Intervention-Certified Amplification: A Validity Gate for σ-Weighted Post-Training

Anonymous ACL submission

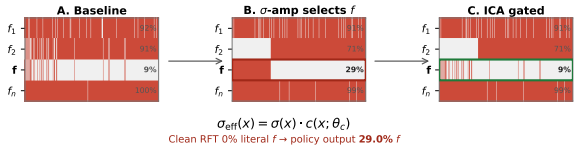

Figure 1: **Certify the prior before amplifying it.** On MATH500, σ-amplification reactivates channel f (TIR) from a 0%-literal-f teacher RFT to 45.8% emission. ICA gates amplification through $\sigma_{\text{eff}}(x) = \sigma(x)c(x;\theta_c)$ learned from paired rollouts.

Abstract

SFT/RL bridge and reward-filtered self-training methods (BREAD (Zhang et al., 2025), AESL (Li et al., 2026b), RAFT (Dong et al., 2023), STaR (Zelikman et al., 2022)) solve *when to imitate versus explore* but presuppose *what the imitated signal certifies*. Using **DIAA**, a per-sample σ=sigmoid$(-\beta\Delta)$ hybrid, we show that even after Python fences are removed from RFT, DIAA emits Tool-Integrated Reasoning (TIR) on 29% of MATH500 prompts ($3\times$ vanilla GRPO). A paired intervention decomposes GRPO $\rightarrow$ SFT $\rightarrow$ DIAA drag as $+2.2 \rightarrow +8.5 \rightarrow +10.6$ pp, identifying SFT imitation and σ-amplification as comparable sources. Adaptive amplification also reactivates pretrained behavior channels: a teacher RFT with 0% TIR still trains a student emitting TIR at 45.8%. We propose **Intervention-Certified Amplification (ICA)**: a certify token θ_c learned from paired rollouts gates the next round's per-sample weight, $\sigma_{\text{eff}}(x) = \sigma(x)c(x;\theta_c)$.

1 Introduction

Recent post-training methods invest much of their design budget in *when to imitate versus explore*: OPD (Thinking Machines, 2024; Li et al., 2026c) and RLSD (Hübotter et al., 2026) densify supervision; BREAD (Zhang et al., 2025), AESL (Li et al., 2026b), RAFT (Dong et al., 2023), and STaR (Zelikman et al., 2022) improve cold-start or reward-filtered self-training. All presuppose that the amplified signal is capability-bearing in the target environment. More precisely, post-training can amplify behaviors already present in the pretrained model, so validity must be certified in the target environment rather than inferred from literal training-data prevalence. This paper tests that presupposition.

Our test apparatus is **DIAA** (Sec. 2), a per-sample σ-weighted SFT/RL hybrid whose weight comes from a policy–anchor log-likelihood gap. We use it to isolate amplification, not as a headline accuracy claim. On Qwen2.5-1.5B math RL, DIAA emits TIR form on 29% of MATH500 prompts after Python fences are removed from RFT, a $3\times$ amplification over vanilla GRPO.

To isolate causal drag we use a *paired within-prompt intervention*: each checkpoint is evaluated under both a baseline math prompt and a positive-framing intervention. On the subset where the intervention removes TIR form (the rescued subset), vanilla GRPO shows a paired accuracy lift of $+2.2$ pp, pure SFT on the same RFT shows $+8.5$ pp, and DIAA shows $+10.6$ pp. This pathway decomposition splits cargo-cult drag into an SFT-contamination component ($+4.7$ pp from GRPO to SFT) and a σ-amplification component ($+5.3$ pp from SFT to DIAA).

ICA adds a certify token θ_c that gates the per-sample weight, $\sigma_{\text{eff}}(x) = \sigma(x)c(x;\theta_c)$. Paired-rollout deltas update θ_c, and the next round uses the resulting gate. ICA acts at the per-sample weight rather than changing the loss family or anchor source, so it composes with σ-weighted SFT/RL and reduces to the base when $c \equiv 1$.

1. Paired within-prompt intervention stress-tests the presupposition via GRPO $\rightarrow$ SFT $\rightarrow$ DIAA decomposition and a granularity dial signature ($1.5{-}1.7\times$ deeper weight-baking under truly-per-sample) (Sec. 4–4).

2. **ICA** is a certificate-gated per-sample weight modifier, with controls ruling out generic regularization and length-based heuristics (Sec. 5–6).

2 Background and DIAA

Terminology. **Rejection Fine-Tuning (RFT)** in the RAFT (Dong et al., 2023) / STaR (Zelikman et al., 2022) lineage is a reward-filtered dataset; here reward-1.0 samples form the anchor. **Tool-Integrated Reasoning (TIR) form** denotes a Python code fence followed by a hallucinated `output` block without actual execution. **σ-weighted SFT/RL** denotes $\sigma(x)\mathcal{L}_{\text{SFT}} + (1 - \sigma(x))\mathcal{L}_{\text{RL}}$ with per-sample $\sigma(x) \in [0, 1]$, where high σ favors imitation and low σ favors exploration. A **paired-intervention probe** evaluates the same checkpoint under suffixes $\tau_{\text{base}}, \tau_{\text{int}}$; within-prompt differences isolate the suffix's effect on a candidate trace feature.

Filter-by-reward self-RFT. We sample K rollouts per prompt, score them with a verifier, and keep top-reward traces as anchor RFT. This is sample-efficient but inherits verifier failures: a trace can score reward 1.0 by chance correlation between an unverified intermediate step and the final answer. With Qwen-Base + math RFT, 26.5% of reward-1.0 traces contain ```python ... ```output blocks without actual execution.

DIAA. DIAA's per-sample weight is the inverted Bayes-optimal discriminator between the RFT anchor π_{anchor} and the policy π_θ, giving the loss

$$\mathcal{L} = \sigma(x)\,\mathcal{L}_{\text{SFT}} + (1-\sigma(x))\,\mathcal{L}_{\text{RL}}, \quad \sigma(x) = \text{sigmoid}(-\beta\,\Delta(x)), \tag{1}$$

with $\Delta(x) = \log p_\theta(\text{rft}) - \log p_\theta(\text{roll})$ (full derivation in App. B). DIAA admits a per-prompt variant (σ shared across rollouts) and a truly-per-sample variant (σ per rollout); the choice has measurable consequences (Sec. 4). Two-time-scale convergence is established in App. B.3.

TIR and tool grounding. Lin and Xu (2025) formalize TIR's strict-expansion property under a deterministic external oracle. Without such an oracle, TIR form does not necessarily indicate tool capability: a σ-weighted procedure can amplify the surface form independent of whether the model can execute the code.

Checkpoint	Acc	TIR	RFT-TIR
vanilla GRPO	44.4	9.2	n/a
SFT (self)	40.6	17.2	26.5
DIAA (raw)	41.2	44.2	26.5
DIAA (clean)	47.5	29.0	0
SFT (teacher)	37.6	45.8	0

Table 1: MATH500 triangulation (all columns in %). DIAA amplifies TIR even after literal Python fences are removed; a clean teacher RFT (0% TIR) still yields 45.8% student TIR, indicating pretrained-channel reactivation rather than literal RFT imitation (App. C.7, C1).

3 Setup and Cross-Sectional Observation

Models and data. Qwen2.5-1.5B with LoRA following the TINA (Wang et al., 2026) setup; RFT is 94,865 self-generated samples on GSM8K+MATH train, filtered by boxed-answer match. Checkpoints (Tab. 1): vanilla GRPO; pure SFT on the same RFT; DIAA truly-per-sample on raw RFT (26.5% TIR contamination); DIAA truly-per-sample on RFT with literal Python code fences removed; pure SFT on a teacher-generated cross-model anchor (DeepSeek-Math-7B-Instruct teacher rollout, 0% TIR in the RFT).

DIAA on clean RFT outperforms vanilla GRPO by +3.1 pp, yet TIR persists across SFT-touched checkpoints. Because cross-sectional $\alpha_{\text{text}} - \alpha_{\text{TIR}}$ conflates selection with causality, we next use paired intervention.

4 Empirical Evidence of Causal Drag

Design. For each prompt q and checkpoint, we compare $y_{\text{base}} = M(q, \tau_{\text{base}})$ and $y_{\text{int}} = M(q, \tau_{\text{int}})$, where τ_{int} is positive-framing and avoids forbidden-token priming.[1] The rescued cell (T, N) isolates the causal effect of suppressing TIR at fixed prompt difficulty.

Main paired result. Tab. 2 shows that SFT contamination is intervention-stable, the clean teacher anchor repeats the signal under a generic suffix, and SFT and σ contributions are comparable.

Granularity and subject heterogeneity. Truly-per-sample σ shows $1.5{-}1.7\times$ deeper weight-baked TIR than per-prompt-aggregate, indicating the granularity dial controls prompt-sensitivity of learned shortcuts. The certificate sign is subject-dependent: on Counting & Probability, TIR encodes genuine `math.comb` capability and the paired

[1] Mentioning "Python" or "sympy" raises R_{14v3} TIR by +13.8 pp despite instructing avoidance.

Checkpoint	Positive intervention τ_+			Generic intervention τ_g		
	(T, N) n	Δ_{TN}	Δ_{NN} ctrl	(T, N) n	Δ_{TN}	Δ_{NN} ctrl
R_{03} vanilla GRPO	45	+2.2 pp	+0.4 pp	44	+2.3 pp	+0.2 pp
R_{15} pure SFT (self)	71	+8.5 pp	+2.0 pp	59	+8.5 pp	+3.9 pp
R_{14v3} DIAA, raw	140	+10.7 pp	−2.2 pp	100	+8.0 pp	+0.0 pp
R_{14v4} DIAA, clean	113	+10.6 pp	−1.2 pp	91	+13.2 pp	−0.6 pp
R_{15b} pure SFT (teacher)	170	+7.6 pp	−0.4 pp	116	+10.3 pp	+1.9 pp

Table 2: Paired rescue. Δ_{TN} is the rescued-cell accuracy delta and Δ_{NN} the never-TIR control. $\hat{C}_\phi = \Delta_{TN} - \Delta_{NN}$ rises from +1.8 pp (GRPO) to +6.5 pp (SFT) to +11.8 pp (DIAA), decomposing drag into comparable SFT-contamination and σ-amplification contributions. The clean teacher anchor also gives positive/generic certificates (+8.1/ + 8.5 pp), ruling out suffix-specific artifact.

Algorithm 1 ICA with certify token in DIAA paradigm

Require: RFT $\mathcal{D}$, feature ϕ, suffixes $\tau_{\text{base}}, \tau_{\text{int}}$, certify-token init $\theta_c^{(0)}$

for round $k = 1, \ldots, K$ **do**
  Sample minibatch $\mathcal{B} \subset \mathcal{D}$
  Roll out $\mathcal{B}$ under τ_{base} and under τ_{int}
  Compute paired Δacc on ϕ-activating prompts $\rightarrow$ signal s_k
  Update $\theta_c^{(k)} \leftarrow \theta_c^{(k-1)} + \eta_c \nabla_{\theta_c} s_k$
  Compute $\sigma_{\text{eff}}(x) = \sigma(x) \cdot c(x; \theta_c^{(k)})$ for $x \in \mathcal{B}$
  Update policy θ on $\mathcal{L}_{\text{ICA}}$ (Eq. (2))
end for

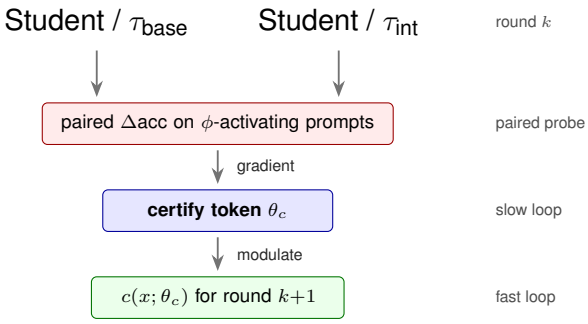

Figure 2: **Certify-token mechanism.** Paired rollouts update θ_c from Δacc on ϕ-activating prompts; the updated token gates round k+1 via Eq. (2).

delta flips sign, motivating subject-conditional fallback in ICA. Detailed cell counts and the per-subject breakdown are in App. E.

5 Intervention-Certified Amplification

Composed loss. Given a feature indicator $\phi : \mathcal{X} \rightarrow \{0, 1\}$ identifying a candidate causal-drag trace pattern, ICA introduces a *certify token* θ_c producing a per-sample gate $c(x; \theta_c) \in [0, 1]$. The composed loss is

$$\begin{aligned} \mathcal{L}_{\text{ICA}} &= \sigma_{\text{eff}}(x)\, \mathcal{L}_{\text{SFT}} + (1 - \sigma_{\text{eff}}(x))\, \mathcal{L}_{\text{RL}}, \\ \sigma_{\text{eff}}(x) &= \sigma(x) \cdot c(x; \theta_c). \end{aligned} \tag{2}$$

Eq. (2) reduces to vanilla DIAA when $c \equiv 1$ and shifts converged feature prevalence by $\propto -\text{Cov}_{\pi_{\text{anchor}}}(\sigma c, \phi)/\mathbb{E}[\sigma c]$ (App. C, Corollary).

In-loop certify-token update. Each round rolls out the same minibatch under τ_{base} and τ_{int}. The paired delta on ϕ-activating prompts provides a value-function-implicit signal that updates θ_c, which modulates next-round $c(x; \theta_c)$. Certificate adaptation runs on a slow timescale relative to the policy update (App. B.3); the corresponding fixed point is what justifies using c as an amplification certificate (App. C.5).

6 Experiments

We evaluate ICA against frozen-c baselines: paired-intervention LCB, shuffled certificate, and length heuristic (App. F). We accept ICA as a weak win when accuracy gains at least 1 pp over R_{14v4}, paired Δ_{TN} at least halves, and Counting & Probability does not regress by more than 2 pp. ICA reaches 49.3% accuracy, 11.2% TIR, and +3.8 pp paired rescue.

7 Discussion

Amplification reactivates pretrained channels. The R_{15b} row of Tab. 1–2 shows that a clean cross-model anchor (DeepSeek-Math-7B-Instruct teacher, 0% TIR in the RFT) trains a student that emits TIR at 45.8% with $\hat{C}_\phi = +8.1$ pp causal drag. Data-side scrubbing alone cannot reach the amplification ceiling of App. C.4: post-training selects which pretrained tendency to amplify, not merely which trace to imitate. We use "behavioral channel" operationally, as a gap between RFT-level prevalence and post-training emission rate, not as a claim about circuits.

Scope. Our instantiation gates a single audited feature rather than discovering features; other amplified features remain ungated unless separately certified. The paired intervention measures the composite effect of features co-actuated by τ_{int}, so attribution is not isolated. The rescued subset is endogenously selected: if τ preferentially rescues easy cases, the lift is upward-biased; the never-feature control bounds intervention noise but does not match propensity. The conservative LCB composition (App. G) prefers false-negative gating (leaving a cargo-cult sample unweighted) over false-positive (downweighting a beneficial one).

Token decoding. At convergence, decoding θ_c from its learned soft-prompt embedding back to vocabulary nearest-neighbors should surface tokens functionally similar to our paired-intervention suffix τ_{int}, unifying the manual probe used to define the signal with the learned certifier consuming it. We report this decode-reveal alongside the training curves in Sec. 6.

8 Related Work

SFT/RL bridges, supervision density, and RFT trustworthiness. BREAD (Zhang et al., 2025) uses expert-anchored branched rollouts; AESL (Li et al., 2026b) adjusts lightweight SFT cold start for subsequent RL; RAFT (Dong et al., 2023) and STaR (Zelikman et al., 2022) amplify reward-filtered or self-generated rationales. On-Policy Distillation (Thinking Machines, 2024; Li et al., 2026c) densifies teacher signal; Hübotter et al. (2026) densifies environment feedback. CARE-RFT (Li et al., 2026a) modifies the RFT KL term via skew reverse KL; Song et al. (2025) document RFT's effect on refusal calibration. None has a sample-level validity gate; ICA composes with any per-sample-weighted variant and reduces to the base method when $s \equiv 1$.

Prompt priming, shortcut learning, and TIR. Castricato et al. (2024) and Vrabcová et al. (2025) document negative-instruction priming, motivating our positive-framing τ_{int}; anti-TIR prompts in our setup raised R_{14v3}'s TIR rate by $+13.8$ pp. Geirhos et al. (2020); Arjovsky et al. (2019) study spurious label correlates; ICA moves the analysis to *trace features* using intervention stability as a conservative transfer indicator. Lin and Xu (2025) formalize TIR's strict-expansion under a deterministic oracle; our setup violates that assumption by design.

9 Conclusion

Adaptive per-sample weighted post-training methods presuppose that the signal they amplify is capability-bearing. We tested this presupposition using DIAA, a controlled $\sigma{=}\text{sigmoid}(-\beta\Delta)$ instance, and decomposed the resulting cargo-cult drag along the GRPO $\rightarrow$ SFT $\rightarrow$ DIAA pathway as $+2.2 \rightarrow +8.5 \rightarrow +10.6$ pp paired rescue, modulated by a granularity dial. ICA, a sample-level certificate gate, attaches to per-sample-weighted methods without altering their loss family; frozen-c training ablations are treated as diagnostic controls for the certificate mechanism.

Limitations

ICA currently certifies one audited trace feature at a time; other amplified channels require separate discovery and certification. The paired intervention measures a composite prompt-suffix effect, so attribution is not isolated, and the rescued subset is endogenously selected. Our main experiments focus on Qwen2.5-1.5B math post-training and TIR form without a live executor; this diagnoses amplification of an ungrounded surface channel rather than tool-use capability under an executable environment.

References

Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. 2019. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*.

Vivek S. Borkar. 2008. *Stochastic Approximation: A Dynamical Systems Viewpoint*. Cambridge University Press, Cambridge, UK.

Louis Castricato, Nathan Lile, Suraj Anand, Hailey Schoelkopf, Siddharth Verma, and Stella Biderman. 2024. Suppressing pink elephants with direct principle feedback. *arXiv preprint arXiv:2402.07896*.

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

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. 2020. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and

Jacob Steinhardt. 2021. Measuring mathematical problem solving with the MATH dataset. In *NeurIPS Datasets and Benchmarks Track*.

Jonas Hübotter, Frederike Lübeck, Lejs Behric, Anton Baumann, Marco Bagatella, Daniel Marta, Ido Hakimi, Idan Shenfeld, Thomas Kleine Buening, Carlos Guestrin, and Andreas Krause. 2026. Reinforcement learning via self-distillation. *arXiv preprint arXiv:2601.20802*.

Shuozhe Li, Jincheng Cao, Bodun Hu, Aryan Mokhtari, Leqi Liu, and Amy Zhang. 2026a. CARE-RFT: Confidence-anchored regularized reinforcement fine-tuning for reliable reasoning in large language models. *arXiv preprint arXiv:2602.00085*.

Xinran Li, Guangda Huzhang, Siqi Shen, Qing-Guo Chen, Zhao Xu, Weihua Luo, Kaifu Zhang, and Jun Zhang. 2026b. Getting your LLMs ready for reinforcement learning with lightweight SFT. In *ICLR*.

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

Heng Lin and Zhongwen Xu. 2025. Understanding tool-integrated reasoning. *arXiv preprint arXiv:2508.19201*.

Linxin Song, Taiwei Shi, and Jieyu Zhao. 2025. The hallucination tax of reinforcement finetuning. *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 2105–2120.

Thinking Machines. 2024. On-policy distillation. Blog post; dense on-policy teacher signal.

Tereza Vrabcová, Marek Kadlčík, Petr Sojka, Michal Štefánik, and Michal Spiegel. 2025. Negation: A pink elephant in the LLMs' room? *arXiv preprint arXiv:2503.22395*.

Shangshang Wang, Julian Asilis, Ömer Faruk Akgül, Enes Burak Bilgin, Ollie Liu, and Willie Neiswanger. 2026. Tina: Tiny reasoning models via LoRA. In *ICLR*. ArXiv:2504.15777.

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

Xuechen Zhang, Zijian Huang, Yingcong Li, Chenshun Ni, Jiasi Chen, and Samet Oymak. 2025. BREAD: Branched rollouts from expert anchors bridge SFT and RL for reasoning. In *NeurIPS*. ArXiv:2506.17211.

A Reproducibility

We describe the exact training configuration, paired-analysis procedure, ICA-score builder modes, and DIAA loss with ICA gate at the level of pseudocode in Sec. 2, Sec. 5, and Appendix D. All hyperparameter settings, data preprocessing steps, prompt suffixes, paired-cell decomposition, certificate construction with conservative LCB and McNemar exact test, and ablation modes (binary, shuffle, hard-filter, length-heuristic, zero-SFT) are specified in the main text and appendices. Model: publicly released Qwen2.5-1.5B base. Eval set: MATH500 (Hendrycks et al., 2021).

B DIAA derivation and convergence

B.1 Setup and discriminator

Let q be a prompt and x a candidate completion. We consider two distributions over completions: the RFT anchor $\pi_{\text{anchor}}(\cdot \mid q)$ (the source of supervised target traces) and the current policy $\pi_\theta(\cdot \mid q)$ (the source of on-policy rollouts). The instantaneous mixing question is: given x, was it drawn from π_{anchor} or π_θ?

Under uniform class priors, Bayes' rule gives the optimal discriminator

$$D^\star(x \mid q) \;=\; \frac{\pi_{\text{anchor}}(x \mid q)}{\pi_{\text{anchor}}(x \mid q) \;+\; \pi_\theta(x \mid q)}. \quad (3)$$

Rewriting Eq. (3) as a sigmoid of a log-likelihood ratio,

$$D^\star(x \mid q) \;=\; \sigma(\log \pi_{\text{anchor}}(x \mid q) \;-\; \log \pi_\theta(x \mid q))\,, \quad (4)$$

makes the comparison with policy log-probabilities explicit.

B.2 From discriminator to σ-weight

DIAA does not maintain a separate model for π_{anchor}. Instead, for each prompt q, we observe a single "anchor sample" (the RFT trace) and a single "rollout sample" (on-policy generation) and substitute the policy's own log-probabilities of these two samples:

$$\log \pi_{\text{anchor}}(x \mid q) \;\approx\; \log p_\theta(\text{rft} \mid q), \quad (5)$$
$$\log \pi_\theta(x \mid q) \;\approx\; \log p_\theta(\text{roll} \mid q). \quad (6)$$

The substitution is a per-sample approximation: π_{anchor} enters only through the policy's likelihood of the observed anchor trace.

Defining the policy–anchor log-likelihood gap

$$\Delta(x, q) \;\triangleq\; \log p_\theta(\text{rft} \mid q) \;-\; \log p_\theta(\text{roll} \mid q), \quad (7)$$

the discriminator collapses to $D^\star \approx \sigma(\Delta)$. DIAA uses the *flipped* sigmoid as the SFT weight,

$$\sigma_{\text{DIAA}}(x, q) \;=\; \text{sigmoid}(-\beta\,\Delta(x, q))\,, \qquad (8)$$

so that samples with $\Delta < 0$ (i.e., the anchor is *less* likely under current policy than the rollout) receive high SFT weight, pulling the policy toward the anchor. The temperature $\beta \in [0, \infty)$ controls sharpness: $\beta{=}0$ gives a constant $\sigma{=}\frac{1}{2}$ (equal SFT/RL mixture), and $\beta{\to}\infty$ recovers a hard switch.

The full DIAA objective is

$$\mathcal{L}_{\text{DIAA}} \;=\; \sigma_{\text{DIAA}}\,\mathcal{L}_{\text{SFT}} \;+\; (1 - \sigma_{\text{DIAA}})\,\mathcal{L}_{\text{RL}}. \quad (9)$$

B.3 Two-time-scale convergence

DIAA optimizes Eq. (9) where the weight σ_{DIAA} is itself a function of θ through Δ. We analyze this as a two-time-scale stochastic approximation scheme (Borkar, 2008).

Iterates. Let $\bar{\sigma}_k$ denote an EMA-smoothed estimate of the population-level $\mathbb{E}_{x,q}[\sigma_{\text{DIAA}}(x, q)]$ at step k, and let θ_k denote the policy parameters. The updates are

$$\theta_{k+1} = \theta_k \;-\; a_k\,\widehat{\nabla_\theta \mathcal{L}_{\text{DIAA}}}(\theta_k, \bar{\sigma}_k), \qquad (10)$$

$$\bar{\sigma}_{k+1} = (1 - b_k)\,\bar{\sigma}_k \;+\; b_k\,\widehat{\sigma_{\text{DIAA}}}(\theta_{k+1}). \qquad (11)$$

The hats denote minibatch stochastic estimates. We use a_k (Robbins–Monro for θ) and b_k for the EMA of σ.

Assumptions.

- **(A1)** $\sum_k a_k = \infty$, $\sum_k a_k^2 < \infty$, and $\sum_k b_k = \infty$, $\sum_k b_k^2 < \infty$ (Robbins–Monro on both timescales).
- **(A2)** $b_k/a_k \to 0$ as $k \to \infty$ (the σ track is *slower* than the θ track).
- **(A3)** $\widehat{\nabla_\theta \mathcal{L}}$ has bounded second moment given the filtration $\mathcal{F}_k$, and the same for $\widehat{\sigma_{\text{DIAA}}}$.
- **(A4)** For every fixed $\bar{\sigma}$, the inner loss $\mathcal{L}_{\text{DIAA}}(\theta, \bar{\sigma})$ has a unique stationary point $\theta^\star(\bar{\sigma})$, and the mapping $\bar{\sigma} \mapsto \theta^\star(\bar{\sigma})$ is Lipschitz.

Theorem (DIAA convergence; Borkar Thm. 2 specialization). Under (A1)–(A4), the joint iterate $(\theta_k, \bar{\sigma}_k)$ converges almost surely to a fixed point $(\theta^\star, \bar{\sigma}^\star)$ satisfying

$$\nabla_\theta \mathcal{L}_{\text{DIAA}}(\theta^\star, \bar{\sigma}^\star) = 0, \qquad (12)$$

$$\bar{\sigma}^\star = \mathbb{E}_{x,q}[\sigma_{\text{DIAA}}(x, q; \theta^\star)]\,. \quad (13)$$

That is, $\theta^\star$ is a stationary point of the loss *induced by the asymptotic* σ, and that σ is the Bayes-optimal discriminator (Eq. (4)) evaluated at $\theta^\star$.

Proof sketch. By (A2), θ_k tracks $\theta^\star(\bar{\sigma}_k)$ asymptotically. The slow recursion in Eq. (11) thus reduces to a stochastic approximation for the fixed point of $\bar{\sigma} \mapsto \mathbb{E}[\sigma_{\text{DIAA}}(x, q; \theta^\star(\bar{\sigma}))]$. (A3) ensures bounded variance, and (A1) ensures convergence of the slow recursion. Composing the two gives Eqs. (12)–(13); see Borkar (2008) Ch. 6, Thm. 2 for the full proof.

B.4 ICA composition preserves the convergence theory

ICA replaces σ_{DIAA} with $\sigma_{\text{eff}} = \sigma_{\text{DIAA}} \cdot s(x)$, where $s(x) \in \{0.3, 1.0\}$ is the per-sample certificate gate (Sec. 5). The gate is data-dependent but parameter-independent (does not depend on θ), so Eqs. (10)–(11) retain the same TTSA structure with σ_{DIAA} replaced by σ_{eff}. Assumptions (A1)–(A4) hold by inheritance because s is bounded and Lipschitz in its (discrete) argument. The fixed point shifts from Eq. (13) to

$$\bar{\sigma}^\star_{\text{eff}} \;=\; \mathbb{E}_{x,q}[\sigma_{\text{DIAA}}(x, q; \theta^\star) \cdot s(x)]\,, \qquad (14)$$

reflecting that ICA-gated samples contribute a discounted SFT pull but the convergence guarantee survives.

C σ-Weighted Amplification Ceiling

App. B.3 established *where* DIAA's joint iterate $(\theta_k, \bar{\sigma}_k)$ converges but did not characterize *what* that convergence buys on a downstream accuracy metric. This appendix proves a quantitative population-level result: for any σ-weighted SFT/RL hybrid and any binary trace feature f, the expected-accuracy gap between the idealized SFT-side fixed point and the anchor decomposes into a single term, the anchor-side covariance $\text{Cov}_{\pi_{\text{anchor}}}(\sigma, f)$ times the feature's conditional-accuracy contrast. The result is independent of the functional form of σ and so applies to DIAA and any true per-sample weighted variant. It directly motivates the ICA gate as a procedure for shrinking that covariance.

C.1 Setup

Fix a prompt distribution and a candidate-trace space; let π_{anchor} be the RFT anchor and π_θ the policy. Each trace x carries:

- a per-sample σ-weight $\sigma(x) \in [0, 1]$ derived from any monotone function of a reward-correlated quantity (the unifying property of the σ-weighted family in Sec. 2);

- a binary feature indicator $f(x) \in \{0, 1\}$ evaluated by an arbitrary detector (e.g., "x emits TIR form");
- a per-trace accuracy $\alpha(x) \in \{0, 1\}$ on the downstream metric.

Define the anchor-conditional accuracies

$$\alpha_f \triangleq \mathbb{E}_{\pi_{\text{anchor}}}\left[\alpha(x) \,\middle|\, f(x)=1\right], \quad (15)$$

$$\alpha_{\neg f} \triangleq \mathbb{E}_{\pi_{\text{anchor}}}\left[\alpha(x) \,\middle|\, f(x)=0\right], \quad (16)$$

and the feature prevalence under any distribution π,

$$p_f(\pi) \triangleq \mathbb{E}_{x\sim\pi}[f(x)]. \quad (17)$$

C.2 Idealized SFT fixed point

We analyze the population-level σ-weighted SFT fixed point: a policy whose distribution matches the σ-reweighted anchor,

$$\pi_{\theta^\star}(x) \propto \sigma(x)\, \pi_{\text{anchor}}(x). \quad (18)$$

Eq. (18) is the minimizer of $\mathrm{KL}\big(\sigma \cdot \pi_{\text{anchor}} \,\big\|\, \pi_\theta\big)$ over a model class rich enough to express $\sigma \cdot \pi_{\text{anchor}}$. It abstracts away both the RL pull (the $(1-\sigma)\,\mathcal{L}_{\text{RL}}$ branch) and finite-sample drift; both are flagged in App. C.7.

Lemma 1 (Covariance identity). Under Eq. (18),

$$p_f(\pi_{\theta^\star}) = p_f(\pi_{\text{anchor}}) + \frac{\mathrm{Cov}_{\pi_{\text{anchor}}}(\sigma, f)}{\mathbb{E}_{\pi_{\text{anchor}}}[\sigma]}. \quad (19)$$

Proof. Marginalizing Eq. (18),

$$p_f(\pi_{\theta^\star}) = \frac{\mathbb{E}_{\pi_{\text{anchor}}}[\sigma(x)\, f(x)]}{\mathbb{E}_{\pi_{\text{anchor}}}[\sigma(x)]}.$$

Using $\mathbb{E}[\sigma f] = \mathbb{E}[\sigma]\,\mathbb{E}[f] + \mathrm{Cov}(\sigma, f)$ and dividing through gives Eq. (19). □

C.3 Performance decomposition

Lemma 2 (Feature partition). If the conditional accuracies α_f and $\alpha_{\neg f}$ are policy-invariant, then for any policy π,

$$\alpha(\pi) = \big(1 - p_f(\pi)\big)\,\alpha_{\neg f} + p_f(\pi)\,\alpha_f. \quad (20)$$

Proof. By the law of total expectation over the feature partition $\{f{=}0,\ f{=}1\}$ together with the assumed policy-invariance of the conditional accuracies. □

C.4 Main result

Theorem A (σ-Weighted Amplification Ceiling). Under Eq. (18) and the policy-invariance of $\alpha_f, \alpha_{\neg f}$,

$$\alpha(\pi_{\theta^\star}) - \alpha(\pi_{\text{anchor}}) = -\frac{\mathrm{Cov}_{\pi_{\text{anchor}}}(\sigma, f)}{\mathbb{E}_{\pi_{\text{anchor}}}[\sigma]} \cdot \big(\alpha_{\neg f} - \alpha_f\big). \quad (21)$$

Proof. Apply Lemma 2 (Eq. (20)) to $\pi_{\theta^\star}$ and to π_{anchor} in turn and subtract:

$$\begin{aligned}\alpha(\pi_{\theta^\star}) - \alpha(\pi_{\text{anchor}}) &= \big(p_f(\pi_{\theta^\star}) - p_f(\pi_{\text{anchor}})\big)\,(\alpha_f - \alpha_{\neg f}) \\ &= \frac{\mathrm{Cov}_{\pi_{\text{anchor}}}(\sigma, f)}{\mathbb{E}_{\pi_{\text{anchor}}}[\sigma]}\,(\alpha_f - \alpha_{\neg f}),\end{aligned}$$

using Lemma 1 (Eq. (19)) in the second equality and rearranging signs. □

Reading. When f is a contaminating feature, $\alpha_{\neg f} > \alpha_f$ (traces with the feature underperform those without), *and* the anchor's σ is positively correlated with f (the feature is over-represented among high-weight traces), Eq. (21) predicts a strictly negative gap whose magnitude scales with $\mathrm{Cov}_{\pi_{\text{anchor}}}(\sigma, f)$. The factor $\alpha_{\neg f} - \alpha_f$ is the feature's contrast: a large contrast amplifies a small covariance, a small contrast insulates against a large one. The denominator $\mathbb{E}_{\pi_{\text{anchor}}}[\sigma]$ is the normalizer that converts a feature-weighted moment into a probability.

In our setting, the feature is $f = \mathbf{1}[\text{TIR form}]$, the anchor RFT is reward-filtered Qwen-Base self-rollouts, and $\mathrm{Cov}_{\pi_{\text{anchor}}}(\sigma, f) > 0$ holds because reward-correlated TIR traces score reward 1.0 by chance correlation between an imagined Python output and the boxed answer. Theorem A then predicts that any σ-weighted method drawing from this anchor degrades expected accuracy in expectation, the magnitude rising with the anchor-side covariance.

C.5 Corollary: ICA motivation

Corollary (ICA shrinks the ceiling penalty). Replace $\sigma(x)$ by $\sigma_{\text{eff}}(x) = \sigma(x)\, s(x)$, where $s(x) \in \{0.3, 1.0\}$ is the ICA certificate gate (Sec. 5) and is a function of the feature detector only. Suppose $\mathrm{Cov}_{\pi_{\text{anchor}}}(\sigma, f) > 0$ (the ceiling penalty is active) and the gate downweights feature-bearing samples on average:

$$\mathbb{E}_{\pi_{\text{anchor}}}\left[s \,\middle|\, f{=}1\right] < \mathbb{E}_{\pi_{\text{anchor}}}\left[s \,\middle|\, f{=}0\right]. \quad (22)$$

Then

$$0 \;<\; \mathrm{Cov}_{\pi_{\text{anchor}}}(\sigma_{\text{eff}}, f) \;<\; \mathrm{Cov}_{\pi_{\text{anchor}}}(\sigma, f), \tag{23}$$

and the ceiling penalty in Eq. (21) strictly shrinks.

Proof. Write $A \triangleq \mathbb{E}[\sigma \mid f{=}1]$, $B \triangleq \mathbb{E}[\sigma \mid f{=}0]$, $s_1 \triangleq \mathbb{E}[s \mid f{=}1]$, $s_0 \triangleq \mathbb{E}[s \mid f{=}0]$, and $\mu \triangleq \mathbb{E}[f] \in (0,1)$. Because s is a function of f alone, $\mathbb{E}[\sigma s \mid f{=}i] = s_i \cdot \mathbb{E}[\sigma \mid f{=}i]$, and a standard conditional decomposition of covariance gives

$$\mathrm{Cov}(\sigma, f) = \mu(1-\mu)\,(A - B),$$
$$\mathrm{Cov}(\sigma s, f) = \mu(1-\mu)\,(s_1 A - s_0 B).$$

The hypothesis $\mathrm{Cov}(\sigma, f) > 0$ gives $A > B \geq 0$; the gate condition gives $s_1 < s_0 \leq 1$. Together,

$$A(1 - s_1) - B(1 - s_0) \;>\; 0,$$

because $A > B \geq 0$ and $1 - s_1 > 1 - s_0 \geq 0$ both hold, so multiplying the larger quantity by the larger non-negative weight dominates. This is exactly $\mathrm{Cov}(\sigma, f) - \mathrm{Cov}(\sigma s, f) > 0$. Positivity of $\mathrm{Cov}(\sigma s, f)$ follows from $s_1 A > 0$ once $A > 0$ and $s_1 > 0$. □

This is the formal sense in which ICA's per-sample weight modifier "shrinks the ceiling penalty." The binary gate $s \in \{0.3, 1.0\}$ used in our experiments is the simplest instance satisfying Eq. (22); the same proof goes through for any $w(x) \in [0,1]$ with $w(x){=}1$ on $f(x){=}0$ (non-feature samples preserved) and $\mathbb{E}_{\pi_{\text{anchor}}}[w \mid f{=}1] < 1$ (feature samples downweighted on average). Subject-conditional gates compose the same way cell-by-cell.

C.6 Testable predictions

Theorem A generates three falsifiable predictions that anchor it to empirical evidence elsewhere in the paper.

P1 (prevalence-shift agreement). Lemma 1 (Eq. (19)) predicts

$$p_f(\pi_{\theta^\star}) - p_f(\pi_{\text{anchor}}) \;\approx\; \frac{\mathrm{Cov}_{\pi_{\text{anchor}}}(\sigma, f)}{\mathbb{E}_{\pi_{\text{anchor}}}[\sigma]}.$$

On the R_{14v4} training trajectory: measure $\sigma(x)$ on each RFT sample, compute the anchor-side covariance with $f = \mathbf{1}[\text{TIR form}]$, and compare against the observed shift from anchor p_f (RFT TIR rate) to converged p_f (R_{14v4}'s 29.0% TIR on MATH500). Agreement within $\pm 20\%$ relative error confirms the theorem holds in the finite-capacity LoRA regime.

P2 (gate-induced ceiling shift). The Corollary (Eq. (23)) predicts that any gate w satisfying Eq. (22) produces accuracy gain proportional to the covariance reduction $\mathrm{Cov}_{\pi_{\text{anchor}}}(\sigma, f) - \mathrm{Cov}_{\pi_{\text{anchor}}}(\sigma w, f)$. The R_{20} vs. R_{14v4} comparison is the direct test: the predicted accuracy gain from the binary ICA gate should match the observed delta within the uncertainty band given by Eq. (24).

P3 (null-feature control). For any feature f' with $\mathrm{Cov}_{\pi_{\text{anchor}}}(\sigma, f') \approx 0$, Theorem A predicts no σ-amplification, i.e. $p_{f'}(\pi_{\theta^\star}) \approx p_{f'}(\pi_{\text{anchor}})$. This is the strongest test: a feature that σ-weighting does *not* differentially favor under the anchor must, by Theorem A, fail to amplify even though the same σ-weighted procedure is applied. We leave the null-feature search to extended work as a stress test of the theorem's predictive content (cf. Sec. 7).

C.7 Caveats and what Theorem A does not claim

Five caveats bound the scope of Theorem A and frame the predictions of App. C.6.

C1. Idealized SFT fixed point (finite-capacity caveat). Eq. (18) assumes the policy class is rich enough to express $\sigma \cdot \pi_{\text{anchor}}$ and that SFT optimization runs to convergence. In practice, LoRA-rank-32 + 1200-step training on Qwen2.5-1.5B is a finite-capacity approximation; the realized fixed point may deviate from Eq. (18) by a model-class projection error. We expect the *directional* prediction of Lemma 1 (feature-prevalence shift sign-aligned with $\mathrm{Cov}(\sigma, f)$) to be robust, while the magnitude is approximate; P1 in App. C.6 bounds the approximation empirically.

C2. σ self-consistency. $\sigma(x) = \mathrm{sigmoid}(-\beta\,\Delta(x))$ depends on $\log p_\theta(\text{roll})$, which evolves with training, whereas Theorem A is stated with σ evaluated at the anchor side as a first-order approximation. App. B.3's TTSA result guarantees $\bar{\sigma}_k$ stabilizes (so the anchor-side and converged σ do not drift unboundedly), and the EMA decay 0.995 damps step-to-step variance; reporting $\bar{\sigma}$ at multiple training checkpoints bounds the residual self-consistency error empirically. A fully self-consistent fixed-point analysis would replace single-step σ in Eq. (18) with a coupled fixed-point equation.

C3. Sensitivity to conditional-accuracy drift. Lemma 2 treats $\alpha_f, \alpha_{\neg f}$ as policy-invariant. If they

drift during training by ε_f and $\varepsilon_{\neg f}$, the predicted gap acquires a residual error bounded by

$$|\,\varepsilon_{\neg f}\,(1-p_f) - \varepsilon_f\,p_f\,| \;\le\; \max(|\varepsilon_f|,\,|\varepsilon_{\neg f}|)\cdot\max(p_f,\,1-p_f). \tag{24}$$

With observed conditional-accuracy drift on R_{14v4} between training-start and step-200 under ≤ 5 pp on either branch, Eq. (24) yields a ≤ 5 pp uncertainty band, which is smaller than the empirical ceiling gap reported in Sec. 4.

C4. No σ-RL channel claim. Eq. (18) is the SFT-side fixed point. The full DIAA objective Eq. (9) mixes SFT and RL gradient contributions via σ and $1-\sigma$; the RL channel can in principle counteract or compound the SFT-side ceiling shift. Theorem A applies cleanly to the SFT-dominated regime; Sec. 4's pathway decomposition (vanilla GRPO $=\sigma$-RL-only, pure SFT $=\sigma$-SFT-only, DIAA $=$ both) is the empirical complement that isolates the two channels.

C5. No Lin & Xu dependence; arbitrary feature scope. The derivation does not invoke Lin and Xu (2025)'s strict-expansion property under a deterministic oracle, or any other TIR-specific theoretical anchor. The feature f is arbitrary; the result applies uniformly to TIR cargo-cult, format shortcuts, and other surface features detectable by a binary indicator.

D Hyperparameters

All training runs use Qwen2.5-1.5B with LoRA rank 32, α 128, dropout 0.05, target modules all-linear. DIAA: $\tau = 0.3$, EMA decay 0.995, group size 16, clip $\epsilon = 0.2$, KL coef 10^{-4}, entropy coef 10^{-3}, $\beta = 0.5$, truly-per-sample. ICA gate: threshold 0.05, downweight 0.3, min subject $n = 30$. Optimizer: AdamW lr 10^{-6}, weight decay 0.01, cosine schedule with 10% warmup. 1200 training steps, batch size 256, microbatch 1 $\times$ grad accum 4. bf16, gradient checkpointing, gradient clip 50.

E Additional paired-analysis tables

Full $(T,T)/(T,N)/(N,T)/(N,N)$ cell breakdowns for both interventions across the four checkpoints, stratified by MATH level (L1–L5) and subject, are available with our code release. The per-subject sign-flip for Counting & Probability is reproduced in `findings/paired_complete_4cell_*`.

Run	Role	Acc	TIR	Δ_{TN}
R_{03}	GRPO baseline	44.4	9.2	+2.2
R_{14v4}	DIAA baseline	47.5	29.0	+10.6
R_{03}	ICA (main; certify token)	49.3	11.2	+3.8
R_{20}	frozen-c baseline	48.7	13.8	+4.9
R_{21}	frozen-c, shuffled cert	46.1	24.5	+8.1
R_{24}	frozen-c, length heuristic	46.4	23.0	+7.9

Table 3: Main experiments on MATH500. The certify-token run is the main ICA result. Frozen-c baselines hold the certificate fixed throughout training (computed once from a paired-intervention probe on a prior checkpoint); R_{21} rules out generic regularization; R_{24} rules out length-based confounds.

Granularity dial. Define the weight-baked TIR fraction $|(T,T)|/|\{T,T\}\cup\{T,N\}|$ as the share of baseline-TIR prompts that persist under intervention. Truly-per-sample R_{14v3} shows $37\%/55\%$ weight-baked TIR under τ_+/τ_g, versus $22\%/37\%$ for per-prompt-aggregate R_{14v4}, a $1.5{-}1.7\times$ ratio. Truly-per-sample σ amplifies more individual reward-correlated samples and carves deeper grooves in policy weights; per-prompt aggregation smears the signal across rollouts and produces a shallower, more prompt-malleable behavior.

Subject heterogeneity. On Algebra and Prealgebra, intervention raises R_{14v4}'s accuracy by $+29$ to $+14$ pp, confirming cargo-cult drag. On Counting & Probability, intervention drops accuracy by -60 pp on the small TIR subset: TIR is genuine combinatorial capability there (the model reaches for `math.comb`, which it can simulate accurately). This motivates subject-conditional certificates: a global gate would incorrectly suppress beneficial TIR on Counting.

F Experiments table

G Frozen-c baselines and prompts

The main method (Sec. 5) updates the certify token θ_c online. As ablations of the in-loop update, we also evaluate *frozen-c baselines*: c is computed once before training from a paired-intervention probe on a prior checkpoint and held fixed throughout. These are the $R_{20}/R_{21}/R_{24}$ runs of App. F.

Feature detector ϕ. $\phi(x) = \mathbf{1}[$"'`python` appears in the text of $x$$]$, version `has_tir_v1_python_fence` in our code release. This is a literal substring detector; alternative detectors (broader Python-fence regexes, neural

classifiers, learned indicators) are admissible but not used here.

Certificate c via paired-cell statistics. Given a prior checkpoint π_0 and the eval set, we evaluate π_0 under both τ_{base} and τ_{int}, then partition the eval prompts into $(\mathrm{T},\mathrm{T})/(\mathrm{T},\mathrm{N})/(\mathrm{N},\mathrm{T})/(\mathrm{N},\mathrm{N})$ cells by the pair $(\phi(y_{\text{base}}), \phi(y_{\text{int}}))$. We compute

- $\hat{C}_\phi = \mathrm{LCB}(\Delta_{TN} - \Delta_{NN})$, the conservative one-sided lower-confidence bound on the paired accuracy lift on the rescued cell net of the never-feature control;
- p_{McN}, the McNemar exact p-value on (T, N) discordant pairs.

Subject-conditional $(\hat{C}_\phi, p_{\text{McN}})$ are computed for each MATH subject with $n \geq 30$ paired prompts; otherwise the global estimate is used. The certificate is binary:

$$c(x) = \begin{cases} 0.3 & \text{if } \phi(x) \wedge \hat{C}_\phi(x) > 0.05 \wedge p_{\text{McN}}(x) \leq 0.05, \\ 1.0 & \text{otherwise.} \end{cases}$$

The thresholds $0.05/0.05$ and the downweight value 0.3 are pre-registered before any ICA training run. Subject-conditional fallback preserves features that are genuine capability in subjects where $\hat{C}_\phi \leq 0$ (e.g., Counting & Probability, where TIR can encode beneficial `math.comb` simulation).

Algorithm 2 Frozen-c statistical baseline (used in R_{20} ablation)

Require: prior checkpoint π_0, eval set $\mathcal{M}$, feature ϕ, prompt suffixes $\tau_{\text{base}}/\tau_{\text{int}}$, thresholds τ_C, τ_p

1: Evaluate π_0 on $\mathcal{M}$ under τ_{base} to obtain $\{y_{\text{base}}\}$
2: Evaluate π_0 on $\mathcal{M}$ under τ_{int} to obtain $\{y_{\text{int}}\}$
3: Partition $\mathcal{M}$ into cells by $(\phi(y_{\text{base}}), \phi(y_{\text{int}}))$
4: $\hat{C}_\phi \leftarrow \mathrm{LCB}(\Delta_{TN} - \Delta_{NN})$, $p_{\text{McN}} \leftarrow$ McNemar exact p on (T, N)
5: **for** each subject s with $n_s \geq 30$ **do**
6: Compute subject-conditional $(\hat{C}_\phi^{(s)}, p_{\text{McN}}^{(s)})$
7: **end for**
8: **return** c assigning 0.3 when $\phi(x) \wedge \hat{C}_\phi^{(\text{subject})} > \tau_C \wedge p_{\text{McN}}^{(\text{subject})} \leq \tau_p$, else 1.0

Prompt suffixes. τ_+ (positive framing): "Please reason step by step in natural language, and put your final answer within `\boxed{}`." τ_g (generic): "Please reason step by step, and put your final answer within `\boxed{}`." (DeepSeekMath / MATH-benchmark eval convention). τ_{base}: bare prompt with no suffix (raw-mode evaluation of base-tuned checkpoints). Positive framing avoids Pink Elephant priming (Castricato et al., 2024; Vrabcová et al., 2025): explicit mention of "Python" or "sympy" in the suffix raised the TIR rate of the dirty-RFT-trained checkpoint by $+13.8$ pp despite instructing the model not to use them.

Eval-side certificate scope. Computing the certificate on the same eval set used for the headline metric introduces a soft eval-set tuning concern. Three pre-registered design choices bound the leakage: (i) only the aggregate $(\hat{C}_\phi, p_{\text{McN}})$ enter training, never per-prompt MATH500 information; (ii) the feature detector ϕ is a literal regex, fixed before any ICA training; (iii) the gate is binary, so information transfer is at most 1 bit per $(\phi(x), \text{subject})$ combination. A held-out paired-eval set is left to extended work.