±3.9**	20.46±0.91	32.23±0.84
FKL + Utility (Ours)	**6.9±3.7**	**2.7±1.7**	27.8±3.9	**63.0±1.7**	19.1±2.0	**24.0±1.3**	7.3±3.5	**21.54±1.02**	**34.41±0.86**
Qwen3-4B-Base ← Qwen3-30B-A3B-Instruct-2507									
FKL (Uniform)	**24.0±5.3**	21.0±5.7	**60.1±5.2**	84.0±1.5	32.0±2.7	48.5±1.8	24.8±7.4	42.04±1.7	55.85±1.2
FKL + Utility (Ours)	22.9±5.5	**22.9±5.9**	57.0±5.1	**85.5±1.6**	**32.7±2.9**	**47.6±1.9**	**33.5±7.2**	**43.2±1.8**	**56.1±1.2**

Table 7: Forward-KL mathematical-reasoning performance across three student–teacher pairs. Benchmark columns report Avg@k, while aggregate columns report macro- and pooled-average Avg@k. Values are reported as mean ± standard error. The best result within each model pair is shown in bold.