OC-OPD: OBSERVATION-COORDINATED SUPERVISION FOR ON-POLICY DISTILLATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Reverse and forward Kullback–Leibler (KL) divergences provide complementary supervision in on-policy distillation: reverse KL refines the student's active token rankings, and forward KL allocates mass to teacher-supported candidates. At student-generated prefixes, we form a four-observation matrix by treating the student and teacher as both candidate sources and evaluators. The diagonal self-observations measure uncertainty, while source-centered cross-observations recover the two KL objectives. We establish an exact identity equating the Fisher inverse-metric coupling of their logit gradients with the sum of these contrasts, yielding a common local descent direction for distinct positive distributions. Observation-coordinated on-policy distillation (OC-OPD) uses observations to set preferences, closed-form bargaining to coordinate gradients, and distributional discrepancy to scale the update. Across three heterogeneous student settings (non-thinking generation, thinking generation, and base-model initialization), OC-OPD achieves strong gains in mathematical reasoning and transfers effectively to out-of-distribution scientific question answering and code generation.

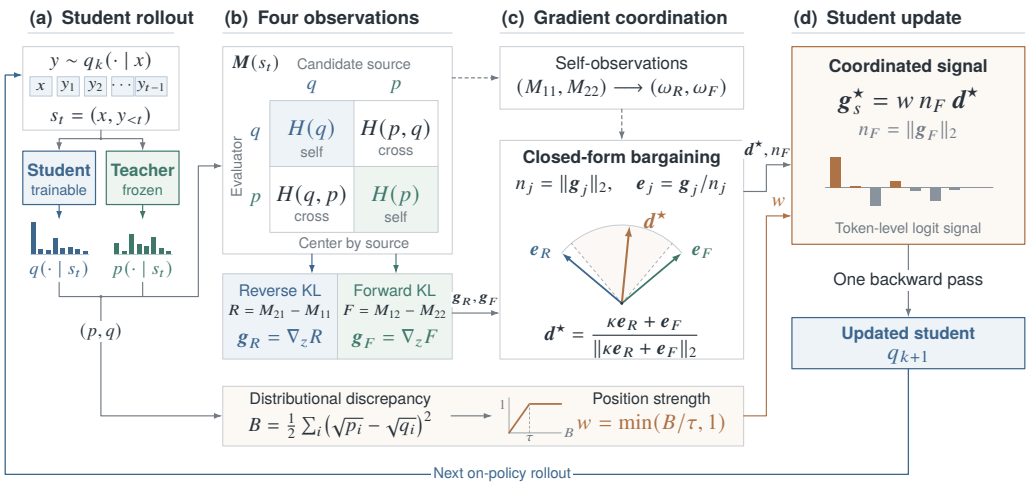


Figure 1: **Training pipeline of OC-OPD. (a)** Student rollouts provide prefixes at which the student and frozen teacher produce aligned next-token distributions. **(b)** Four self- and cross-observations at the shared prefix yield source-centered KL gradients for ranking/refinement and coverage. **(c)** Self-observation values set uncertainty-based priorities for closed-form gradient bargaining, while distributional discrepancy controls supervision strength. **(d)** The resulting bounded logit signal updates the student through a single backward pass. The updated student then generates new rollouts.

1 INTRODUCTION

On-policy distillation (OPD) trains a student at its own generated prefixes, where a frozen teacher provides supervision for the same next-token decision (Ross et al., 2011; Agarwal et al., 2024a). At each prefix, the aligned student distribution q and teacher distribution p define two complementary

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.

objectives: reverse KL, $D_{\mathrm{KL}}(q\|p)$, refines the relative allocation among student-supported candidates, while forward KL, $D_{\mathrm{KL}}(p\|q)$, moves probability toward teacher-supported candidates even when the student assigns them little mass (Agarwal et al., 2024a; Gu et al., 2024). How should these corrections be coordinated within one update at the same student-visited state?

Using either KL direction alone emphasizes only one role. Recent OPD methods also adapt supervision using uncertainty, token importance, teacher–student mismatch, and divergence routing (Ko et al., 2026; Jin et al., 2026; Lee et al., 2026; Xu et al., 2026; Yu et al., 2026). These signals indicate where or how to supervise, but coordinating the two corrections additionally requires knowing whether their gradients admit joint improvement and how their relative influence should respond to the current predictive state.

We connect state information to candidate-level updates through a source–evaluator view of p and q. Each distribution can supply candidate mass, and either model can evaluate that mass, yielding a 2×2 system of self- and cross-observations. Its diagonal entries are student and teacher entropies; its off-diagonal entries are cross-entropies. Centering each candidate source's column by its self-observation recovers $D_{\mathrm{KL}}(q\|p)$ from the student-sourced column and $D_{\mathrm{KL}}(p\|q)$ from the teacher-sourced column. The four values describe uncertainty and cross-model evaluation, while the contrasts connect those values to the refinement and coverage gradients. The observation system organizes the existing distributions to expose this connection and guide their joint update.

The resulting gradients have a stronger compatibility property than a generic pair of objectives. Let g_R and g_F be the logit gradients of reverse and forward KL, and let $J(q)$ be the categorical Fisher matrix. For positive p and q, we establish the exact identity $g_R^\top J(q)^+ g_F = D_{\mathrm{KL}}(q\|p) + D_{\mathrm{KL}}(p\|q)$. The right-hand side is the total excess of cross-observations over self-observations and is positive whenever $p \neq q$. Thus the gradients cannot be exactly opposed: a sufficiently small step in a common descent direction locally decreases both KL objectives. This identity supplies a geometric basis for coordinating refinement and coverage at a shared prefix.

We develop observation-coordinated on-policy distillation (OC-OPD) around three decisions: *preference*, *direction*, and *strength*. The diagonal self-observations set state-dependent preferences between the two corrections. A closed-form bargaining rule coordinates their normalized gradients within the common-descent region. A symmetric distributional discrepancy then scales the position's supervision strength, attenuating updates near teacher–student agreement. Together, these components yield a bounded token-level logit signal applied with one student backward pass. Figure 1 summarizes the observation $\rightarrow$ contrast $\rightarrow$ coordination $\rightarrow$ strength pipeline.

We evaluate OC-OPD on mathematical reasoning in three Qwen3 student settings (Qwen Team, 2025a), all trained with fresh student rollouts: non-thinking students at 1.7B and 4B scales, thinking-generation students, and a weaker base-model initialization. In the main non-thinking setting, the 1.7B student exceeds the strongest compared baseline by $+7.02$ Avg@32 and $+13.34$ Pass@32 points on average across AIME24–26; the 4B student exceeds EOPD (Jin et al., 2026) by $+2.74$ and $+3.33$ points, respectively (Table 1). Gains persist in the other two settings, and GPQA-Diamond (Rein et al., 2023) and MBPP+ (Liu et al., 2023) assess transfer to scientific question answering and code generation.

Our contributions are threefold. **(1)** We describe token-level OPD through a source–evaluator system whose self- and cross-observations characterize the teaching state and whose source-centered contrasts recover the two KL corrections. **(2)** We prove an exact Fisher inverse-metric identity linking those contrasts to their logit gradients and establish common local descent for every distinct positive teacher–student pair. **(3)** We develop OC-OPD to determine the preference, direction, and strength of token-level supervision, and demonstrate its effectiveness across heterogeneous student settings and out-of-domain evaluations.

2 Related Work

KL directions in on-policy distillation. Knowledge distillation transfers predictive structure from a teacher to a student (Hinton et al., 2015; Kim & Rush, 2016). OPD evaluates teacher guidance at student-generated prefixes (Ross et al., 2011; Agarwal et al., 2024a), where both models produce aligned next-token distributions. GKD studies flexible divergence objectives at these states (Agarwal et al., 2024a), and MiniLLM develops reverse-KL distillation under student sampling (Gu et al.,

2024). Related methods explore skewed divergences and adaptive sampling (Ko et al., 2024), contrastive teacher–student responses (Ko et al., 2025), and interleaved generation (Xu et al., 2025). Reverse KL refines relative probabilities among student-supported candidates, while forward KL transfers mass toward teacher-supported candidates. EOPD conditions forward-KL supervision on teacher entropy (Jin et al., 2026), and SEAD adjusts the emphasis of the two KL directions across tokens (Lee et al., 2026). Our formulation derives both directions from source-centered observations of the same distribution pair and characterizes their joint local gradient geometry.

Observation-based adaptive supervision. At a shared prefix, either model can supply candidate mass and either model can evaluate it. These choices produce four observations: two self-entropies describe the predictive states of the student and teacher, and two cross-entropies describe cross-model evaluation. Related methods use local information to determine how supervision is assigned. REOPOLD uses entropy-based token selection and relaxed reward shaping (Ko et al., 2026); SMOPD selects informative positions for self-distillation (Jiang & Huang, 2026); and TIP combines student uncertainty with divergence to assess token importance (Xu et al., 2026). EOPD and SEAD use entropy in their divergence choices (Jin et al., 2026; Lee et al., 2026), while TIDE characterizes student-excess and student-deficit mismatch (Yu et al., 2026). Region-dependent objectives, adaptive targets, bounded transformations, and trust-region control further shape local supervision (Jia et al., 2026; Jang et al., 2026; Zhao et al., 2026a; Xing et al., 2026). Our source–evaluator matrix places the distribution-derived signals in a shared coordinate system: the diagonal entries describe self-uncertainty, the off-diagonal entries describe cross-evaluation, and source-centering recovers the two KL corrections. Appendix G summarizes representative methods in this notation.

Gradient coordination. Multi-objective optimization constructs common update directions through minimum-norm descent (Sener & Koltun, 2018), gradient projection (Yu et al., 2020), conflict-averse optimization (Liu et al., 2021), and bargaining (Nash, 1950; Navon et al., 2022; Wu & Harandi, 2024). Here, both KL gradients arise from the aligned student–teacher distributions and their source-centered observation contrasts. An exact Fisher-geometric identity relates their local coupling to the sum of these contrasts (Amari, 1998; Nielsen, 2020). We use this relationship to coordinate candidate-level refinement and coverage, with bargaining yielding a closed-form direction conditioned on the observation state. Scalarization and normalized gradient mixing provide complementary comparison points (Kurin et al., 2022).

3 Preliminaries and Motivation

This section formalizes shared-context OPD, organizes the aligned student and teacher distributions into four observations, and shows how their source-centered contrasts recover the ranking and coverage signals of reverse and forward KL.

3.1 Shared-context OPD as observation system

OPD evaluates supervision along student-generated responses by querying the teacher at the same prefixes (Agarwal et al., 2024a; Gu et al., 2024). We distinguish this choice of training states from the information available at each visited state. In the shared-context setting, both models condition on the same prompt and student prefix, yielding an aligned pair of next-token predictive distributions. Our aim is to organize the information in this pair into effective supervision.

For a prompt x, let q_θ and p_T denote the student and frozen teacher. A representative reverse-KL OPD objective is

$$\begin{aligned} \mathcal{J}_{\mathrm{OPD}}(\theta; x) &= D_{\mathrm{KL}}\big(q_\theta(\cdot \mid x) \| p_T(\cdot \mid x)\big) \\ &= \mathbb{E}_{y \sim q_\theta(\cdot|x)} \left[\sum_{t=1}^{|y|} \log \frac{q_\theta(y_t \mid s_t)}{p_T(y_t \mid s_t)} \right], \qquad s_t = (x, y_{<t}). \end{aligned} \tag{1}$$

Student-sampled log ratios estimate this objective. At iteration k, write $q_k = q_{\theta_k}$ and collect valid response positions $\mathcal{S}_k$ from fresh student rollouts. A token-local surrogate assigns each sampled

token a teacher-to-student log-ratio credit:

$$\widehat{\mathcal{L}}_{\mathrm{OPD},k}(\theta) = -\frac{1}{N}\sum_{t\in\mathcal{S}_k} \mathrm{sg}[A_{k,t}] \log q_\theta(y_t \mid s_t), \qquad A_{k,t} = \log \frac{p_T(y_t \mid s_t)}{q_k(y_t \mid s_t)}, \tag{2}$$

where $N = |\mathcal{S}_k|$ and sg denotes stop-gradient. Prefixes and credits are fixed during differentiation. Full-vocabulary implementations evaluate the corresponding expectation over candidates at each collected prefix (Agarwal et al., 2024a). We use this objective as a reference and derive the four observations directly from the aligned distributions, making student-sourced and teacher-sourced supervision explicit.

3.2 Four observations from two distributions

At a collected prefix $s = (x, y_{<t})$, write $q = q_{\theta_k}(\cdot \mid s)$ and $p = p_T(\cdot \mid s)$ over an aligned vocabulary. The fixed pair p and q supplies two candidate sources—student and teacher—and two evaluators—student and teacher. Pairing each source with each evaluator yields four expected observations. For $H(a, b) = -\sum_i a_i \log b_i$ and $H(a) = H(a, a)$, define

$$M(s) = \begin{pmatrix} H(q) & H(p,q) \\ H(q,p) & H(p) \end{pmatrix} = -\sum_{i\in\mathcal{V}} \begin{pmatrix} q_i \log q_i & p_i \log q_i \\ q_i \log p_i & p_i \log p_i \end{pmatrix}. \tag{3}$$

Rows index evaluators and columns index candidate sources, both ordered as student, teacher. Accordingly, M_{11} records the student evaluating student-sourced mass, M_{12} the student evaluating teacher-sourced mass, M_{21} the teacher evaluating student-sourced mass, and M_{22} the teacher evaluating teacher-sourced mass. The diagonal self-observations are the student and teacher entropies and quantify distributional dispersion. The off-diagonal cross-observations measure how well one model accommodates candidates emphasized by the other.

The matrix provides a relational reparameterization of the same p, q pair by separating the distribution supplying candidate mass from the distribution scoring it. Teacher-sourced expectations use the teacher distribution at the same student-visited prefix. The matrix records state-level expectations, while the underlying p and q retain the candidate-level resolution needed for an update. It gathers common adaptive OPD signals in one object: student entropy is M_{11}, teacher entropy is M_{22}, and teacher–student mismatch appears as the off-diagonal excess over the corresponding self-observation. Table 6 in appendix G compares how different methods are described in this framework.

3.3 Source-centered contrasts reveal ranking and coverage

Centering each column on its candidate source's self-observation isolates the difference between the two evaluators:

$$\begin{aligned} R &= M_{21} - M_{11} = H(q,p) - H(q) = D_{\mathrm{KL}}(q\|p), \\ F &= M_{12} - M_{22} = H(p,q) - H(p) = D_{\mathrm{KL}}(p\|q). \end{aligned} \tag{4}$$

Thus reverse and forward KL arise as centered cross-observations from the student-sourced and teacher-sourced columns, respectively.

Student-sourced contrast: rank and refine active support. For reverse KL, the token-level credit $A_i = \log(p_i/q_i)$ compares teacher support with the student's current allocation. Since $A_i > A_j$ exactly when $p_i/q_i > p_j/q_j$, this credit ranks student-supported candidates by relative underallocation. The student-sourced contrast thereby refines the alternatives the student already explores. Its logit gradient is weighted by q_i, concentrating the correction on candidates that carry student probability.

Teacher-sourced contrast: recover underrepresented support. Forward KL evaluates teacher-sourced candidates under the student. Its logit gradient $q-p$ directly reallocates mass toward teacher-supported alternatives, including those with very small student probability. This coverage correction allocates student capacity across the teacher-supported set (Gu et al., 2024). In our full-vocabulary construction, the teacher-sourced gradient is evaluated directly at the visited prefix, yielding the exact candidate-level correction from the aligned distributions.

From complementary contrasts to a coordination problem. The matrix thus provides both statistics describing the current state and contrasts specifying candidate-level corrections. Existing entropy gates, token filters, and mismatch criteria can be interpreted as selecting or weighting particular regions of this object (Ko et al., 2026; Jin et al., 2026; Lee et al., 2026; Xu et al., 2026; Yu et al., 2026). Together, the source-centered contrasts invite a geometric account of when their ranking/refinement and coverage gradients share a descent direction. Section 4 establishes this compatibility; Section 5 uses self-observation, compatibility, and discrepancy to determine preference, direction, and magnitude.

4 Theory: From Observations to Compatible Gradients

The matrix view yields two source-centered corrections whose candidate-level gradients can be studied together. Their ranking/refinement and coverage directions admit common descent, with the observation matrix providing an explicit certificate of compatibility.

4.1 From observation entries to candidate-level gradients

For current-state coefficients a_s, b_s, define the local matrix loss

$$\begin{aligned}\ell_s(\theta) &= \left\langle \mathrm{sg}\left[\begin{pmatrix} -a_s & b_s \\ a_s & -b_s \end{pmatrix}\right], M_\theta(s) \right\rangle_F \\ &= \mathrm{sg}[a_s]\big[H(q_\theta, p) - H(q_\theta)\big] + \mathrm{sg}[b_s]\big[H(p, q_\theta) - H(p)\big].\end{aligned} \tag{5}$$

Writing the coefficient matrix as W_s, its columns preserve each candidate source's self-observation as a baseline. The resulting contrasts are $R = D_{\mathrm{KL}}(q\|p)$ and $F = D_{\mathrm{KL}}(p\|q)$; Appendix A.1 derives these identities from the observations.

Differentiating the matrix with respect to the current student logits yields

$$\begin{aligned} g_R &= \nabla_z(M_{21} - M_{11}) = q \odot (\log(q/p) - R\mathbf{1}), \\ g_F &= \nabla_z(M_{12} - M_{22}) = q - p, \\ \nabla_z \ell_s &= -a_s \nabla_z M_{11} + b_s \nabla_z M_{12} + a_s \nabla_z M_{21} - b_s \nabla_z M_{22} = a_s g_R + b_s g_F. \end{aligned} \tag{6}$$

The student self-observation and both cross-observations supply direct derivatives. The frozen teacher's self-observation has $\nabla_z M_{22} = 0$ and conditions the coefficients through its value. Thus all four entries contribute to the update. Section 5 specifies the coefficients; Appendix A.1 gives the entrywise derivatives and parameter update.

These derivatives make the complementary roles in Section 3.3 explicit. For a fixed positive teacher and $q_j \to 0$, their coordinates satisfy $g_{R,j} \to 0$ and $g_{F,j} \to -p_j$. The teacher-sourced contrast therefore retains a direct correction for candidates with very small student probability. Coordinating the two directions refines the student's current allocation while preserving a correction toward teacher-favored alternatives.

4.2 Compatibility certified by the observation contrast

Let $J(q) = \mathrm{diag}(q) - qq^\top$ be the categorical Fisher matrix(Amari, 1998; Nielsen, 2020), with Moore–Penrose pseudoinverse $J(q)^+$. The following identity relates the source-centered gradients through the observation matrix.

Theorem 1 (Observation–gradient identity)**.** *For positive p, q, the logit gradients in Eq.* (6) *satisfy*

$$g_R^\top J(q)^+ g_F = \sum_i (q_i - p_i) \log \frac{q_i}{p_i} = M_{12} + M_{21} - M_{11} - M_{22} = R + F. \tag{7}$$

If $p \neq q$, both gradients have positive norm and their Euclidean cosine is strictly greater than -1. A unit direction d therefore exists with $g_R^\top d > 0$ and $g_F^\top d > 0$.

The total observation excess certifies a common descent direction for both acute and obtuse Euclidean gradient configurations. The identity applies to every aligned pair in Section 3.2. Appendix B gives the proof, and Appendix C extends compatibility to a finite family of differentiable strictly convex f-divergences. We use observation values to select a direction within this feasible set and determine the update's magnitude. The appendix H.2 provides empirical analysis of the theory.

5 METHOD: OBSERVATION-COORDINATED UPDATES

At each student-visited prefix, OC-OPD follows an observation → contrast → coordination → strength pipeline. The aligned distributions first determine $M(s)$, and column-centering yields the ranking/refinement and coverage contrasts together with their candidate-level gradients. The diagonal self-observations specify preferences between these roles; the compatibility geometry then selects a direction that improves both normalized contrasts. Finally, a symmetric discrepancy between p and q scales the resulting update.

Algorithm 1 Observation-Coordinated On-Policy Distillation

Require: Frozen teacher p_T, student q_θ
Require: Entropy scale h_0, preference floor ϵ
Require: Discrepancy scale τ, learning rate η
Require: Training dataset $\mathcal{D}$, prompt batch size B

for each training iteration k **do**
  Set $q_k \leftarrow q_{\theta_k}$ and $\mathcal{R} \leftarrow \varnothing$; sample B prompts $\mathcal{B} \subset \mathcal{D}$.
  for each prompt $x \in \mathcal{B}$ **do**
    Sample $\mathbf{y} = (y_1, \ldots, y_{|\mathbf{y}|}) \sim q_k(\cdot \mid x)$.
    for $t = 1$ **to** $|\mathbf{y}|$ **do**
      $s_t \leftarrow (x, y_{<t})$, $q \leftarrow q_k(\cdot \mid s_t)$, $p \leftarrow p_T(\cdot \mid s_t)$.
      $M \leftarrow \begin{bmatrix} H(q) & H(p,q) \\ H(q,p) & H(p) \end{bmatrix}$, $R \leftarrow M_{21} - M_{11}$, $F \leftarrow M_{12} - M_{22}$.
      $g^\star_{s_t} \leftarrow 0$.
      if $p \neq q$ **then**
        $g_R \leftarrow q \odot \big(\log(q/p) - R\mathbf{1}\big)$, $g_F \leftarrow q - p$.
        Compute (ω_R, ω_F) from (M_{11}, M_{22}) using h_0, ϵ.
        Compute closed-form bargaining direction $d^\star$ from $g_R, g_F, \omega_R, \omega_F$.
        $w \leftarrow \min\big\{\frac{1}{2\tau}\sum_i (\sqrt{p_i} - \sqrt{q_i})^2,\, 1\big\}$, $g^\star_{s_t} \leftarrow w\|g_F\|_2 d^\star$.
    Store $(x, \mathbf{y}, \{g^\star_{s_t}\}_{t=1}^{|\mathbf{y}|})$ in $\mathcal{R}$.
  Collect valid prefixes $\mathcal{S}_k$ from $\mathcal{R}$.
  $L_k(\theta) \leftarrow \dfrac{1}{|\mathcal{S}_k|} \displaystyle\sum_{s \in \mathcal{S}_k} \sum_i \mathrm{sg}[g^\star_{s,i}] \log q_\theta(i \mid s)$.
  Update student to θ_{k+1} with one backward pass on L_k at learning rate η.

5.1 OBSERVATION-CONDITIONED GRADIENT COORDINATION

For $h_0, \epsilon > 0$, the diagonal observations determine normalized entropies and positive preferences:

$$\begin{aligned} h_S &= \frac{M_{11}}{M_{11} + h_0}, \qquad h_T = \frac{M_{22}}{M_{22} + h_0}, \\ \nu_R &= \epsilon + (1 - h_T) h_S, \qquad \nu_F = \epsilon + h_T, \qquad \omega_j = \frac{\nu_j}{\nu_R + \nu_F}, \quad j \in \{R, F\}. \end{aligned} \tag{8}$$

When the teacher is diffuse, this rule gives coverage greater influence; when the teacher is concentrated and the student diffuse, it gives ranking/refinement greater influence. Both sources retain positive preference (Appendix A.2).

For $p \neq q$, set $n_j = \|g_j\|_2$, $e_j = g_j / n_j$, and $c = e_R^\top e_F$. Gradient bargaining (Navon et al., 2022) selects a direction by maximizing weighted relative improvement:

$$d^* = \arg \max_{\substack{\|d\|_2 \le 1 \\ e_R^\top d > 0,\ e_F^\top d > 0}} \left\{ \omega_R \log(e_R^\top d) + \omega_F \log(e_F^\top d) \right\}. \tag{9}$$

Theorem 1 guarantees that the feasible set is nonempty. Its unique solution is

$$\begin{aligned} \kappa &= \frac{c(\omega_R - \omega_F) + \sqrt{c^2(\omega_R - \omega_F)^2 + 4\omega_R\omega_F}}{2\omega_F}, \\ d^* &= \frac{\kappa e_R + e_F}{D}, \qquad D = \sqrt{\kappa^2 + 2\kappa c + 1}. \end{aligned} \tag{10}$$

The self-observations thus set priorities, while the current gradient geometry converts them into compatible gains. Equal preferences yield $(e_R + e_F)/\|e_R + e_F\|_2$. Appendix D proves the closed

form and supplies its numerically stable evaluation; Section 6.6 defines comparisons between the adaptive and symmetric directions.

5.2 Position strength and realized observation coefficients

Let $B = \frac{1}{2}\sum_i(\sqrt{p_i} - \sqrt{q_i})^2$ and $w = \min(B/\tau, 1)$ for $\tau > 0$. Expanding the direction and position strength gives the complete logit signal:

$$\begin{aligned} g_s^* &= w n_F d^* = \frac{\min\left\{\frac{1}{2\tau}\sum_i(\sqrt{p_i} - \sqrt{q_i})^2,\, 1\right\}}{\sqrt{\kappa^2 + 2\kappa c + 1}} \left[\frac{n_F \kappa}{n_R} g_R + g_F\right], \\ g_s^* &= a_s g_R + b_s g_F, \qquad a_s = \frac{w n_F \kappa}{D n_R}, \qquad b_s = \frac{w}{D}. \end{aligned} \tag{11}$$

These coefficients instantiate W_s in Eq. (5). The derivatives supply the candidate directions, the self-observation values condition their relative influence, and w controls the contribution of the position. The coefficients remain fixed during differentiation.

Update properties. The signal obeys $\|g_s^*\|_2 \le \sqrt{2}$. For $q_j < p_j$, its correction satisfies $-g_{s,j}^* \ge w\omega_F(p_j - q_j)$, retaining a positive limiting bound as $q_j \to 0$ for a fixed interior teacher. The coordinated direction has positive inner product with both source-centered gradients. Near agreement, discrepancy gating gives cubic attenuation in $\|q - p\|_2$; with aligned gradients, the update reduces to $w(q - p)$ (Appendices E, F, and D.1). The component comparisons examine these three aspects of the update: candidate support, direction, and magnitude.

We accumulate the signed surrogate over valid response positions and differentiate it with one student backward pass. Appendix H gives the explicit batch objective, parameter gradient, and numerical implementation.

6 Experiments

6.1 Experimental settings

Models and training data. Students are Qwen3-1.7B and Qwen3-4B, with Qwen3-4B-Base used for the base-initialization study (Qwen Team, 2025a). We use *Qwen3-30B-A3B-Instruct-2507* and *Qwen3-30B-A3B-Thinking-2507* as teachers (Qwen Team, 2025b;c). The main setting pairs the Instruct teacher with non-thinking students; auxiliary settings use the Thinking teacher with thinking students and the Instruct teacher with Qwen3-4B-Base. All methods train on DAPO-Math-17k (Yu et al., 2025) with fresh student rollouts; distillation methods query a frozen teacher at the same student prefixes.

Evaluation benchmarks and metrics. We evaluate AIME24–26 for mathematical reasoning and GPQA-Diamond (Rein et al., 2023) and MBPP+ (Liu et al., 2023) for out-of-distribution transfer. Avg@32 is the mean accuracy over 32 responses per problem and Pass@32 the fraction of problems solved by at least one; OOD columns report Avg@32 in Table 1 and Pass@32 in Table 2. We compare with GRPO (Shao et al., 2024), OPD (Agarwal et al., 2024b), OPSD (Zhao et al., 2026b), EOPD (Jin et al., 2026), and KD (Hinton et al., 2015; Kim & Rush, 2016), as applicable. The 4B non-thinking setting also reports pure RKL and FKL objectives. Within each comparison, methods share initialization, prompts, rollout budget, decoding, and evaluation protocol, and teacher-based methods share the same teacher.

6.2 Main results

Performance on mathematical reasoning. Table 1 shows consistent gains for both non-thinking student scales. For Qwen3-1.7B, OC-OPD outperforms GRPO, OPD, and OPSD on all six AIME metrics, beating the strongest baseline by +7.02 Avg@32 and +13.34 Pass@32 points on average. For Qwen3-4B, it exceeds EOPD by +2.74 Avg@32 and +3.33 Pass@32 points across AIME24–26. OC-OPD is best or tied-best in five of six columns, with RKL higher only on AIME24 Avg@32 by 0.10 points. Thus, the gains cover both average accuracy and multi-sample success.

Table 1: **Results for non-thinking students** (accuracy %). Shaded rows denote OC-OPD. Δ reports the Qwen3-4B gain over EOPD in percentage points, computed from the displayed scores.

Methods	Math (I.D.)						O.O.D.	
	AIME 2024		AIME 2025		AIME 2026		GPQA_D	MBPP+
	Avg@32	Pass@32	Avg@32	Pass@32	Avg@32	Pass@32	Avg@32	
Teacher: Qwen3-30B-A3B-Instruct-2507								
Qwen3-1.7B	14.16	43.33	9.90	30.00	7.71	23.33	29.59	56.03
+ GRPO	33.96	70.00	25.31	46.67	26.04	50.00	—	—
+ OPD	32.08	73.33	20.52	50.00	23.54	46.67	—	—
+ OPSD	33.33	73.33	22.50	53.33	21.35	56.67	—	—
+ OC-OPD	41.67	80.00	34.38	66.67	30.31	76.67	34.01	59.01
Qwen3-4B	23.85	56.67	22.40	56.67	15.10	53.33	39.60	65.55
+ OPD (RKL)	61.04	86.67	51.04	76.67	52.19	80.00	43.43	64.29
+ OPD (FKL)	55.42	83.33	53.75	73.33	54.17	83.33	40.91	65.55
+ EOPD	60.10	86.67	53.13	83.33	54.69	83.33	45.45	64.81
+ OC-OPD	60.94	93.33	56.67	83.33	58.54	86.67	46.80	66.01
Δ vs. EOPD	+0.84	+6.66	+3.54	+0.00	+3.85	+3.34	+1.35	+1.20

Combining the two KL directions. Neither KL direction dominates at 4B: RKL leads FKL by 5.62 Avg@32 points on AIME24, whereas FKL is stronger on AIME25–26 and keeps MBPP+ at initialization. OC-OPD retains both strengths: it is within 0.10 points of RKL on AIME24, exceeds the stronger one by +2.92 and +4.37 Avg@32 points on AIME25–26, and surpasses both OOD. It also beats the adaptive linear mixture on all six AIME metrics (Table 4), so the gain comes from coordination rather than mixing alone.

6.3 Out-of-distribution evaluation

Generalization beyond mathematical reasoning. Although trained only on DAPO-Math-17k, OC-OPD also improves out-of-distribution performance (Table 1, last two columns). Relative to its initialization, the 1.7B student gains +4.42 points on GPQA-Diamond and +2.98 points on MBPP+. At 4B, it reaches 46.80 and 66.01, exceeding EOPD by +1.35 and +1.20 points. On MBPP+, OC-OPD stays above initialization, whereas EOPD and OPD (RKL) drop by 0.74 and 1.26 points.

6.4 Thinking teacher and thinking students

Table 2: **Thinking-model results** (accuracy %). Shaded rows denote OC-OPD.

Methods	Math (I.D.)						O.O.D.	
	AIME24		AIME25		AIME26		GPQA_D	MBPP+
	avg@32	pass@32	avg@32	pass@32	avg@32	pass@32	pass@32	
Qwen3-1.7B	50.94	83.33	37.19	73.33	38.65	70.00	77.78	84.66
+ OPD	48.54	86.67	40.21	73.33	42.08	73.33	79.29	83.60
+ OC-OPD	52.81	86.67	46.46	76.66	48.85	76.67	81.82	85.45
Qwen3-4B	72.71	86.67	65.52	83.33	61.04	83.33	84.34	87.04
+ OPD	78.13	83.33	65.94	80.00	64.17	83.33	81.31	87.57
+ OC-OPD	80.21	93.33	66.98	86.67	66.77	90.00	85.35	88.36

Performance with thinking generation. Table 2 shows that the advantage of OC-OPD persists with thinking generation. Relative to OPD, the 1.7B student improves by +5.76 Avg@32 and +2.22 Pass@32 points on average across AIME24–26, while the 4B student gains +1.91 and +7.78 points, respectively. OC-OPD also exceeds OPD on both GPQA-Diamond and MBPP+ at both model scales. These results show that the proposed update is not specific to non-thinking decoding and remains effective when the student already produces explicit reasoning traces.

6.5 BASE-MODEL INITIALIZATION

Table 3: Pass@8 performance comparison on AIME with the Qwen3-4B-Base student. Shaded rows denote OC-OPD. Bold denotes the best value in each column.

Methods	AIME 2024	AIME 2025
	Pass@8	Pass@8
Qwen3-4B-Base	23.33	23.33
+ KD	26.67	23.33
+ OPD	33.33	30.00
+ EOPD	33.33	33.33
+ OC-OPD	**36.67**	**36.67**

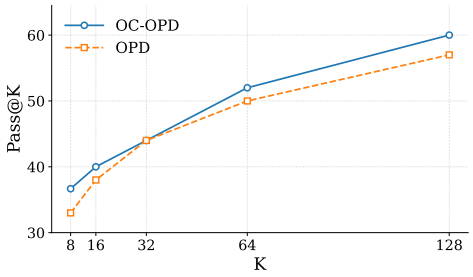


Figure 2: Pass@k performance comparison between OPD and OC-OPD on AIME24 with the Qwen3-4B-Base student.

Performance from base-model initialization. We further test whether the gains depend on instruction-tuned initialization by distilling into Qwen3-4B-Base. Table 3 shows that OC-OPD reaches 36.67 Pass@8 on both AIME24 and AIME25, improving the base model by +13.34 points and EOPD by +3.34 points on each benchmark. Across sampling budgets (Figure 2), OC-OPD exceeds OPD at four of five tested values of k and ties at $k = 32$, so the advantage persists from a substantially weaker initialization.

6.6 ABLATION STUDY

Table 4: **Ablation study.** The student is Qwen3-4B with non-thinking generation and the teacher is Qwen3-30B-A3B-Instruct-2507. Bold denotes the best value in each column.

Method	AIME 2024		AIME 2025		AIME 2026	
	Avg@32	Pass@32	Avg@32	Pass@32	Avg@32	Pass@32
OC-OPD	60.94	**93.33**	**56.67**	**83.33**	**58.54**	**86.67**
Forward KL (FKL)	55.42	83.33	53.75	73.33	54.17	83.33
Reverse KL (RKL)	**61.04**	86.67	51.04	76.67	52.19	80.00
Adaptive linear mixture	60.21	86.67	53.02	80.00	53.96	80.00
Equal-preference	58.65	86.67	53.54	80.00	56.67	83.33
w/o position gate	60.63	90.00	54.06	80.00	56.67	83.33

Effect of individual design choices. Table 4 shows that no ablated objective matches the full OC-OPD across AIME24–26. RKL is only 0.10 points higher on AIME24 Avg@32, but is clearly worse on AIME25–26 and on Pass@32, while FKL shows a similar coverage gap. Adaptive linear mixing and equal-preference weighting recover part of the performance but remain below the full method. Removing the position gate is the strongest ablation, yet it reduces Pass@32 by about 3.33 points on all three benchmarks. These comparisons show that direction-aware preferences and position-sensitive weighting provide complementary benefits. Detailed definitions are given in Appendix I.4.

7 CONCLUSION

OC-OPD organizes teacher and student predictions at each student-generated prefix into four self- and cross-observations, yielding a unified supervision state. Source-centered contrasts recover the complementary ranking and coverage corrections of reverse and forward KL. The observation–gradient identity establishes their local compatibility, while observation-conditioned coordination combines them in a bounded update that retains a correction for teacher-supported candidates with low student probability. Experiments across non-thinking, thinking, and base-initialized students show consistent gains in mathematical reasoning. Scientific question answering and code generation further assess transfer beyond the mathematical training domain.

AI USE STATEMENT

Generative AI tools were used solely for limited assistance with language polishing, formatting, routine coding, and debugging. The research, analysis, interpretation of results, and preparation of the final content were carried out by the human authors. All results and citations were independently verified by the authors, who take full responsibility for the work.

ETHICS STATEMENT

This work studies on-policy distillation using publicly available datasets and pre-trained open-source models, and all experiments follow the licenses of the corresponding resources. Because distillation can transfer characteristics of the teacher into the student, downstream use should include task-appropriate evaluation of factuality, bias, and safety together with standard deployment safeguards.

REPRODUCIBILITY STATEMENT

Full proofs appear in Appendices. Appendix H describes the gradient implementation, and Appendix I provides training and sampling hyperparameters, evaluation protocols, and controlled comparisons.

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*, 2024a. URL https://arxiv.org/abs/2306.13649.

Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stanczyk, Sabela Ramos Garea, Matthieu Geist, and Olivier Bachem. On-policy distillation of language models: Learning from self-generated mistakes. In *The Twelfth International Conference on Learning Representations (ICLR)*. OpenReview.net, 2024b.

Shun-ichi Amari. Natural Gradient Works Efficiently in Learning. *Neural Computation*, 10(2): 251–276, 1998. doi: 10.1162/089976698300017746. URL https://doi.org/10.1162/089976698300017746.

Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. MiniLLM: Knowledge Distillation of Large Language Models. In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2306.08543v4.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the Knowledge in a Neural Network. *arXiv preprint arXiv:1503.02531*, 2015. URL https://arxiv.org/abs/1503.02531.

Ijun Jang, Jewon Yeom, Juan Yeo, Hyunggyu Lim, and Taesup Kim. Stable On-Policy Distillation through Adaptive Target Reformulation. *arXiv preprint arXiv:2601.07155*, 2026. URL https://arxiv.org/abs/2601.07155.

Nan Jia, Haojin Yang, Xing Ma, Jiesong Lian, Shuailiang Zhang, Weipeng Zhang, Ke Zeng, Xunliang Cai, and Zequn Sun. Asymmetric On-Policy Distillation: Bridging Exploitation and Imitation at the Token Level. *arXiv preprint arXiv:2605.06387*, 2026. URL https://arxiv.org/abs/2605.06387v4.

Chenyang Jiang and Changhan Huang. SMOPD: Selective Token-Entropy Masking for Dirty-History Multi-Turn On-Policy Self-Distillation. *arXiv preprint arXiv:2608.14647*, 2026. URL https://arxiv.org/abs/2608.14647.

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. *arXiv preprint arXiv:2603.07079*, 2026. URL https://arxiv.org/abs/2603.07079.

Yoon Kim and Alexander M. Rush. Sequence-Level Knowledge Distillation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 1317–1327. Association for Computational Linguistics, 2016. doi: 10.18653/v1/D16-1139. URL https://aclanthology.org/D16-1139/.

Jongwoo Ko, Sungnyun Kim, Tianyi Chen, and Se-Young Yun. DistiLLM: Towards Streamlined Distillation for Large Language Models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pp. 24872–24895. PMLR, 2024. URL https://proceedings.mlr.press/v235/ko24c.html.

Jongwoo Ko, Tianyi Chen, Sungnyun Kim, Tianyu Ding, Luming Liang, Ilya Zharkov, and Se-Young Yun. DistiLLM-2: A Contrastive Approach Boosts the Distillation of LLMs. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267, pp. 31044–31062. PMLR, 2025. URL https://proceedings.mlr.press/v267/ko25a.html.

Jongwoo Ko, Sara Abdali, Young Jin Kim, Tianyi Chen, and Pashmina Cameron. Scaling Reasoning Efficiently via Relaxed On-Policy Distillation. *arXiv preprint arXiv:2603.11137*, 2026. URL https://arxiv.org/abs/2603.11137.

Vitaly Kurin, Alessandro De Palma, Ilya Kostrikov, Shimon Whiteson, and M. Pawan Kumar. In Defense of the Unitary Scalarization for Deep Multi-Task Learning. In *Advances in Neural Information Processing Systems*, 2022. URL https://arxiv.org/abs/2201.04122.

Chia-Hsuan Lee, Zelei Cheng, Yu Wang, Renkun Ni, Sambit Sahu, Shi-Xiong Zhang, and William Campbell. SEAD: Competence-Aware On-Policy Distillation via Entropy-Guided Supervision. *arXiv preprint arXiv:2606.28562*, 2026. URL https://arxiv.org/abs/2606.28562.

Bo Liu, Xingchao Liu, Xiaojie Jin, Peter Stone, and Qiang Liu. Conflict-Averse Gradient Descent for Multi-task Learning. In *Advances in Neural Information Processing Systems*, volume 34, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/9d27fdf2477ffbff837d73ef7ae23db9-Abstract.html.

Jiawei Liu, Chunqiu Steven Xia, Yuyao Wang, and Lingming Zhang. Is your code generated by chatGPT really correct? rigorous evaluation of large language models for code generation. In *Advances in Neural Information Processing Systems*, 2023.

John F. Nash, Jr. The Bargaining Problem. *Econometrica*, 18(2):155–162, 1950. URL https://www.jstor.org/stable/1907266.

Aviv Navon, Aviv Shamsian, Idan Achituve, Haggai Maron, Kenji Kawaguchi, Gal Chechik, and Ethan Fetaya. Multi-Task Learning as a Bargaining Game. In *International Conference on Machine Learning*, 2022. URL https://proceedings.mlr.press/v162/navon22a.html.

Frank Nielsen. An Elementary Introduction to Information Geometry. *Entropy*, 22(10):1100, 2020. doi: 10.3390/e22101100. URL https://doi.org/10.3390/e22101100.

Qwen Team. Qwen3: Think Deeper, Act Faster. Official model release, 2025a. URL https://qwenlm.github.io/blog/qwen3/.

Qwen Team. Qwen3-30B-A3B-Instruct-2507. Official model card, 2025b. URL https://huggingface.co/Qwen/Qwen3-30B-A3B-Instruct-2507.

Qwen Team. Qwen3-30B-A3B-Thinking-2507. Official model card, 2025c. URL https://huggingface.co/Qwen/Qwen3-30B-A3B-Thinking-2507.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. Gpqa: A graduate-level google-proof q&a benchmark. *arXiv preprint arXiv:2311.12022*, 2023.

Stéphane Ross, Geoffrey Gordon, and Drew Bagnell. A Reduction of Imitation Learning and Structured Prediction to No-Regret Online Learning. In *International Conference on Artificial Intelligence and Statistics*, pp. 627–635, 2011. URL https://proceedings.mlr.press/v15/ross11a.html.

Ozan Sener and Vladlen Koltun. Multi-Task Learning as Multi-Objective Optimization. In *Advances in Neural Information Processing Systems*, volume 31, 2018. URL https://proceedings.neurips.cc/paper/2018/hash/432aca3a1e345e339f35a30c8f65edce-Abstract.html.

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

Jing Wu and Mehrtash Harandi. MUNBa: Machine Unlearning via Nash Bargaining. *arXiv preprint arXiv:2411.15537*, 2024. URL https://arxiv.org/abs/2411.15537. Revised 2025.

Xingrun Xing, Haoqing Wang, Boyan Gao, Ziheng Li, and Yehui Tang. Trust Region On-Policy Distillation. *arXiv preprint arXiv:2606.01249*, 2026. URL https://arxiv.org/abs/2606.01249.

Wenda Xu, Rujun Han, Zifeng Wang, Long T. Le, Dhruv Madeka, Lei Li, William Yang Wang, Rishabh Agarwal, Chen-Yu Lee, and Tomas Pfister. Speculative Knowledge Distillation: Bridging the Teacher-Student Gap Through Interleaved Sampling. In *International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2410.11325v3.

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.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, 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, Weinan Dai, 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. *arXiv preprint arXiv:2503.14476*, 2025. URL https://arxiv.org/abs/2503.14476.

Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn. Gradient Surgery for Multi-Task Learning. In *Advances in Neural Information Processing Systems*, 2020. URL https://arxiv.org/abs/2001.06782.

Zichao Yu, Chengzhi Yu, Shengze Xu, Yujin Han, Bingqing Jiang, Xu Wang, and Difan Zou. Mismatch Matters: On-Policy Distillation Beyond Token Agreement. *arXiv preprint arXiv:2608.09836*, 2026. URL https://arxiv.org/abs/2608.09836.

Anhao Zhao, Junlong Tong, Yingqi Fan, Ping Nie, Wenjie Li, and Xiaoyu Shen. PowerOPD: Stabilizing On-Policy Distillation with Bounded Power Transformation. *arXiv preprint arXiv:2606.17199*, 2026a. URL https://arxiv.org/abs/2606.17199.

Siyan Zhao, Zhihui Xie, Mengchen Liu, Jing Huang, Guan Pang, Feiyu Chen, and Aditya Grover. Self-distilled reasoner: On-policy self-distillation for large language models. *arXiv preprint arXiv:2601.18734*, 2026b.

A NOTATION, ASSUMPTIONS, AND THE OBSERVATION MATRIX

Table 5: Notation used throughout the paper. All local quantities refer to the same current prefix.

Symbol	Meaning
$s = (x, y_{<t})$	A prefix visited by the current student behavior policy
$p, q \in \Delta_V^\circ$	Frozen teacher and current student distributions with positive support
$z \in \mathbb{R}^V$	Student logits, with $q = \mathrm{softmax}(z)$
M	Self-/cross-entropy observation matrix in Eq. (3)
R, F	Student-sourced and teacher-sourced observation contrasts
g_R, g_F	Their Euclidean gradients with respect to current logits
$J(q)$	$\mathrm{diag}(q) - qq^\top$, with kernel $\mathrm{span}\{\mathbf{1}\}$
n_j, e_j	Gradient norm $\|g_j\|_2$ and normalized gradient g_j/n_j
c	Euclidean cosine $e_R^\top e_F$
ω_R, ω_F	Positive bargaining preferences summing to one
d^*	Unit-norm bargaining direction, used with a minus sign in descent
B, w	Squared Hellinger discrepancy and position weight
g^*	Final logit signal $wn_F d^*$
$\mathrm{sg}[\cdot]$	Stop-gradient operator evaluated at the current snapshot
A_s, K_{st}	Parameter-to-logit Jacobian and block coupling $A_s A_t^\top$

Interior support. The theoretical domain is $\Delta_V^\circ = \{q \in \mathbb{R}^V : q_i > 0, \sum_i q_i = 1\}$. Theorems 1 and 5 hold throughout this open simplex, including arbitrarily small positive coordinates. The expansion around agreement and the coordinate limit $g_{R,j} \to 0$ use a fixed interior teacher. Boundary limits are taken through interior distributions, with divergence values interpreted in the extended-real sense at boundary points.

Freezing conventions. The teacher probabilities, prefix samples, preferences, gate, and bargaining solution are detached during the local optimization step. The student probabilities inside R, F in Eq. (5), or inside $\log q_\theta$ in Eq. (18), remain differentiable. The differentiated objective is conditional on the current prefix distribution. Fresh student sampling supplies the state distribution for each subsequent update.

What each observation contributes. M_{11} supplies student dispersion and the student-source baseline. M_{22} supplies teacher dispersion and the teacher-source baseline. The cross-observations M_{21} and M_{12} evaluate each source under the other model. Their within-source differences define centered supervision signals, and their total excess supplies the geometric certificate. Frozen teacher entropy influences the adaptive coefficients through its numerical value.

Organized observations and candidate resolution. The matrix provides a canonical organization of four expected scores. The aligned distributions p, q provide candidate-level probabilities, likelihood ratios, and gradient geometry. OC-OPD uses this information jointly: w determines position strength and g^* allocates the learning signal. Every candidate remains represented in the gradient construction.

A.1 FROM THE FOUR OBSERVATIONS TO A PARAMETER GRADIENT

The convention $H(a, b) = -\sum_i a_i \log b_i$ averages the evaluator's negative log probability under candidate source a. Each column fixes that source while changing the evaluator. The centered contrasts in Eq. (4) therefore admit the representations

$$R = H(q, p) - H(q) = \sum_i q_i \log(q_i/p_i) = D_{\mathrm{KL}}(q\|p),$$

$$F = H(p, q) - H(p) = \sum_i p_i \log(p_i/q_i) = D_{\mathrm{KL}}(p\|q).$$

These identities express the source-centered observation costs in standard divergence notation.

Entrywise derivatives. At a fixed prefix, let $q = q_\theta = \mathrm{softmax}(z)$ and hold p fixed. Direct differentiation of the four entries gives

$$\begin{aligned}
\nabla_z M_{11} &= -q \odot \big(\log q + H(q)\mathbf{1}\big),\\
\nabla_z M_{12} &= q - p,\\
\nabla_z M_{21} &= -q \odot \big(\log p + H(q,p)\mathbf{1}\big),\\
\nabla_z M_{22} &= 0.
\end{aligned} \tag{12}$$

For the first identity, $\nabla_q H(q) = -(\log q + \mathbf{1})$ and multiplication by $J(q)$ subtracts the q-mean. For the third, $\nabla_q H(q,p) = -\log p$, so the same centering gives the stated expression. The second is the usual cross-entropy logit derivative, and the last follows because the teacher is frozen.

Substituting these derivatives into the matrix loss yields

$$\begin{aligned}
\nabla_z \ell_s &= -a_s \nabla_z M_{11} + b_s \nabla_z M_{12} + a_s \nabla_z M_{21} - b_s \nabla_z M_{22}\\
&= a_s q \odot \big(\log(q/p) - R\mathbf{1}\big) + b_s(q-p).
\end{aligned} \tag{13}$$

The student self-observation contributes the centering term through its derivative. Combining it with the teacher's evaluation gives the student-sourced direction g_R. For the teacher-sourced contrast, the constant $H(p)$ establishes the zero-at-agreement baseline, and $H(p,q)$ supplies the derivative g_F.

Values condition the update; derivatives provide its direction. All four entries are evaluated at the current snapshot. The diagonal values condition the preference rule in Eq. (15); the two excesses and their derivatives define the candidate-level objectives and geometry. Teacher entropy influences the numerical coefficients through the conditioning path, with a zero direct derivative. Alignment and gate saturation can make some coefficient sensitivities vanish at particular pairs, while the observation-to-gradient map retains all four inputs.

The coefficient structure W_s pairs each source's self- and cross-observation through one realized coefficient. The frozen-coefficient convention differentiates M_θ with a_s, b_s held at their current values. In the product-rule expansion, the derivative of these fixed coefficients is zero, yielding exactly $a_s \nabla_\theta R + b_s \nabla_\theta F$. Recomputing the coefficients after each student update refreshes their dependence on the current observations.

Exact batch gradient. Let $\mathcal{S}_k$ contain the N_k valid response positions in a fresh student batch, and evaluate the controller at θ_k. With $A_s = \partial z_s / \partial \theta$, define

$$\begin{aligned}
\mathcal{L}_k(\theta) &= \frac{1}{N_k} \sum_{s \in \mathcal{S}_k} \big\langle \mathrm{sg}[W_s(\theta_k)], M_\theta(s) \big\rangle_F,\\
\nabla_\theta \mathcal{L}_k(\theta)|_{\theta_k} &= \frac{1}{N_k} \sum_{s \in \mathcal{S}_k} A_s^\top \big(a_s g_{R,s} + b_s g_{F,s}\big) = \frac{1}{N_k} \sum_{s \in \mathcal{S}_k} A_s^\top g_s^*.
\end{aligned} \tag{14}$$

For plain SGD, $\theta_{k+1} = \theta_k - \eta_k \nabla_\theta \mathcal{L}_k(\theta_k)$. Other optimizers consume the same gradient with their own parameter-space transformation. The coefficients are deterministic functions of current distributions. Each update refreshes the matrix and coefficient computation, with the complete autoregressive prefix serving as the conditioning state.

A.2 Complete current-state coefficient rule

Self-observations to preferences. Define

$$\begin{aligned}
h_S &= \frac{M_{11}}{M_{11} + h_0}, \qquad h_T = \frac{M_{22}}{M_{22} + h_0},\\
\nu_F &= \epsilon + h_T, \qquad \nu_R = \epsilon + (1 - h_T) h_S, \qquad \omega_j = \frac{\nu_j}{\nu_R + \nu_F}, \quad j \in \{R, F\}.
\end{aligned} \tag{15}$$

Here $h_0, \epsilon > 0$. The ν_j are unnormalized preferences; they are distinct from the realized coefficients a_s, b_s below.

Increasing h_T shifts preference toward the teacher-sourced signal; increasing h_S at fixed h_T shifts it toward the student-sourced signal. A concentrated teacher paired with a diffuse student therefore emphasizes the teacher's assessment of student-distributed candidates. A diffuse teacher gives greater priority to representing its distributed support. The positive floor retains both views when both distributions are concentrated, including when their favored candidates differ. This continuous entropy-aware parameterization instantiates the preference layer of the controller. The compatibility and bargaining results hold for every positive preference pair, separating the structural guarantee from the specific preference map.

Cross-observation excesses to direction. Compute g_R, g_F from Eq. (6), $n_j = \|g_j\|_2$, $e_j = g_j/n_j$, and $c = e_R^\top e_F$. Let κ, D be the closed-form quantities in Eq. (25). Then $d^* = (\kappa e_R + e_F)/D$. The ratio κ is an algebraic solution of the two-objective bargaining problem; it is distinct from the iteration index k.

Discrepancy to position strength. With $\tau > 0$, set

$$B = \frac{1}{2}\sum_i (\sqrt{p_i} - \sqrt{q_i})^2 \in [0,1],$$
$$w = \arg\max_{0\le u\le 1} \left\{ uB - \frac{\tau}{2}u^2 \right\} = \mathrm{clip}(B/\tau, 0, 1). \tag{16}$$

This bounded discrepancy is computed directly from the aligned distributions and controls position strength. It complements the matrix's organization of expected scores with a full-distribution measure of agreement. Hellinger-based supervision also appears in sampled excess-token reward shaping (Yu et al., 2026); here it determines the position-level coefficient. Appendix E proves its unique scalar optimum.

Realized loss coefficients. Expanding $g_s^* = w n_F d^*$ in the source-centered observation gradients gives

$$a_s = \frac{w n_F \kappa}{D n_R}, \qquad b_s = \frac{w}{D}, \qquad g_s^* = a_s g_R + b_s g_F. \tag{17}$$

Preferences, realized influence, and total strength have complementary roles: ω_j values relative gains, the gradient angle determines the compromise, and $w n_F$ sets its norm. Near agreement, the small discrepancy smoothly attenuates the final signal at any preference. The coefficients above make Eq. (5) an exact current-point realization of the controller.

A convenient gradient surrogate. A numerically direct realization computes the vector g_s^* and uses

$$\widetilde{\mathcal{L}}_s(\theta) = \sum_i \mathrm{sg}[g_{s,i}^*] \log q_\theta(i \mid s). \tag{18}$$

Since $\mathbf{1}^\top g_s^* = 0$, its current logit gradient is exactly g_s^*; Appendix H expands the derivative. The matrix loss and signed surrogate have identical current gradients. Their scalar values follow their respective definitions.

B Complete proof of the compatibility identity

Lemma 2 (Logit differentiation and the tangent space)**.** *Let $q = \mathrm{softmax}(z)$. Then $\partial q/\partial z = J(q)$, $J(q)\mathbf{1} = 0$, and $J(q)$ is positive definite on $\mathcal{T} = \{v : \mathbf{1}^\top v = 0\}$. For any $u, v \in \mathcal{T}$,*

$$u^\top J(q)^+ v = u^\top \mathrm{diag}(q)^{-1} v. \tag{19}$$

Proof. The softmax derivative is $\partial q_i/\partial z_j = q_i(\mathbb{1}_{i=j} - q_j)$. For any x, $x^\top J x = \sum_i q_i x_i^2 - (\sum_i q_i x_i)^2 = \mathrm{Var}_q(x_i)$. Since every $q_i > 0$, this variance is zero exactly when x is constant. Thus the kernel is $\mathrm{span}\{\mathbf{1}\}$ and restriction to $\mathcal{T}$ is positive definite.

For $v \in \mathcal{T}$, define $x = \mathrm{diag}(q)^{-1} v$. Then

$$Jx = v - q(q^\top \mathrm{diag}(q)^{-1} v) = v - q\mathbf{1}^\top v = v.$$

The vector J^+v is the unique zero-sum solution of this equation; any two solutions differ by a constant vector. Hence $J^+v = x + \alpha\mathbf{1}$ for some scalar α. Multiplication by $u^\top$, with $u \in \mathcal{T}$, proves Eq. (19). This equality identifies the bilinear forms restricted to the tangent space $\mathcal{T}$. □

Lemma 3 (KL gradients and their zero sets)**.** *With the teacher frozen, Eq.* (6) *holds. Each gradient is zero if and only if $p = q$.*

Proof. Let $\ell_i = \log(q_i/p_i)$. Differentiating $R = \sum_i q_i\ell_i$ with respect to q gives $\ell + \mathbf{1}$. Therefore

$$\nabla_z R = J(\ell + \mathbf{1}) = q \odot \ell - q(q^\top \ell) = q \odot (\ell - R\mathbf{1}).$$

For $F = \sum_i p_i \log p_i - \sum_i p_i \log q_i$, the derivative of $\log q_i$ gives $\nabla_z F = q - p$. Each gradient sums to zero. If $g_F = 0$, equality of distributions is immediate. If $g_R = 0$, positivity of q_i implies $\ell_i = R$ for every i, so $q_i = e^R p_i$. Normalization forces $e^R = 1$, and $q = p$. The converse follows by substitution. □

Proof of Theorem 1. By Lemmas 2 and 3,

$$g_R^\top J^+ g_F = \sum_i (q_i - p_i)(\ell_i - R) \tag{20}$$

$$= \sum_i q_i \log(q_i/p_i) + \sum_i p_i \log(p_i/q_i) \tag{21}$$

$$= R + F. \tag{22}$$

The R-multiple disappears because the distributions have equal total mass. Substituting their definitions in Eq. (3) proves the matrix contrast.

If $p \neq q$, strict nonnegativity of KL implies $R + F > 0$, and both gradients are nonzero. Suppose their Euclidean cosine were -1. Then $g_R = -\alpha g_F$ for some $\alpha > 0$, and

$$g_R^\top J^+ g_F = -\alpha g_F^\top J^+ g_F < 0,$$

contradicting the identity. Thus $c > -1$. The vector

$$d_0 = \frac{e_R + e_F}{\sqrt{2 + 2c}}$$

has unit norm and satisfies $e_R^\top d_0 = e_F^\top d_0 = (1 + c)/\sqrt{2 + 2c} > 0$. This is an explicit common-descent witness. The case $c = 1$ is permitted and simply gives $d_0 = e_R = e_F$. □

Metric-dependent gradient geometry. Equation (7) weights coordinate products by $1/q_i$, giving small-student-probability coordinates substantial influence. The Euclidean inner product weights these coordinates uniformly. The identity therefore certifies compatibility in the Fisher inverse metric across both acute and obtuse Euclidean configurations. At each interior pair with $p \neq q$, the Euclidean cosine satisfies $c > -1$; the resulting angular margin $1 + c$ is pair-dependent.

C Extension to strictly convex divergences

The following extension identifies a monotonicity principle underlying the compatibility identity. It establishes a shared descent direction for a finite family of smooth, strictly convex distributional contrasts.

Theorem 4 (Monotone divergence compatibility)**.** *Let $p, q \in \Delta_V^\circ$, $p \neq q$, and let $f, h : (0, \infty) \to \mathbb{R}$ be continuously differentiable strictly convex functions with $f(1) = h(1) = 0$. Define*

$$D_f(q\|p) = \sum_i p_i f(r_i), \quad r_i = q_i/p_i, \quad g_f = \nabla_z D_f(q\|p),$$

and similarly define g_h. Then

$$g_f = q \odot \big(f'(r) - \mathbb{E}_q[f'(r)]\mathbf{1}\big), \tag{23}$$

$$g_f^\top J^+ g_h = \frac{1}{2} \sum_{i,j} q_i q_j \big(f'(r_i) - f'(r_j)\big)\big(h'(r_i) - h'(r_j)\big) > 0. \tag{24}$$

For any finite collection $f_1, \ldots, f_m$ satisfying these assumptions, there is a single direction d such that $g_{f_j}^\top d > 0$ for every j.

Proof. The derivative of D_f with respect to q_i is $f'(r_i)$. Multiplying by the softmax Jacobian gives Eq. (23). Applying Lemma 2,

$$g_f^\top J^+ g_h = \sum_i q_i(f'(r_i) - \mathbb{E}_q f')(h'(r_i) - \mathbb{E}_q h') = \mathrm{Cov}_q(f'(r), h'(r)).$$

For arbitrary sequences a_i, b_i, expansion of the double sum gives

$$\mathrm{Cov}_q(a, b) = \tfrac{1}{2} \sum_{i,j} q_i q_j (a_i - a_j)(b_i - b_j).$$

The derivative of a differentiable strictly convex function on an interval is strictly increasing: equality of its derivative at two distinct points would force it to be constant between them and the function to be affine on that interval. Hence every summand in Eq. (24) is nonnegative. Since $p \neq q$, at least two ratios r_i differ, so at least one summand is strictly positive. This proves the pairing claim.

For a finite family, set $v = J^+ \sum_{j=1}^m g_{f_j}$. For every k,

$$g_{f_k}^\top v = \sum_j g_{f_k}^\top J^+ g_{f_j} > 0.$$

In particular $v \neq 0$, so $d = v/\left\|v\right\|_2$ is a unit common-descent direction. □

Generators and assumptions. The generators $f(r) = r \log r$ and $h(r) = -\log r$ recover the two source-centered observation contrasts. Adding an affine function to a generator preserves its logit gradient for normalized distributions. Differentiability supplies the gradient representation, and strict convexity supplies the strictly increasing derivatives used in the positivity argument. For general differentiable convex generators, the covariance is nonnegative; strict positivity holds when at least one positively weighted pair contributes a strictly positive product.

Compatibility and the bounded controller. The witness $J^+ \sum_j g_{f_j}$ establishes feasibility for the finite family. Its norm reflects the inverse metric and can grow near the simplex boundary. OC-OPD uses a unit Euclidean budget and the explicit two-direction solution to implement bounded local updates. The extension explains the geometric principle, while the source-centered observation construction specifies the candidate populations and statistical meaning of the implemented directions.

D EXISTENCE, UNIQUENESS, AND THE CLOSED-FORM BARGAINING SOLUTION

We prove Theorem 5 and two useful consequences. Let $u_j(d) = e_j^\top d$, $\omega_R + \omega_F = 1$, and $\omega_j > 0$ throughout this section.

Theorem 5 (Unique closed-form compromise)**.** *Under Theorem 1 with $p \neq q$ and $\omega_R, \omega_F > 0$, Eq.* (9) *has a unique solution. Set*

$$\kappa = \frac{c(\omega_R - \omega_F) + \sqrt{c^2(\omega_R - \omega_F)^2 + 4\omega_R\omega_F}}{2\omega_F}, \qquad D = \sqrt{\kappa^2 + 1 + 2\kappa c}. \tag{25}$$

Then $d^* = (\kappa e_R + e_F)/D$, and both $g_j^\top d^*$ are strictly positive.

Existence. Theorem 1 gives a feasible point d_0. On the closed unit ball both utilities lie in $[-1, 1]$. For a maximizing sequence with objective at least that of d_0, both utilities stay bounded away from zero: approaching zero would send a positive-weight logarithm to $-\infty$, while the other logarithm is at most zero. A convergent subsequence of this sequence therefore has a limit in the strictly positive-utility domain. Continuity proves that the maximum is attained.

The norm constraint is active. If a feasible d has $\left\|d\right\|_2 < 1$, then $d/\left\|d\right\|_2$ is feasible and increases the objective by $-\log\left\|d\right\|_2 > 0$. Every optimum consequently has unit norm.

Uniqueness. Suppose distinct optima d_1, d_2 exist. Their midpoint is feasible and, by concavity of the log utilities, has objective at least the optimal value. Strict convexity of the Euclidean ball implies $\|(d_1 + d_2)/2\|_2 < 1$ for distinct unit vectors. Radially enlarging the midpoint strictly increases the objective, a contradiction. This proof covers collinear gradients with $c = 1$ as well as linearly independent gradients.

First-order conditions. The objective is differentiable at the optimum. Its gradient is

$$\nabla_d \Phi(d) = \frac{\omega_R e_R}{u_R(d)} + \frac{\omega_F e_F}{u_F(d)}.$$

Writing the multiplier for $(\|d\|_2^2 - 1)/2 \le 0$ as λ, the KKT condition is $\nabla\Phi(d) = \lambda d$. Taking the inner product with the unit vector d gives $\lambda = \omega_R + \omega_F = 1$. Hence

$$d = ae_R + be_F, \qquad a = \frac{\omega_R}{u_R(d)} > 0, \quad b = \frac{\omega_F}{u_F(d)} > 0. \tag{26}$$

Here a, b are coefficients of the unit direction in normalized gradients; the state-indexed a_s, b_s in Eq. (17) multiply the unnormalized gradients in the final update. It follows that

$$a^2 + cab = \omega_R, \qquad b^2 + cab = \omega_F. \tag{27}$$

For $\kappa = a/b > 0$, eliminating b yields

$$\omega_F \kappa^2 + c(\omega_F - \omega_R)\kappa - \omega_R = 0.$$

The constant term is negative and the leading coefficient positive, so exactly one root is positive. This is Eq. (25). Since the actual unique optimum has this positive ratio, normalizing $\kappa e_R + e_F$ reconstructs it. Conversely, this vector satisfies the first-order condition and positive-utility constraints and is therefore optimal by concavity. This completes the proof of Theorem 5.

Corollary 6 (Scale invariance and proportional fairness)**.** *The direction d^* is invariant to positive rescaling of either objective at fixed preferences. For every feasible v with $\|v\|_2 \le 1$,*

$$\omega_R \frac{u_R(v) - u_R(d^*)}{u_R(d^*)} + \omega_F \frac{u_F(v) - u_F(d^*)}{u_F(d^*)} \le 0. \tag{28}$$

Proof. For $\alpha_j > 0$, normalization gives $\alpha_j g_j / \|\alpha_j g_j\|_2 = e_j$, so each utility and its maximizing direction are invariant. By the KKT equation,

$$\sum_j \omega_j \frac{u_j(v)}{u_j(d^*)} = {d^*}^\top v \le 1.$$

Subtracting $\sum_j \omega_j = 1$ proves Eq. (28). The inequality characterizes the weighted relative local gains of the two source-centered contrasts. □

Equal weights and the simplest strong baseline. If $\omega_R = \omega_F = 1/2$, then $\kappa = 1$ for every $c > -1$. Thus $d^* = (e_R + e_F)/\|e_R + e_F\|_2$. This direction is already a feasible common descent direction, independently of uncertainty. The appropriate incremental question is whether observation-dependent preferences yield more useful learning than this symmetric baseline, at a common norm, gate, learning rate search budget, and teacher cost. For $c = e_R^\top e_F$, the direct normalized-gradient mixture $v = \alpha e_R + (1 - \alpha)e_F$ has unnormalized local gains

$$e_R^\top v = \alpha + (1 - \alpha)c, \qquad e_F^\top v = \alpha c + (1 - \alpha).$$

These expressions characterize how unequal mixing preferences interact with the current gradient geometry.

Stable root evaluation. Let $A = c(\omega_R - \omega_F)$ and $Q = \sqrt{A^2 + 4\omega_R\omega_F}$. An algebraically equivalent computation is

$$\kappa = \begin{cases} (A + Q)/(2\omega_F), & A \ge 0, \\ 2\omega_R/(Q - A), & A < 0. \end{cases} \tag{29}$$

The second branch avoids subtracting nearly equal positive quantities. Computing Q with a hypot routine further reduces overflow risk. The reference implementation corrects roundoff by clipping the computed cosine to $[-1, 1]$ and normalizes the resulting vector directly. Theorem 1 gives $c > -1$ for each distinct interior pair.

D.1 ALIGNED-GRADIENT AND BINARY REDUCTIONS

Proposition 7 (Exact reduction under aligned gradients)**.** *If $p \neq q$ and $c = 1$, then $g^* = w(q - p)$ for every positive preference pair.*

Proof. When $e_R = e_F$, Eq. (9) is maximized by their common unit vector, independently of the preferences. A binary logit tangent space is one-dimensional; Theorem 1 rules out opposite directions, so the normalized gradients coincide. □

Distinct normalized candidate directions create room for preference-dependent coordination. A binary vocabulary yields the exact reduction $g^* = w(q - p)$. The directions also align to first order near agreement, as established in Appendix F. Mismatch and gradient angle together characterize the positions at which coordination changes the signal.

E STRENGTH, LOCAL DESCENT, AND CANDIDATE-DEFICIT GUARANTEES

Proposition 8 (Local descent, boundedness, and teacher-deficit signal)**.** *For $p \neq q$, sufficiently small isolated logit steps $z^+ = z - \eta g^*$ strictly decrease both R and F. Moreover,*

$$\|g^*\|_2 = w\,\|q - p\|_2 \leq \sqrt{2}, \qquad -g_j^* \geq w\omega_F(p_j - q_j) \quad \textit{whenever } q_j < p_j. \tag{30}$$

For a fixed positive teacher with $p_j = \rho > 0$, as $q_j \to 0$, Eq. (15) gives

$$\liminf_{q_j \to 0}(-g_j^*) \;\geq \rho \frac{\epsilon}{1 + 2\epsilon} \min\left\{ \frac{1 - \sqrt{1 - \rho}}{\tau}, 1 \right\} > 0. \tag{31}$$

The coordinate lower bound concerns a logit correction. The softmax Jacobian translates this correction into probability changes, whose scale depends on the current student probabilities.

Position-weight optimum. The objective $uB - \tau u^2/2$ has second derivative $-\tau < 0$, so its constrained maximum is unique. The unconstrained stationary point is $u = B/\tau$; projection onto $[0, 1]$ gives Eq. (16). Because $B > 0$ for distinct distributions, $w > 0$ whenever $p \neq q$. If $p = q$, the selected weight is zero.

Norm bound. For distributions p, q,

$$\|p - q\|_2^2 = \sum_i p_i^2 + \sum_i q_i^2 - 2\sum_i p_i q_i \leq 2,$$

since each sum of squares is at most one and the cross term is nonnegative. Since $0 \leq w \leq 1$ and $\|d^*\|_2 = 1$, Eq. (30) follows. The universal constant is a supremum over pairs approaching distinct point masses. With a bounded parameter-to-logit Jacobian, the corresponding parameter gradient is bounded by the product of its operator norm and this logit bound.

An explicit local descent condition. Fix z with $p \neq q$, and let L_j be a Lipschitz constant of $\nabla_z j$ on a ball of radius $r > 0$ around z, for $j \in \{R, F\}$. Such finite constants exist on every compact logit neighborhood for a fixed positive teacher. Set $m_j = g_j^\top d^* > 0$ and $a = wn_F > 0$. The descent lemma gives

$$j(z - \eta a d^*) \leq j(z) - \eta a m_j + \frac{L_j}{2}\eta^2 a^2. \tag{32}$$

Thus both losses strictly decrease whenever

$$0 < \eta < \min\left\{ \frac{r}{a}, \frac{2m_R}{L_R a}, \frac{2m_F}{L_F a} \right\}, \tag{33}$$

with a term interpreted as $+\infty$ if its Lipschitz constant is zero. The restriction $\eta a < r$ keeps the segment inside the ball. This condition relates the admissible local step to the positive gain margin and the smoothness of each source-centered contrast.

Lower bound for a teacher-supported deficit. From Eq. (26), $b = \omega_F/(e_F^\top d^*) \geq \omega_F$, because $0 < e_F^\top d^* \leq 1$. For $q_j < p_j$, $g_{F,j} = q_j - p_j < 0$. Also $\log(q_j/p_j) < 0$ and $R \geq 0$, so $g_{R,j} < 0$. The coefficients a, b are positive; hence

$$-g_j^* = -wn_F(ae_{R,j} + be_{F,j}) \tag{34}$$

$$\geq -wn_F be_{F,j} = wb(p_j - q_j) \geq w\omega_F(p_j - q_j). \tag{35}$$

This argument uses the signs of the individual deficit coordinates and the bargaining KKT equation. The vocabulary construction evaluates every deficit coordinate and supplies the deterministic correction in Eq. (30).

Uniformly positive source preferences. Because $0 \leq h_S, h_T < 1$,

$$\nu_R + \nu_F = 2\epsilon + h_T + (1 - h_T)h_S \leq 1 + 2\epsilon, \qquad \nu_R, \nu_F \geq \epsilon.$$

Both preferences therefore obey $\omega_j \geq \epsilon/(1 + 2\epsilon) > 0$.

The limit as student support vanishes. Fix $p_j = \rho \in (0, 1)$. Cauchy–Schwarz on the remaining coordinates implies

$$\sum_i \sqrt{p_i q_i} \leq \sqrt{\rho q_j} + \sqrt{(1 - \rho)(1 - q_j)}.$$

Consequently,

$$\liminf_{q_j \to 0} B \geq 1 - \sqrt{1 - \rho} > 0.$$

Combining this bound, continuity and monotonicity of $\min(B/\tau, 1)$, the preference floor, and $p_j - q_j \to \rho$ gives Eq. (31). This proves Proposition 8.

Proposition 9 (Finite virtual-step probability-ratio control)**.** *Let $q' = \mathrm{softmax}(z - \eta g^*)$, for any $\eta \geq 0$. Then*

$$\left|\log \frac{q_i'}{q_i}\right| \leq 2\eta\|g^*\|_\infty \leq 2\sqrt{2}\eta \quad \textit{for every } i. \tag{36}$$

In particular, both $D_{\mathrm{KL}}(q'\|q)$ and $D_{\mathrm{KL}}(q\|q')$ are at most $2\eta\|g^*\|_\infty$.

Proof. Write $v = -\eta g^*$. Then $\log(q_i'/q_i) = v_i - \log\sum_j q_j e^{v_j}$. The logarithm of the weighted exponential average lies between $\min_j v_j$ and $\max_j v_j$. The absolute difference is therefore at most $\max_j v_j - \min_j v_j \leq 2\|v\|_\infty$. Taking expectations under either distribution gives the KL bounds. The bound describes an isolated logit step. □

F Near-agreement asymptotics and position strength

Proposition 10 (Cubic attenuation)**.** *Fix $p \in \Delta_V^\circ$ and $h_0, \epsilon, \tau > 0$. Let $q = p + \delta \in \Delta_V^\circ$, $\mathbf{1}^\top\delta = 0$, and $\|\delta\|_2 \to 0$. Then*

$$g^* = \frac{1}{8\tau}\left(\sum_i \frac{\delta_i^2}{p_i}\right)\delta + O(\|\delta\|_2^4). \tag{37}$$

In particular, $\|g^*\|_2 = \Theta(\|\delta\|_2^3)$ along any nonzero fixed perturbation direction.

Proof. Because p is fixed in the interior, Taylor expansions are uniform in a sufficiently small neighborhood. For each coordinate,

$$\sqrt{p_i + \delta_i} - \sqrt{p_i} = \frac{\delta_i}{2\sqrt{p_i}} + O(\delta_i^2).$$

Squaring and summing gives

$$B = \frac{1}{8}\sum_i \frac{\delta_i^2}{p_i} + O(\|\delta\|_2^3). \tag{38}$$

For small enough δ, $B < \tau$, so the gate is unsaturated and $w = B/\tau$.

Also, $\log(1+\delta_i/p_i) = \delta_i/p_i + O(\delta_i^2)$ and $R = O(\|\delta\|_2^2)$. Equation (6) therefore gives

$$g_F = \delta, \qquad g_R = \delta + O(\|\delta\|_2^2).$$

Writing $v = \delta/\|\delta\|_2$ for nonzero δ, it follows that $e_F = v$ and $e_R = v + O(\|\delta\|_2)$. The preferences are continuous near p and uniformly positive. The positive root κ is consequently bounded above and away from zero there. Normalization of $\kappa e_R + e_F$ gives $d^* = v + O(\|\delta\|_2)$, and hence

$$n_F d^* = \delta + O(\|\delta\|_2^2).$$

Multiplying by Eq. (38) divided by τ yields

$$g^* = \frac{1}{8\tau}\left(\sum_i \frac{\delta_i^2}{p_i}\right)\delta + O(\|\delta\|_2^4),$$

as claimed. Positive definiteness of $\mathrm{diag}(p)^{-1}$ supplies the matching upper and lower cubic orders along any fixed nonzero direction. □

Interpretation. Near agreement, the two source-centered gradients agree to first order. Preference-dependent changes enter at higher order, and the gate controls the leading attenuation. Equation (37) quantifies the resulting cubic allocation of supervision to small discrepancies, including its slower fine-matching scale. A positive gate floor would restore a first-order term. The unit-gate ablation isolates the contribution of discrepancy-dependent attenuation.

Behavior at agreement and concentrated predictions. For $p = q$, the exact update is zero at every entropy level, preserving the matching distribution. Concentrated predictions with different favored candidates can have large Hellinger discrepancy and receive strong supervision. Position strength thus responds to cross-model mismatch, while self-observations condition the relative candidate-source influence.

G Comparison related works

Table 6: Comparison of entropy and token-selection-based OPD methods under our four-observation framework. • denotes an observation directly used for token selection or divergence routing, while ◦ denotes an observation involved only through a derived KL contrast.

Method	$\mathbf{M_{11}}$ $H(q)$	$\mathbf{M_{12}}$ $H(p,q)$	$\mathbf{M_{21}}$ $H(q,p)$	$\mathbf{M_{22}}$ $H(p)$
REOPOLD	•	–	–	–
TIP	•	–	–	–
EOPD	–	–	–	•
SEAD	•	–	–	•
OC-OPD (Ours)	•	•	•	•

H Implementation details

Algorithm 1 gives the main training loop. The following details specify its sampling, masking, and numerical realization.

H.1 Training semantics and numerical implementation

Current-policy sampling. The mathematical formulation samples from the full current student distribution q at temperature one. The sampled batch is fixed for one student update, after which fresh rollouts are collected.

Alignment and masking. Teacher and student distributions share vocabulary semantics and next-token positions. Both models evaluate the exact student continuation. The loss sums over valid response positions, including the response end token. The signed token-vocabulary surrogate is summed over these positions and divided by their global count N. This preserves the position strength in Eq. (11).

Degeneracies and stable normalization. Away from agreement, the two gradient norms are nonzero by Theorem 1. The stable quadratic-root expression in Appendix D supports accurate evaluation under unequal preferences. Direct normalization of the combined vector gives the final gradient at the prescribed magnitude.

Loss realization. Because g^* is zero-sum,

$$\frac{\partial}{\partial z_j} \sum_i \mathrm{sg}[g_i^*] \log q_i = g_j^* - q_j \sum_i g_i^* = g_j^*.$$

Thus the signed surrogate produces the desired current logit gradient with one student backward pass. The coefficients and controller computations are detached; student log probabilities retain their computation graph. The teacher remains frozen. The frozen-coefficient convention realizes the matrix differential in Eq. (6) at the current snapshot.

H.2 Real-update diagnostics

Purpose. The geometric analysis in the main text is expressed in terms of a change in student logits, while optimization acts on shared neural-network parameters. We therefore measure whether the logit-space quantities predict the change produced by a real optimizer step. This diagnostic uses a Qwen3-1.7B non-thinking student and the Qwen3-30B-A3B-Instruct-2507 teacher. Training starts from the same zero-LoRA initialization used by the corresponding non-thinking experiments.

Quantities measured. For a fixed token prefix, let $p = \mathrm{softmax}(z_T)$ be the frozen teacher distribution and let $q_0 = \mathrm{softmax}(z_0)$ denote the student distribution immediately before the optimizer update. We define

$$R(z_0) = \mathrm{KL}(q_0 \,\|\, p), \qquad F(z_0) = \mathrm{KL}(p \,\|\, q_0), \tag{39}$$

$$g_R = q_0 \odot \big(\log q_0 - \log p - R(z_0)\big), \qquad g_F = q_0 - p. \tag{40}$$

After the production backward pass and AdamW update, we evaluate the same prefix again to obtain z_1, set $\Delta z = z_1 - z_0$, and compare the predictions

$$\widehat{\Delta R} = g_R^\top \Delta z, \qquad \widehat{\Delta F} = g_F^\top \Delta z \tag{41}$$

with the measured changes

$$\Delta R = R(z_1) - R(z_0), \qquad \Delta F = F(z_1) - F(z_0). \tag{42}$$

All quantities are computed over the vocabulary. A negative predicted or measured change denotes descent for the corresponding objective. We report the Pearson correlation between predicted and measured changes, the fraction of token positions for which their signs agree, and whether the prediction correctly identifies simultaneous descent of both objectives.

Probe construction. Measurements are taken every ten successful optimizer updates, always using the same teacher target before and after the update. We use three probe types. The *on-policy* probe samples positions from the current rollout. The *fixed in-domain* probe reuses fixed prompt–continuation pairs drawn from the training domain. The *HMMT25 OOD* probe contains four held-out prompts with fixed short continuations and is used only to diagnose distribution shift.

Table 7 and Figure 3 show that the local approximation remains accurate after an update to real network parameters. Pearson correlations range from 0.9751 to 0.9993, while individual-objective sign agreement ranges from 98.00% to 100.00%. Joint sign accuracy is at least 99.00% for the two fixed probes and 99.57% on on-policy rollouts. The few sign errors occur close to zero, where the measured change is small. These results empirically validate the local predictive role of $g^\top \Delta z$ under the observed optimizer steps.

Table 7: **Agreement between first-order predictions and measured KL changes.** Correlations and sign agreement are computed over token positions. "Both" is the accuracy of predicting whether RKL and FKL decrease simultaneously.

Probe	Updates	Tokens	r_R	Sign_R	r_F	Sign_F	Both
On-policy rollout	19	464	0.9977	99.57	0.9989	99.57	99.57
Fixed in-domain	25	300	0.9751	99.67	0.9992	98.00	99.00
HMMT25 OOD	25	400	0.9987	99.00	0.9993	100.00	99.25

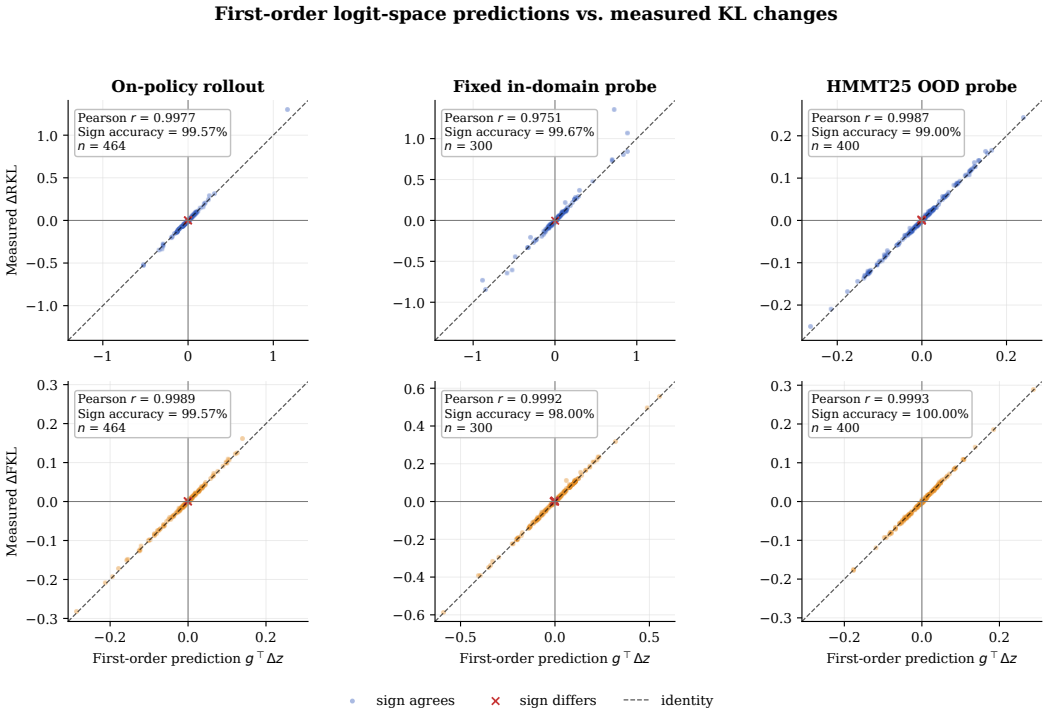


Figure 3: RKL (top) and FKL (bottom) are shown separately for on-policy, fixed in-domain, and HMMT25 OOD probes. Each point is a token position evaluated immediately before and after the same real optimizer step. The dashed line denotes $y = x$ and red crosses indicate sign disagreement.

Interpretation and scope. The diagnostic supports the specific bridge used by our analysis: for the optimizer steps observed during training, the logit-space terms remain accurate predictors of the KL changes induced by a shared-parameter neural-network update. Agreement on current rollouts, fixed in-domain examples, and held-out HMMT25 prompts indicates that this fidelity is not confined to one probe distribution.

I Experimental Details

I.1 Models and Experimental Organization

The teacher checkpoints are Qwen3-30B-A3B-Thinking-2507 and Qwen3-30B-A3B-Instruct-2507. The main experiment pairs the Instruct teacher with non-thinking Qwen3-1.7B and Qwen3-4B students. The first auxiliary experiment pairs the Thinking teacher with thinking students at both sizes; the second pairs the Instruct teacher with Qwen3-4B-Base. Within each comparison block, methods use the same student initialization, teacher, prompt source, training budget, generation/evaluation configuration, and metric implementation; only the learning objective or controller component being compared is changed.

Each iteration generates fresh student continuations and aligns teacher and student distributions on the collected prefixes. The four observations, gradient geometry, and coefficients are recomputed at every valid response position. The teacher remains frozen, and the backward pass updates the student. The selected student regime determines the generation configuration of its training and evaluation continuations, and all methods in that regime use the same prompt wrapper and decoding protocol.

I.2 Training and sampling hyperparameters

Table 8: **Hyperparameters for OC-OPD.** Prompt and response lengths are measured in tokens. A dash denotes the fresh-rollout setting, with one optimization step per rollout batch.

Hyperparameter	OC-OPD
General	
Max prompt length	8192
GPUs	$4\times$ NVIDIA A800 80GB
Training precision	BF16
Optimization	
Global batch size	128
Mini-batch size	32
Mini-batches	4
Group size	1
Off-policy epoch	—
Optimizer	AdamW
Learning rate	1×10^{-5}
LR schedule	Constant
Warmup ratio	0
Weight decay	0
Adam betas	$(0.9, 0.999)$
Gradient clipping	Global L2 norm 1.0
Sampling	
Rollout max response length	32768
Rollout temperature	1.0
Rollout top-p	1.0
Validation max response length	32768
Validation temperature	1.0
Validation top-p	1.0
Method-specific	
h_0	1.0
ϵ	0.05
τ	0.01

I.3 Prompt templates

The prompt wrapper is fixed by the student generation regime and is shared by all compared methods in that regime. The placeholder {`problem`} is replaced by the benchmark problem text. We use the following templates.

Non-thinking.

```
{problem}\n\nPlease solve the following problem and put your final answer within\n\n\boxed{}.
```

Thinking.

```
{problem}\n\nPlease reason step by step, and put your final answer within\n\n\boxed{}.
```

Base.

```
Problem:\n{problem}
```

I.4 CONTROLLED OBJECTIVES AND ABLATIONS

The mathematical ablations use Qwen3-30B-A3B-Instruct-2507 with the Qwen3-4B non-thinking student on DAPO-Math-17k, matching the 4B setting in Table 1. All ablations reuse the same initialization, prompts, optimization and sampling setup, and evaluation budget. Table 9 reports the key controls. The comparison design separates the two information paths in Eq. (5): derivatives of the observations form the candidate-level gradients, while observation values condition their mixture.

Table 9: **Ablation study on mathematical reasoning.** The student is Qwen3-4B with non-thinking generation and the teacher is Qwen3-30B-A3B-Instruct-2507. Scores are percentages; bold denotes the best value in each column.

Method	AIME 2024		AIME 2025		AIME 2026	
	Avg@32	Pass@32	Avg@32	Pass@32	Avg@32	Pass@32
OC-OPD	60.94	**93.33**	**56.67**	**83.33**	**58.54**	**86.67**
Forward KL (FKL)	55.42	83.33	53.75	73.33	54.17	83.33
Reverse KL (RKL)	**61.04**	86.67	51.04	76.67	52.19	80.00
Adaptive linear mixture	60.21	86.67	53.02	80.00	53.96	80.00
Equal-preference	58.65	86.67	53.54	80.00	56.67	83.33
w/o position gate	60.63	90.00	54.06	80.00	56.67	83.33

Full OC-OPD attains the highest Pass@32 on all three AIME editions, while RKL is slightly higher on AIME24 Avg@32. The single-KL controls, adaptive linear mixture, equal-preference variant, and no-gate variant each lose ground on at least one of the coverage metrics. Together, these comparisons support the use of both source-centered directions, observation-conditioned preferences, bargaining-based coordination, and discrepancy-dependent strength.

KL-only controls. Forward KL (FKL) optimizes the teacher-sourced contrast $F = D_{\mathrm{KL}}(p\|q)$, while reverse KL (RKL) optimizes the student-sourced contrast $R = D_{\mathrm{KL}}(q\|p)$. These rows isolate the two candidate-source objectives without the observation-conditioned coordination used by OC-OPD.

Adaptive linear mixture This variant uses the same observation-conditioned preferences ω_R, ω_F, source-centered normalized gradients, and discrepancy gate as OC-OPD, but replaces the bargaining solution with the normalized linear direction $d_{\mathrm{lin}} = (\omega_R e_R + \omega_F e_F)/\|\omega_R e_R + \omega_F e_F\|_2$. The final magnitude remains wn_F, matching the full method.

Equal preferences. Equal-preference bargaining variant fixes $\omega_R = \omega_F = 1/2$ while retaining the source-centered gradients, bargaining rule, and discrepancy gate. It therefore removes observation-conditioned preference adaptation without changing the two candidate sources or the update-strength mechanism.

No position gate. This variant sets $w = 1$, preserving the original preference rule and coordinated direction while removing discrepancy-dependent attenuation of the update magnitude. It directly tests the position-strength component analyzed in Appendix F.