PREFIX-OUTCOME VALIDATION FOR RELIABLE ON-POLICY DISTILLATION OF REASONING MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

On-policy distillation (OPD) provides dense token-level supervision by evaluating student-generated trajectories with a teacher model. However, a globally stronger teacher does not guarantee beneficial local supervision: teacher preferences can conflict with response outcomes, and student-generated prefixes can deviate from teacher-supported reasoning paths. We introduce *Prefix-Outcome Validation* (POV), a selective correction mechanism for dense OPD supervision. POV uses group-relative outcomes to identify conflicting supervision and relative teacher support across consecutive windows to regulate attenuation strength. Only outcome-conflicting supervision is attenuated, with lower prefix support inducing stronger attenuation; all other supervision remains unchanged. The correction requires no tunable mixing coefficient and preserves the sign of the original token-level supervision without increasing its magnitude. Across multiple student–teacher configurations and reasoning benchmarks, POV-OPD improves aggregate mean and best-of-k performance over matched raw OPD. On a benchmark where the teacher underperforms the student, POV-OPD also mitigates the loss of pre-distillation capability. Component ablations demonstrate the complementary value of outcome validation and prefix support. Further token-level analysis shows that POV's selective correction reaches key reasoning functions, including logical structuring, self-checking, and reasoning revision. Together, these findings highlight POV as a simple and effective approach to improving on-policy distillation: without additional process annotations, it uses response outcomes and teacher support to selectively weight dense supervision, reducing conflicting signals while retaining teacher guidance.

1 INTRODUCTION

On-policy distillation (OPD) provides dense supervision for training reasoning models: a student samples responses, and a teacher evaluates the sampled tokens on the prefixes actually visited by the student. The teacher–student log-probability gap supplies a local learning signal throughout each response. Unlike offline distillation on teacher-generated trajectories, OPD trains directly on the student's own reasoning paths. Yet dense supervision is not necessarily reliable supervision. A globally stronger teacher need not provide beneficial guidance at every local decision, and recent evidence (Li et al., 2026; Zhu et al., 2026; Yang et al., 2026b; Xing et al., 2026; Liu et al., 2026; Plyusov et al., 2026; Fu et al., 2026; Zhang et al., 2026) shows that OPD is sensitive to student–teacher compatibility and the distribution of visited states.

We examine this reliability problem from two complementary perspectives. The first is *outcome consistency*: does a local teacher preference agree with the response's group-relative outcome? Verifiable rewards, widely used in reasoning post-training methods such as GRPO Shao et al. (2024); DeepSeek-AI (2025), provide an external criterion for evaluating completed responses. Raw OPD does not consult this evidence: it can discourage sampled tokens in above-average responses or encourage them in below-average responses. More fundamentally, even a teacher with a higher success probability from the same prefix can favor an action with negative outcome advantage for the student, because the two models differ in their ability to complete the subsequent reasoning. A stronger teacher therefore does not guarantee that each local signal helps the student.

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.

The second perspective is *prefix support*: how strongly does the teacher support the student's current trajectory? OPD queries the teacher on student-generated contexts, which can depart from teacher-supported reasoning paths. The teacher still produces a distribution at these prefixes, but the teacher–student probability gap alone does not reveal the level of teacher support: a token can have a positive gap even when both models assign it low probability. Outcome consistency and prefix support thus address different questions—whether supervision conflicts with observed results, and how much trust to place in that supervision along the current trajectory.

We propose *Prefix-Outcome Validation* (POV), a lightweight, selective correction for dense OPD supervision. Outcome conflict determines *whether* to attenuate a signal, while prefix support regulates *how strongly* to attenuate it. POV detects directional conflict using each response's reward relative to the group mean. It estimates prefix support by converting teacher excess surprisal into bounded token scores and averaging them over consecutive windows. Window averaging smooths individual-token fluctuations to obtain a more stable local support estimate. Each window is compared with the highest mean support observed up to and including that window, allowing relative support to decrease after a drop and recover later.

POV attenuates only outcome-conflicting supervision, with lower prefix support producing stronger attenuation; all other supervision remains unchanged. The rule preserves the sign of the original OPD signal without increasing its magnitude, and requires neither a triggering threshold nor a tunable mixing coefficient. OPD remains the primary learning signal, rather than being supplemented with a separate GRPO gradient. Without additional process annotations, POV uses response outcomes and teacher support to reduce conflicting signals while retaining dense teacher guidance.

Our contributions are threefold:

- We show why a stronger teacher need not provide beneficial local guidance, and identify outcome consistency and prefix support as complementary checks on dense supervision.
- We introduce POV, a threshold-free correction that uses outcome conflict to select supervision for attenuation and prefix support to set its strength. Its only structural hyperparameter is the window size, and it preserves the OPD signal's sign without increasing its magnitude.
- Across multiple distillation settings, we demonstrate the effectiveness of POV-OPD over matched raw OPD. Component ablations and token-level analysis further clarify the complementary roles of outcome and prefix signals and reveal where the correction acts during reasoning.

2 Related Work

Outcome-based reinforcement learning. RLHF commonly optimizes a policy against learned or programmatic rewards while constraining policy drift Christiano et al. (2017); Schulman et al. (2017); Ouyang et al. (2022). For mathematical reasoning, automatically verifiable final answers make outcome rewards especially attractive Shao et al. (2024); DeepSeek-AI (2025). GRPO removes the learned value model by normalizing rewards among responses sampled for the same prompt and broadcasting the resulting scalar over each response. Such estimators can optimize expected terminal success, but a single rollout provides no token-specific evidence about which intermediate decisions deserve credit. POV instead uses the group-relative outcome as a validator of dense teacher supervision, not as an additive policy-gradient reward.

Knowledge and on-policy distillation. Knowledge distillation transfers predictions from a stronger teacher to a student Hinton et al. (2015). In sequence generation, offline imitation can suffer from state-distribution mismatch because the student is not trained on the prefixes it will induce at inference time. OPD addresses this problem by sampling from the student and querying the teacher on student-visited prefixes. Recent analysis emphasizes both the promise of this dense signal and its dependence on teacher strength, student compatibility, and rollout states Li et al. (2026). POV leaves the OPD rollout and teacher-scoring pipeline unchanged but questions the stronger assumption that every local gap should receive equal trust.

Credit assignment and process supervision. Chain-of-thought, self-consistency, and self-training methods demonstrate that intermediate reasoning paths strongly affect final accuracy Wei

et al. (2022); Wang et al. (2022); Zelikman et al. (2022). Process supervision directly labels intermediate steps, but collecting or generating reliable process labels is costly and can introduce its own errors Lightman et al. (2023). POV requires neither step annotations nor an additional process reward model. It combines a terminal verifier with teacher probabilities already computed by OPD, using prefix support only to decide how strongly an outcome conflict should attenuate dense credit.

3 Reliability Gaps in On-Policy Distillation

Raw OPD turns teacher–student probability gaps into dense token-level supervision. A larger teacher probability indicates a teacher preference, not necessarily a beneficial decision for the student. We first establish this distinction, then examine two complementary aspects of supervision: consistency with response outcomes and teacher support for the student-generated trajectory.

3.1 Teacher Superiority Does Not Guarantee Beneficial Local Guidance

Consider a shared history h and a token action a. Let $Q^{R,\pi}(h,a)$ denote the expected terminal reward after taking a and continuing with policy π, and let $V^{R,\pi}(h) = \mathbb{E}_{a\sim\pi(\cdot|h)}[Q^{R,\pi}(h,a)]$. The raw OPD signal and the student's outcome advantage are

$$\begin{aligned} D(h,a) &= \log \frac{\pi_T(a \mid h)}{\pi_\theta(a \mid h)}, \\ A^{R,\pi_\theta}(h,a) &= Q^{R,\pi_\theta}(h,a) - V^{R,\pi_\theta}(h). \end{aligned} \tag{1}$$

Their local directions conflict when $D(h,a)A^{R,\pi_\theta}(h,a) < 0$.

Proposition 1. Even if the teacher has a higher success probability than the student from the same history, a positive OPD signal can favor an action with negative student outcome advantage.

Proof. Consider a two-step decision tree with actions a and b at history h, followed by a terminal decision with binary reward. Choose arbitrary parameters $0 < p < q < 1$ and $0 < v < u < w < 1$. Let $\pi_\theta(b \mid h) = p$ and $\pi_T(b \mid h) = q$, with the remaining probability assigned to a. Specify the continuation policies so that

$$\begin{aligned} Q^{R,\pi_\theta}(h,a) &= Q^{R,\pi_T}(h,a) = u, \\ Q^{R,\pi_\theta}(h,b) &= v, \qquad Q^{R,\pi_T}(h,b) = w. \end{aligned} \tag{2}$$

These values are realized in the same tree with the same reward function by assigning the corresponding probabilities to the successful terminal action. Thus, the teacher and student have equal continuation success after a, but the teacher is more capable after b and selects b more often:

$$\begin{aligned} V^{R,\pi_T}(h) &= u + q(w-u) > u, \\ V^{R,\pi_\theta}(h) &= u - p(u-v) < u. \end{aligned} \tag{3}$$

Hence $V^{R,\pi_T}(h) > V^{R,\pi_\theta}(h)$, but action b has conflicting OPD and student-outcome signals:

$$\begin{aligned} D(h,b) &= \log \frac{q}{p} > 0, \\ A^{R,\pi_\theta}(h,b) &= v - \big[(1-p)u + pv\big] = -(1-p)(u-v) < 0. \end{aligned} \tag{4}$$

Taking h as the initial history of a single-prompt task also makes the teacher globally stronger, while preserving this local conflict. □

The conflict holds throughout the stated parameter ranges and arises from different continuation capabilities: an action that the teacher can execute successfully need not be a good choice for the student's subsequent reasoning. Thus, teacher superiority alone does not guarantee that each dense probability gap is aligned with the student's task objective.

3.2 Outcome Consistency: Checking Teacher Preferences Against Results

The OPD gap compares two policies' preferences without explicitly checking the verified outcome of the sampled response. To introduce this missing task-level evidence, consider G student responses to the same prompt, with outcomes $R_i \in \{0, 1\}$ and group-relative performance $g_i = R_i - G^{-1} \sum_{j=1}^{G} R_j$. We call a token's supervision *outcome-conflicting* when

$$g_i D_{i,t} < 0. \tag{5}$$

This covers two cases: a negative OPD signal suppresses a token in a relatively successful response, or a positive signal promotes a token in a relatively unsuccessful response. Raw OPD applies these signals without distinguishing them from signals that agree with the response outcome.

This is an observable response-level consistency check, rather than an identification of A^{R,π_θ}: a successful response can contain mistakes, and an unsuccessful response can contain useful steps. Outcome evidence motivates checking dense supervision, not replacing token-level credit with the terminal reward. Homogeneous groups provide no comparative evidence ($g_i = 0$).

3.3 Prefix Support: Assessing Guidance Along the Student Trajectory

Outcome consistency does not describe how strongly the teacher supports the trajectory on which its guidance is queried. OPD conditions the teacher on student-generated histories, which can depart from the paths the teacher typically produces. Moreover, $D_{i,t}$ is a probability ratio: it can be positive even when both policies assign very little probability to the sampled token. The gap alone therefore does not characterize teacher support for the current trajectory.

Here, prefix support refers to compatibility with the teacher's distribution, not the correctness of the partial solution. A low-support trajectory may use an unfamiliar but valid solution, while a teacher-supported trajectory may still fail. Outcome consistency and prefix support thus describe distinct aspects of supervision and should not be treated as interchangeable correctness tests.

Implication for POV. These observations motivate a division of roles: outcome conflict selects supervision for correction, and prefix support modulates the strength of that correction. POV measures relative changes in teacher support across consecutive windows of the student rollout and applies stronger attenuation to conflicting signals when support drops. A support drop alone does not trigger attenuation, and outcome-compatible supervision remains unchanged. Section 4 defines this selective weighting; component ablations test the combined signals.

4 Prefix-Outcome Validation

4.1 Setup and Outcome Conflict

For prompt x, the student π_θ samples G responses y_i with verified outcomes $R_i \in \{0, 1\}$. At each student history $h_{i,t} = (x, y_{i,<t})$, the teacher and student score the sampled token:

$$D_{i,t} = \log \pi_T(y_{i,t} \mid h_{i,t}) - \log \pi_\theta(y_{i,t} \mid h_{i,t}). \tag{6}$$

Figure 1 shows how POV combines this signed OPD gap with response outcomes and prefix support.

We center outcomes within each prompt group and normalize the token gap:

$$g_i = R_i - \frac{1}{G} \sum_{j=1}^{G} R_j, \qquad \widetilde{D}_{i,t} = \tanh\left(\frac{D_{i,t}}{\mathrm{median}_{s \le T_i} |D_{i,s}| + \varepsilon} \right). \tag{7}$$

This bounded normalization preserves the sign of $D_{i,t}$ and makes conflict strength less sensitive to the gap scale of a particular student–teacher pair. The conflict score is

$$c_{i,t} = [-g_i \widetilde{D}_{i,t}]_+ \in [0, 1]. \tag{8}$$

A negative gap on a relatively successful response or a positive gap on a relatively unsuccessful response triggers correction. Thus g_i supplies the response-level direction, while $\widetilde{D}_{i,t}$ supplies the local teacher–student preference and its normalized strength. When group outcomes are identical, $g_i = 0$, and POV leaves every token's OPD guidance unchanged. Since $D_{i,t}$ varies across tokens while g_i is shared, POV corrects only conflicting positions.

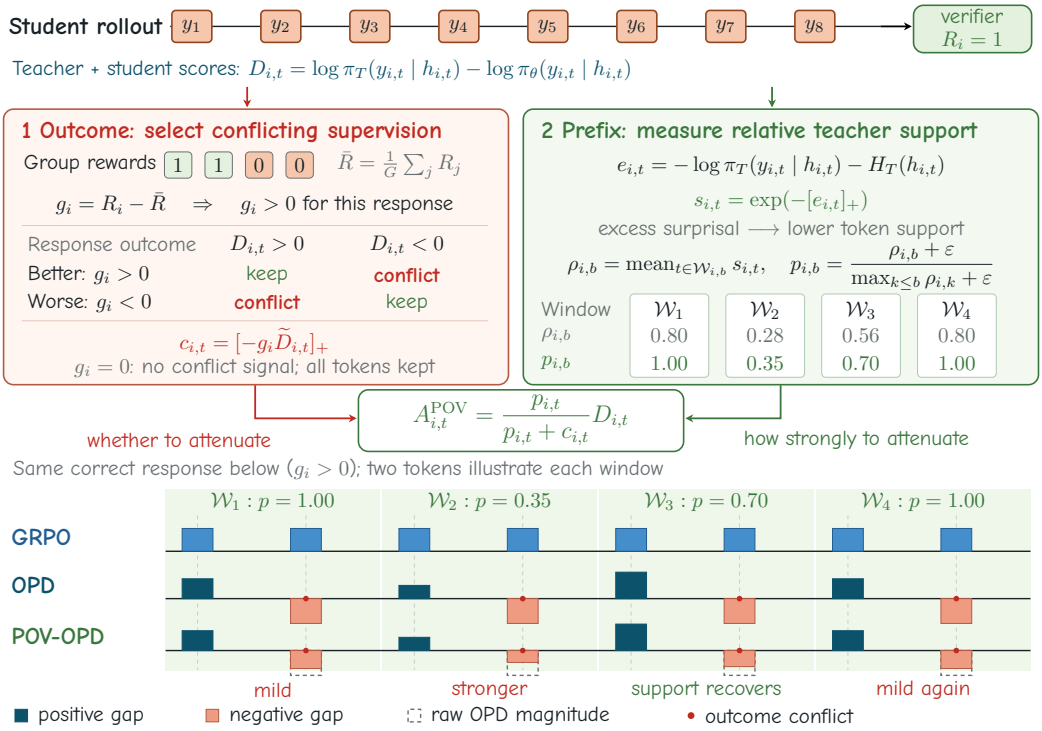


Figure 1: POV-OPD separates conflict selection (left) from attenuation strength (right). Group-relative outcomes identify conflicts using normalized gaps $\widetilde{D}_{i,t}$; teacher excess surprisal yields window-level support relative to the running best. Bottom: the same correct response under GRPO, OPD, and POV-OPD. Equal conflicting gaps receive stronger attenuation when support drops; compatible gaps remain unchanged. Dashed outlines show raw OPD magnitudes on the same scale as POV-OPD. Values are illustrative; each window shows two representative tokens.

4.2 Relative Prefix Support

The teacher is evaluated on student-generated histories. To measure its support for a sampled token relative to its own uncertainty, we subtract the teacher entropy from that token's surprisal:

$$e_{i,t} = -\log \pi_T(y_{i,t} \mid h_{i,t}) - H_T(h_{i,t}), \qquad s_{i,t} = \exp(-[e_{i,t}]_+)\,, \tag{9}$$

where $H_T(h) = -\sum_a \pi_T(a \mid h) \log \pi_T(a \mid h)$. The score is one unless token surprisal exceeds teacher entropy; only the excess lowers support. We average these scores within consecutive windows $\mathcal{W}_{i,b}$ of size W, reducing fluctuations from individual tokens. Each window is compared with the strongest window so far:

$$\rho_{i,b} = \frac{1}{|\mathcal{W}_{i,b}|} \sum_{t \in \mathcal{W}_{i,b}} s_{i,t}, \qquad p_{i,b} = \frac{\rho_{i,b} + \varepsilon}{\max_{k \le b} \rho_{i,k} + \varepsilon}. \tag{10}$$

Each token in window b uses $p_{i,t} = p_{i,b}$, so the estimate responds to overall window support rather than each surprising token. The running maximum makes support relative to the same response, so it can fall and recover without a fixed drift threshold. Here $p_{i,b} = 1$ marks the strongest support so far; lower values do not imply an incorrect partial solution.

4.3 Selective Attenuation

POV uses outcome conflict to select tokens for correction and prefix support to set the strength:

$$A^{\mathrm{POV}}_{i,t} = \frac{p_{i,t}}{p_{i,t} + c_{i,t}} D_{i,t}. \tag{11}$$

For every valid token, $p_{i,t} > 0$ and $c_{i,t} \geq 0$, so the multiplier $p_{i,t}/(p_{i,t} + c_{i,t})$ lies in $(0, 1]$. Hence

$$\operatorname{sign}(A^{\mathrm{POV}}_{i,t}) = \operatorname{sign}(D_{i,t}), \qquad |A^{\mathrm{POV}}_{i,t}| \leq |D_{i,t}|. \tag{12}$$

When $c_{i,t} = 0$, raw OPD is recovered. Otherwise, $c_{i,t}/(p_{i,t} + c_{i,t})$ is the fraction of the raw gap removed, so lower support strengthens attenuation only for conflicting tokens. The detached advantage enters the usual clipped OPD objective; the full objective appears in Appendix B.

5 Experiments

We evaluate whether POV improves reasoning performance over dense OPD under matched training settings. We compare benchmark accuracy across student–teacher configurations and examine where the correction acts during training.

5.1 Experimental Setup

Training data and models. All runs use DAPO-Math-17K (Yu et al., 2025). Our students are DeepSeek-R1-Distill-Qwen-1.5B, DeepSeek-R1-Distill-Qwen-7B (DeepSeek-AI, 2025), and Qwen3-4B initialized from Qwen3-4B-Base (Yang et al., 2025). The 1.5B student is distilled from JustRL-DeepSeek-1.5B (He et al., 2025a) or Skywork-OR1-Math-7B (He et al., 2025b); the 7B student uses the Skywork teacher. The Qwen3-4B student is distilled from a stronger same-architecture teacher trained with GRPO on DAPO-Math-17K.

Baselines and training protocol. We compare POV-OPD with raw sampled-token OPD under matched student initialization, teacher, training data, rollout budget, and optimizer settings. Table 1 also includes Prune-OPD and OPRD. Prune-OPD (Yang et al., 2026b) detects prefix drift through teacher–student top-k token overlap, using the detected drift to down-weight subsequent supervision and dynamically truncate rollouts. OPRD (Yang et al., 2026a) distills hidden-state representations across selected layers on student-generated trajectories, providing direct representation-level supervision that avoids the token-level Monte Carlo estimation variance of output-space supervision. For POV-OPD, we set $W = 128$ throughout. We train the 1.5B models for one epoch before evaluating their reasoning performance. Appendix A gives the training settings and 7B evaluation protocol.

Evaluation. Our evaluation protocol covers all student, teacher, and distilled models on AMC23 (Mathematical Association of America, n.d.), AIME 2024, AIME 2025 (Mathematical Association of America, 2025), and GPQA-Diamond (Rein et al., 2023). We sample 16 responses per problem under shared decoding and scoring settings within each study. For R1-based models, we report Mean@8 (also denoted Avg@8) and Best@8 using size-eight bootstrap draws with replacement. For Qwen3-4B, we report Mean@16 and Best@16. Macro scores equally weight the four benchmarks. Metric and checkpoint details appear in Appendix A.

5.2 Main Results

Selective attenuation improves post-training performance. Table 1 shows that POV-OPD outperforms matched raw OPD on most benchmark–metric pairs and achieves the highest four-benchmark macro Mean@8 and Best@8 among all four distillation methods in each R1-based configuration. Prune-OPD addresses unreliable prefixes through overlap-based pruning, whereas OPRD adds representation-level supervision; neither delivers the same consistent aggregate improvement in these settings. POV instead uses response outcomes to identify conflicting token-level directions and prefix support to control their attenuation, while retaining compatible dense supervision. Its macro lead across both 1.5B teachers and the 7B student highlights the value of this joint correction. Section 5.4 further examines the reasoning contexts in which the correction acts.

POV safeguards student capability under weaker teachers. In the 7B experiment, the teacher is weaker than the original student on GPQA. Raw OPD degrades the student's performance, whereas POV-OPD remains close to the pre-distillation level on both metrics. This contrast highlights the risk of transferring teacher preferences without accounting for capabilities the student already possesses.

Table 1: Matched distillation comparisons (Mean@8 / Best@8, %). Gray student and teacher rows are references only, separated from each matched comparison by a horizontal rule. Bold marks the highest score among the compared distillation methods within each student–teacher configuration, separately for Mean@8 and Best@8; all ties are bold. Each configuration compares OPD, OPRD, Prune-OPD, and POV-OPD. Macro is the mean of the four displayed benchmark scores.

Model / teacher	Role / method	AMC23	AIME24	AIME25	GPQA	Macro
DS-Qwen-1.5B distilled from JustRL-1.5B						
DS-Qwen-1.5B	Student	63.8 / 85.1	27.7 / 53.4	22.1 / 38.2	34.9 / 80.5	37.1 / 64.3
JustRL-DeepSeek-1.5B	Teacher	83.1 / 92.1	53.8 / 75.8	36.9 / 55.3	41.0 / 73.3	53.7 / 74.1
DS-Qwen-1.5B	OPD	77.5 / 90.4	48.1 / 70.2	33.8 / 48.7	37.0 / 78.4	49.1 / 71.9
DS-Qwen-1.5B	OPRD	78.5 / **91.6**	51.3 / 70.0	32.5 / 46.7	37.2 / 75.1	49.9 / 70.9
DS-Qwen-1.5B	Prune-OPD	75.1 / 85.5	49.6 / 68.4	31.3 / 40.0	34.4 / **81.2**	47.6 / 68.8
DS-Qwen-1.5B	**POV-OPD**	**81.4** / 90.8	**51.5 / 73.8**	**36.0 / 53.9**	**39.2** / 77.6	**52.0 / 74.0**
DS-Qwen-1.5B distilled from Skywork-7B						
DS-Qwen-1.5B	Student	63.8 / 85.1	27.7 / 53.4	22.1 / 38.2	34.9 / 80.5	37.1 / 64.3
Skywork-OR1-Math-7B	Teacher	89.2 / 94.5	67.7 / 81.1	53.1 / 71.3	46.1 / 73.4	64.0 / 80.1
DS-Qwen-1.5B	OPD	70.3 / 88.4	37.7 / 66.9	**28.5** / 43.5	42.8 / 77.2	44.8 / 69.0
DS-Qwen-1.5B	OPRD	68.1 / 87.4	36.2 / **69.8**	27.1 / 41.1	39.3 / **78.1**	42.7 / 69.1
DS-Qwen-1.5B	Prune-OPD	68.6 / 88.0	37.9 / 66.7	27.2 / 43.1	41.5 / 75.8	43.8 / 68.4
DS-Qwen-1.5B	**POV-OPD**	**72.4 / 90.1**	**39.2** / 69.2	28.1 / **44.2**	**44.3** / 77.4	**46.0 / 70.2**
DS-Qwen-7B distilled from Skywork-7B						
DS-Qwen-7B	Student	82.2 / 93.2	53.1 / 79.2	39.4 / 60.5	49.1 / 80.7	56.0 / 78.4
Skywork-OR1-Math-7B	Teacher	89.2 / 94.5	67.7 / 81.1	53.1 / 71.3	46.1 / 73.4	64.0 / 80.1
DS-Qwen-7B	OPD	**88.5** / 94.6	64.8 / 80.5	47.9 / 65.1	46.2 / 76.5	61.9 / 79.2
DS-Qwen-7B	OPRD	88.1 / 93.4	64.2 / 79.8	50.0 / 65.1	46.3 / 69.1	62.2 / 76.9
DS-Qwen-7B	Prune-OPD	84.6 / 92.8	60.8 / 80.0	44.2 / 63.3	46.1 / **80.2**	58.9 / 79.1
DS-Qwen-7B	**POV-OPD**	87.8 / **94.9**	**67.3 / 82.7**	**51.9 / 69.6**	**48.6** / 80.1	**63.9 / 81.8**

POV uses outcome evidence to check dense teacher guidance and attenuates conflicting supervision rather than accepting it unconditionally. The outcome signal thus provides a safeguard against weaker teacher guidance: it identifies conflicts between teacher preferences and response outcomes, allowing POV to attenuate conflicting supervision and protect capabilities the student already has.

A stronger teacher does not necessarily yield a stronger student. For the same DS-Qwen-1.5B student, Skywork-7B outperforms JustRL-1.5B on all three mathematics benchmarks, yet the JustRL-distilled student scores higher on all three under both raw OPD and POV-OPD. Teacher benchmark accuracy is therefore insufficient to predict distillation performance. The results suggest that the suitability of teacher supervision for the student's learning process matters alongside the teacher's standalone capability. Supervision reliability thus matters alongside teacher strength.

5.3 QWEN3-4B SCALE TRANSFER

Motivated by the teacher–student compatibility pattern in Section 5.2, we distill a fresh Qwen3-4B student from a GRPO-trained teacher of the same architecture.

Table 2 shows that POV-OPD improves Mean@16 over raw OPD on all four benchmarks and Best@16 on AIME24, AIME25, and GPQA. Both four-benchmark macro metrics improve. These gains show that POV's benefit extends beyond R1-based models. The Qwen3-4B teacher and student share an architecture and originate from the same base model, a setting in which their knowledge and reasoning trajectories might be expected to overlap substantially. Yet this shared origin does not guarantee that every local teacher preference benefits the student. Without outcome validation, raw OPD can reinforce conflicting local preferences, steering the student's updates away from its task objective and leading to poorer post-training performance. By attenuating such outcome-conflicting token-level supervision, POV-OPD reduces misaligned credit, partially counteracts the influence of these misleading local preferences, and achieves better aggregate performance than raw OPD.

Table 2: Qwen3-4B scale-transfer results (Mean@16 / Best@16, %). Base and GRPO rows provide student and teacher references. Within the OPD/POV-OPD pair, bold marks the higher metric value.

Model / teacher	Role / method	AMC23	AIME24	AIME25	GPQA
Qwen3-4B-Base	Student	14.3 / 54.2	4.6 / 22.1	1.9 / 14.7	3.8 / 27.8
Qwen3-4B GRPO	Teacher	40.0 / 71.2	12.7 / 38.6	11.0 / 36.7	14.1 / 57.0
Qwen3-4B	OPD	38.6 / **75.5**	9.8 / 31.3	9.2 / 27.1	9.9 / 51.1
Qwen3-4B	**POV-OPD**	**39.1** / 74.7	**12.1 / 34.5**	**11.3 / 31.7**	**11.3 / 54.1**

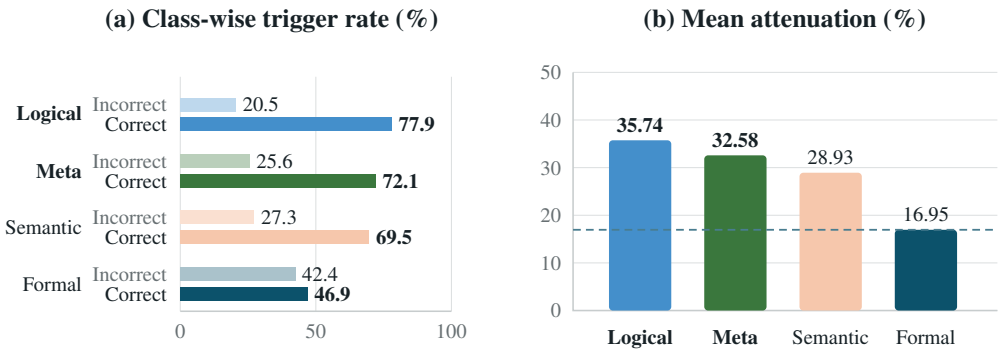


Figure 2: POV correction by token function. (a) Within-class attenuation-trigger rate in verifier-incorrect and verifier-correct responses. (b) Mean attenuation among triggered tokens; the dashed line marks the formal-reasoning mean.

5.4 Token-Function Analysis

To examine how POV changes token-level credit assignment, we distinguish four token functions: formal reasoning, logical structuring, metacognitive activity, and semantic support. We use AI-assisted, rule-based contextual annotation to label training trajectories. Examples and annotation details appear in Appendix C.

Outcome-stratified trigger frequency. Figure 2(a) reports the fraction of tokens in each class that trigger attenuation, separately for verifier-correct and verifier-incorrect responses. Formal-reasoning tokens have similar trigger rates in the two groups, whereas logical, metacognitive, and semantic-support tokens trigger much more often in correct responses. At those positions, raw OPD would suppress tokens in successful responses; POV weakens that conflicting negative credit and protects reasoning transitions and supporting language the teacher disfavors.

Teacher–student mismatch at reasoning transitions. Figure 2(b) measures attenuation strength *among triggered tokens*. Logical, metacognitive, and semantic-support tokens receive mean attenuation of 35.74%, 32.58%, and 28.93%, respectively, well above the 16.95% for formal reasoning. Logical connectors express causal relations, comparisons, and scope, while metacognitive expressions initiate checking and replanning. These transitions are especially sensitive to a mismatch between the teacher's local preference and the student's ability to continue. On a successful trajectory, the teacher may assign lower probability to a token outside its preferred reasoning path, even though the student still reaches the right answer; raw OPD would suppress that choice. On a failed trajectory, the teacher may favor a step that its own continuation could repair, although the student cannot make the same repair; raw OPD would reinforce the step, potentially turning the teacher's preference into a reasoning trap. POV weakens both types of outcome-conflicting guidance, especially when relative prefix support drops, so the teacher's local preferences have less influence on how the student connects, checks, and revises later steps.

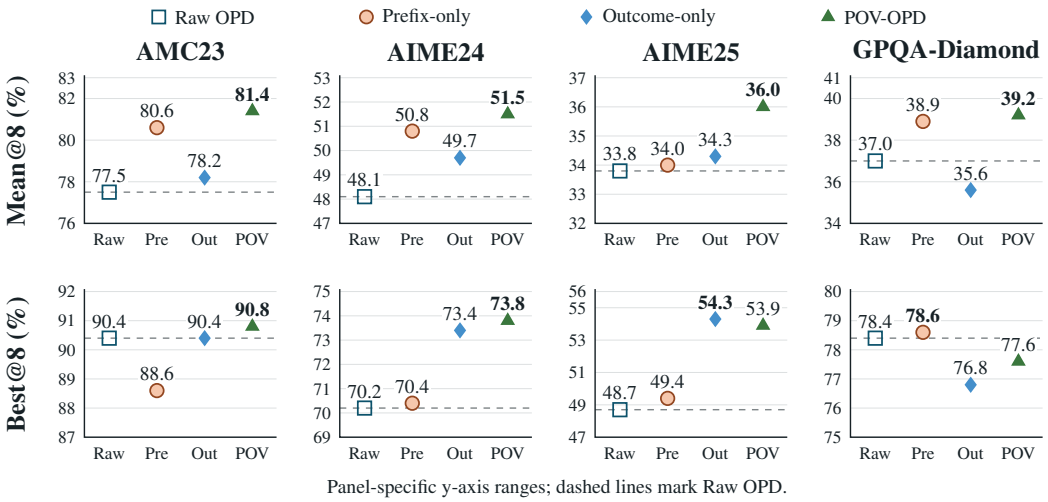


Figure 3: Component ablation across four benchmarks (columns) and Mean@8/Best@8 (rows), with expanded panel-specific linear y-axes. Pre/Out denote prefix-only/outcome-only; dashed lines mark raw OPD. Labels give exact scores (panel maxima in bold). Definitions and full results appear in Appendix A.1.

5.5 Component Ablation

Figure 3 compares raw OPD, prefix-only, outcome-only, and full POV. The ablations support the complementary value of outcome and prefix information: full POV achieves the highest Mean@8 on all four benchmarks, outperforming the other three methods.

The joint design assigns distinct roles to the two signals. Outcome conflict determines whether supervision should be attenuated, preserving compatible OPD signals, while prefix support adapts attenuation strength to the teacher's support for the current trajectory. The Mean@8 gains over outcome-only highlight the added value of prefix-dependent attenuation; the gains over prefix-only further support combining prefix information with outcome-based conflict detection.

The two single-component variants exhibit different patterns in Figure 3. Prefix-only improves Mean@8 over raw OPD on all four benchmarks, showing the value of adapting supervision to local teacher support. Outcome-only yields especially clear Best@8 gains on AIME24 and AIME25, showing that response-level conflict supplies a distinct signal. Full POV combines these roles: it leads on Mean@8 in every benchmark and improves Best@8 in all three math tasks.

The largest Best@8 gains over raw OPD occur on AIME24 and AIME25; variant definitions and full results appear in Appendix A.1.

6 Conclusion

OPD supplies dense supervision for long reasoning chains, but teacher preferences are not uniformly reliable along the prefixes a student actually visits. POV treats this as a two-part validation problem: response outcomes reveal when token-level guidance conflicts with observed success, while relative teacher support determines how strongly that guidance is attenuated. Compatible supervision remains intact, without process annotations or a separate training objective. Across the evaluated distillation settings, POV-OPD improves macro-averaged mean and best-of-k performance over matched raw OPD and better preserves existing student capability when the teacher is weaker on a task. Ablations confirm the complementary value of the two signals, while token-level analysis shows that the correction extends to logical transitions and self-checking. These findings suggest that reliable OPD depends not only on teacher strength, but on whether local supervision agrees with observed outcomes and remains supported along the student's trajectory.

AI USE STATEMENT

Generative AI assisted with manuscript writing, translation, and contextual token annotation. The authors take responsibility for the final content.

REFERENCES

Paul F. Christiano, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences, 2017.

DeepSeek-AI. DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning, 2025.

Zixuan Fu, Bingxiang He, Yuxin Zuo, Haohuan Huang, Jinqian Zhang, Ruhang Xiao, Cheng Qian, Qinyu Luo, Huan-ang Gao, Yudong Wang, Zhiyuan Liu, Ning Ding, and Chaojun Xiao. Rethinking on-policy distillation of large language models II: One training example, 2026. URL https://arxiv.org/abs/2609.04172.

Bingxiang He, Zekai Qu, Zeyuan Liu, Yinghao Chen, Yuxin Zuo, Cheng Qian, Kaiyan Zhang, Weize Chen, Chaojun Xiao, Ganqu Cui, Ning Ding, and Zhiyuan Liu. JustRL: Scaling a 1.5b LLM with a simple RL recipe, 2025a.

Jujie He, Jiacai Liu, Chris Yuhao Liu, Rui Yan, Chaojie Wang, Peng Cheng, Xiaoyu Zhang, Fuxiang Zhang, Jiacheng Xu, Wei Shen, Siyuan Li, Liang Zeng, Tianwen Wei, Cheng Cheng, Bo An, Yang Liu, and Yahui Zhou. Skywork Open Reasoner 1 technical report, 2025b.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network, 2015.

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

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step, 2023.

Yanjiang Liu, Jie Lou, Xinyan Guan, Yuqiu Ji, Hongyu Lin, Ben He, Xianpei Han, Le Sun, Xing Yu, and Yaojie Lu. Your teacher can't help you here: Combating supervision fidelity decay in on-policy distillation, 2026. URL https://arxiv.org/abs/2605.30833.

Mathematical Association of America. MAA invitational competitions. https://maa.org/maa-invitational-competitions/, 2025.

Mathematical Association of America. American mathematics competitions. https://maa.org/student-programs/amc/, n.d. Accessed September 7, 2026.

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

Daniil Plyusov, Alexey Gorbatovski, Alexey Malakhov, Nikita Balagansky, Boris Shaposhnikov, Daria Korotyshova, and Daniil Gavrilov. Trust-region behavior blending for on-policy distillation, 2026. URL https://arxiv.org/abs/2605.31159.

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, 2023.

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

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

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models, 2022.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models, 2022.

Xingrun Xing, Haoqing Wang, Boyan Gao, Ziheng Li, and Yehui Tang. Trust region on-policy distillation, 2026. URL https://arxiv.org/abs/2606.01249.

An Yang et al. Qwen3 technical report, 2025.

Shenzhi Yang, Guangcheng Zhu, Bowen Song, Haobo Wang, Mingxuan Xia, Xing Zheng, Yingfan Ma, Zhongqi Chen, Weiqiang Wang, Junbo Zhao, and Gang Chen. OPRD: On-policy representation distillation, 2026a. URL https://arxiv.org/abs/2606.06021.

Zhicheng Yang, Zhijiang Guo, Yifan Song, Minrui Xu, Yongxin Wang, Yiwei Wang, Xiaodan Liang, and Jing Tang. Prune-OPD: Efficient and reliable on-policy distillation for long-horizon reasoning, 2026b. URL https://arxiv.org/abs/2605.07804.

Qiying Yu et al. DAPO: An open-source LLM reinforcement learning system at scale, 2025.

Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah D. Goodman. STaR: Bootstrapping reasoning with reasoning, 2022.

Ke Zhang, Yunjie Tian, Dongdi Zhao, Yijiang Li, Yuanye Liu, Vishal M. Patel, and Di Fu. On-policy distillation with Best-of-N teacher rollout selection, 2026. URL https://arxiv.org/abs/2605.09725.

Siqi Zhu, Xuyan Ye, Hongyu Lu, Weiye Shi, and Ge Liu. The many faces of on-policy distillation: Pitfalls, mechanisms, and fixes, 2026. URL https://arxiv.org/abs/2605.11182.

A Experimental Details

Training configuration. Table 3 lists the settings shared by raw OPD and POV-OPD in the 1.5B R1-based and Qwen3-4B studies. These distillation runs use DAPO-Math-17K, sampled-token OPD, and a fixed teacher. Within each comparison, the student initialization, training data, rollout budget, and optimizer settings are shared. POV uses $W = 128$.

Table 3: Training and evaluation settings for the 1.5B R1-based and Qwen3-4B distillation studies. Lengths are measured in tokens.

Setting	R1-1.5B	Qwen3-4B
Training		
Training epochs	1	1
Prompts per batch	64	128
Responses per prompt	4	8
Maximum prompt length	1,024	1,024
Maximum response length	8,196	4,096
Learning rate	10^{-6}	10^{-6}
Rollout temperature	1.0	1.0
Actor / teacher model dtype	FP32	BF16
Rollout dtype	BF16	BF16
POV window size W	128	128
Actor KL penalty	None	None
Number of GPUs	8	8
Evaluation		
Responses per problem	16	16
Maximum response length	31,744	4,096
Sampling temperature	0.7	1.0
Top-p	0.95	0.95
Reported mean / best metrics	Mean@8 / Best@8	Mean@16 / Best@16

Evaluation protocol. All models are evaluated on AMC23, AIME 2024, AIME 2025, and GPQA-Diamond with 16 responses per problem and top-$p = 0.95$. Decoding settings are listed in Table 3. Within each study, methods share prompts and answer-verification rules. For the R1-based GPQA evaluations, we extract the final explicit A–D choice from a boxed answer or an answer-labelled statement; missing choices count as incorrect.

For R1-based models, Mean@8 (Avg@8) and Best@8 are defined using eight draws with replacement from the 16 saved responses. Let $z_{q,j} \in \{0, 1\}$ denote the correctness of response j to problem q, and let $\hat{p}_q = \frac{1}{16}\sum_{j=1}^{16} z_{q,j}$. For a benchmark with N problems, the exact expectations under this empirical resampling rule are

$$\text{Mean@8} = \frac{1}{N}\sum_{q=1}^{N} \hat{p}_q, \qquad \text{Best@8} = \frac{1}{N}\sum_{q=1}^{N} \big[1 - (1 - \hat{p}_q)^8\big]. \tag{13}$$

Analytic evaluation computes the exact with-replacement expectations from the 16 saved responses, eliminating Monte Carlo error in the resampling step.

Qwen3-4B uses temperature 1.0, a 4,096-token response limit, and a Thinkless-style rule-based verifier extended to Qwen3/Qwen3.5 answer conventions. We report the evaluator's Mean@16 and Best@16 statistics. Pooled validation accuracy weights all sampled responses equally; macro averages give each benchmark equal weight. Method comparisons use the shared decoding budget and sample size of each study.

A.1 Component Ablation

We compare raw OPD, prefix-only, outcome-only, and full POV-OPD with a DS-Qwen-1.5B student and a JustRL-DeepSeek-1.5B teacher. The comparison uses the one-epoch training setup and

four-benchmark evaluation protocol described above. The variants differ in how they transform the sampled-token OPD gap $D_{i,t}$. Using the prefix support $p_{i,t}$ and outcome conflict $c_{i,t}$ defined in Section 4, their advantages are

$$A_{i,t}^{\mathrm{Raw}} = D_{i,t}, \qquad A_{i,t}^{\mathrm{Prefix}} = \lambda_{\mathcal{B}} p_{i,t} D_{i,t},$$
$$A_{i,t}^{\mathrm{Outcome}} = \frac{D_{i,t}}{1 + c_{i,t}}, \qquad A_{i,t}^{\mathrm{POV}} = \frac{p_{i,t}}{p_{i,t} + c_{i,t}} D_{i,t}. \tag{14}$$

Here $\mathcal{B}$ contains the valid response tokens in a training batch, with

$$\lambda_{\mathcal{B}} = \frac{\sum_{(i,t)\in\mathcal{B}} |D_{i,t}|}{\sum_{(i,t)\in\mathcal{B}} p_{i,t}|D_{i,t}|}.$$

A zero-mass batch retains raw OPD. Prefix-only is a normalized reweighting control: it redistributes token weights according to prefix support while preserving the batch's total absolute advantage mass. Individual token advantages can increase under this normalization. Outcome-only and full POV apply no batch-wide rescaling and leave nonconflicting tokens unchanged. Setting the outcome conflict to zero in full POV recovers raw OPD. Comparing full POV with outcome-only tests the added contribution of prefix-dependent attenuation.

Table 4: Component ablation with a DS-Qwen-1.5B student and a JustRL-DeepSeek-1.5B teacher (Mean@8 / Best@8, %). All variants share the same training and evaluation setup. The Raw OPD and POV-OPD rows reproduce the corresponding results in Table 1. Bold marks the highest score for each benchmark and metric.

Method	AMC23	AIME24	AIME25	GPQA-Diamond
Raw OPD	77.5 / 90.4	48.1 / 70.2	33.8 / 48.7	37.0 / 78.4
Prefix-only	80.6 / 88.6	50.8 / 70.4	34.0 / 49.4	38.9 / **78.6**
Outcome-only	78.2 / 90.4	49.7 / 73.4	34.3 / **54.3**	35.6 / 76.8
POV-OPD	**81.4 / 90.8**	**51.5 / 73.8**	**36.0** / 53.9	**39.2** / 77.6

Table 4 shows that full POV has the highest Mean@8 on all four benchmarks, supporting the complementary value of outcome conflict detection and prefix-dependent attenuation. Full POV also improves Best@8 over raw OPD on all three mathematics benchmarks and achieves the highest Best@8 among the four variants on AMC23 and AIME24. Together, these results favor the joint design for improving average accuracy while increasing multi-sample success on the mathematics benchmarks.

B POV Implementation Details

Outcome normalization. Equation (7) normalizes each response's OPD gaps by their median magnitude. The transformation bounds conflict strength while preserving the gap's sign. Homogeneous groups have $g_i = 0$ and provide no comparative outcome evidence, so their conflict scores are zero.

Teacher support at each token. Equation (9) gives a score of one when a sampled token is no more surprising than the teacher's average uncertainty. Its score decreases smoothly for unusually surprising tokens. Averaging within the windows of Equation (10) reduces local fluctuations. The running maximum includes the current window, so a new maximum has $p_{i,b} = 1$. Later windows may have lower support and subsequently recover. The relative construction needs no drift threshold.

Optimization and cost. The detached advantage in Equation (11) enters the same clipped policy objective as raw OPD:

$$\mathcal{L}_{\mathrm{POV}} = -\sum_{i,t} \min\big(r_{i,t} A_{i,t}^{\mathrm{POV}}, \; \mathrm{clip}(r_{i,t}, 1-\epsilon, 1+\epsilon) A_{i,t}^{\mathrm{POV}}\big), \tag{15}$$

where $r_{i,t} = \pi_\theta(y_{i,t} \mid h_{i,t}) / \pi_{\theta_{\mathrm{old}}}(y_{i,t} \mid h_{i,t})$. POV adds only linear-time reductions over scores already produced by OPD. Its only structural hyperparameter is the prefix window size W; the remaining constants are numerical stabilizers.

C Token-Function Annotation and Examples

Figure 2(b) uses three complete responses from the first training step of the DS-Qwen-1.5B–JustRL-1.5B run. Panel (a) uses 44 naturally completed responses from 12 prompts for which both logged verifier rewards, 0 and 1, occur: 28 responses have reward 0 and 16 have reward 1. Its class-conditional trigger rates are averaged equally across prompts. A token triggers attenuation when $c_{i,t} > 0$; the trigger rate uses all tokens assigned to that class as its denominator, whereas mean attenuation is computed only among triggered tokens.

The contextual functional labels are AI-assisted, rule-based annotations. They distinguish formal computation, logical organization, checking and replanning, and ordinary semantic support. The labels are used only for this diagnostic, not during POV training. Table 5 lists representative tokens; a token's function can depend on its context.

Table 5: Example tokens from the annotated responses, grouped by contextual function.

Function	**Example tokens**
Formal reasoning	`1`, `3`, `+`, `=`, `x`, `Subtract`
Logical structuring	`because`, `Therefore`, `as`, `for`
Metacognitive	`Wait`, `check`, `Alternatively`
Semantic support	`the`, `is`, `problem`, `distance`