

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

Leiyue Chen*, Meng Wei*, Xuefeng Hu, Wei Liu†

ByteDance

{leiyue.ml, xmechira, liujikun63}@gmail.com, {xuefeng.hu}@bytedance.com

Abstract

Post-training methods use training-side signals to decide which supervision to emphasize, but those signals need not identify behavior that is capability-bearing in the target environment. We probe this with **DIAA** (Dual-Implicit Adversarial Amplification), a per-sample $\sigma = \text{sigmoid}(-\beta\Delta)$ SFT/RL hybrid: after every literal Python code fence is deleted from the anchor data, the model still emits unexecuted Tool-Integrated Reasoning (TIR) form on 29% of MATH-500 prompts, three times vanilla GRPO, and suppressing the form recovers accuracy. A paired within-prompt intervention decomposes the removable drag along $\text{GRPO} \rightarrow \text{SFT} \rightarrow \text{DIAA}$ as $+2.2 \rightarrow +8.5 \rightarrow +10.6$ pp, identifying SFT imitation and σ -amplification as comparable sources; a teacher RFT with 0% TIR likewise trains a student emitting TIR at 45.8%. We propose **Intervention-Certified Amplification (ICA)**, which gates the per-sample weight by a certificate derived from paired rollouts, $\sigma_{\text{eff}}(x) = \sigma(x)c(x; \theta_c)$. Frozen and online instantiations cut TIR to 13.8% and 11.2%, more than halve residual drag while improving accuracy, and carry the joint improvement to held-out GSM8K without retuning the certificate.

1 Introduction

When should a model imitate a demonstration, and when should it explore on its own? Recent post-training methods answer it with training-side signals. OPD (Lu and Thinking Machines Lab, 2025; Li et al., 2026c) densifies teacher supervision, Hübottter et al. (2026) densify environment feedback through self-distillation, BREAD (Zhang et al., 2025) splices expert-anchored prefixes, AESL (Li et al., 2026b) tunes a lightweight SFT cold start, and RAFT (Dong et al., 2023) and

Certify a behavior before amplifying it

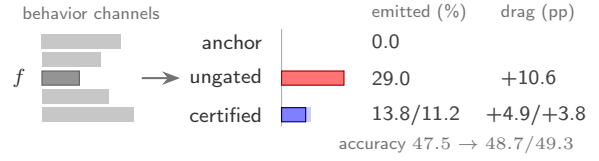


Figure 1: **Certify a behavior before amplifying it.** σ -weighted post-training selects one channel out of the pretrained behavior repertoire and amplifies it; an intervention-derived certificate gates that channel. The channel column on the left is a conceptual depiction and carries no measurement; only the highlighted channel f is instantiated experimentally, as unexecuted Python-fence form (TIR). *Paired drag* is the accuracy recovered on matched prompts when an intervention removes f ; after gating it is residual, so lower is better. Certified values are the frozen and online instantiations of the same gate rather than a ranking; Tab. 3 gives the full comparison, including the ungated baseline and non-certificate controls.

STaR (Zelikman et al., 2022) keep reward-filtered or self-generated rationales. However the signal is formed, it is computed on the training data, while the usefulness it stands in for is realized in the target environment. Whether the two agree is not itself tested.

That presupposition is testable, and cleaning the data does not secure it. We deleted every literal Python code fence from the anchor data used for reinforcement learning on Qwen2.5-1.5B, and the post-trained model wrote Python anyway, on 29% of MATH-500 prompts, three times the rate of vanilla GRPO. Where suppressing the form succeeded, it recovered +10.6 pp of accuracy. The behavior was not explained by literal RFT imitation: post-training reactivated the form from the pretrained model. Validity must therefore be certified in the target environment rather than inferred from literal training-data prevalence.

Our probe is **DIAA** (Sec. 2), a per-sample σ -

* Equal contribution.

† Corresponding author.

weighted SFT/RL hybrid derived from the Bayes-optimal anchor-policy discriminator and used to isolate amplification rather than to claim accuracy. 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 subsequent rounds use 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 $c \equiv 1$.

Roadmap. Sec. 2–3 introduce DIAA as the amplification probe and report the cross-sectional phenomenon; Sec. 4 decomposes the drag along $\text{GRPO} \rightarrow \text{SFT} \rightarrow \text{DIAA}$ and reports a granularity-dial signature; Sec. 5 turns the paired certificate into the gate of Eq. (2); Sec. 6 evaluates its frozen and online instantiations against controls that rule out generic regularization and length-based heuristics.

2 Background and DIAA

Terminology. **Rejection-sampling 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. A **Python fence** is the Markdown delimiter ````python` that opens a code block. **Tool-Integrated Reasoning (TIR) form** denotes reasoning presented inside such a fence together with a claimed output block, with no actual execution. **Cargo-cult drag** is the accuracy recovered on the rescued paired cell when the intervention removes TIR form; the certificate additionally subtracts a never-TIR control to absorb generic suffix effects. **σ -weighted SFT/RL** denotes $\sigma(x)\mathcal{L}_{\text{SFT}} + (1 - \sigma(x))\mathcal{L}_{\text{RL}}$ with $\sigma(x) \in [0, 1]$: no sample is routed exclusively to one branch, and high σ favors imitation. 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. **ICA** is the intervention-derived gate c that modifies the per-sample weight as $\sigma_{\text{eff}} = \sigma c$, with a frozen and an online implementation.

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 Qwen2.5-1.5B 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

$$\begin{aligned}\mathcal{L} &= \sigma(x)\mathcal{L}_{\text{SFT}} + (1 - \sigma(x))\mathcal{L}_{\text{RL}}, \\ \sigma(x) &= \text{sigmoid}(-\beta \Delta(x)), \\ \Delta(x) &= \log p_{\theta}(\text{rft}) - \log p_{\theta}(\text{roll}).\end{aligned}\tag{1}$$

The full derivation is in App. B. Because Δ is computed from the current policy’s log-probabilities of one observed anchor trace and one on-policy rollout, DIAA trains and queries no separate discriminator network. 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, conditional on its stated assumptions, 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. DIAA is therefore an instrument for measuring amplification, not a claim about tool use.

3 Setup and Cross-Sectional Observation

Models and data. Qwen2.5-1.5B base (Yang et al., 2024) with LoRA following TINA’s (Wang et al., 2026) LoRA recipe; 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

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: MATH-500 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).

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 (Shao et al., 2024) 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, a lower TIR cross-section would not by itself certify a better method and a higher one would not prove harm. The paired cells below are the unit of causal evidence.

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 (App. G). The rescued cell (T, N) gives a controlled within-prompt contrast for suppressing TIR at fixed prompt difficulty; the never-TIR cell (N, N) bounds the generic effect of the suffix itself.

Main paired result. Tab. 2 shows that SFT contamination is intervention-stable (R_{15} has the same +8.5 pp rescued delta under both interventions), the clean teacher anchor repeats the signal under a generic suffix (+10.3 pp in TN versus +1.9 pp in NN), and SFT and σ contributions are comparable. Thus interventions targeting only the σ branch can recover at most part of the observed cargo-cult drag. The two suffixes agree in direction and magnitude on R_{14v4} as well (+10.6 versus +13.2 pp rescued, against -1.2 and -0.6 pp controls), which reduces generic prompt confounding but neither eliminates suffix dependence nor shows that TIR is the only affected behavior.

Granularity and subject heterogeneity. Truly-per-sample σ shows $1.5\text{--}1.7\times$ greater intervention persistence under the tested suffixes than per-prompt aggregation. This is consistent with the granularity dial modulating how prompt-sensitive the learned shortcut is. The certificate sign is subject-dependent: on Counting & Probability, TIR encodes genuine `math.comb` capability and the paired delta flips sign, motivating subject-conditional fallback in ICA. Detailed cell counts and the per-subject breakdown are in App. E. The drag is therefore real, decomposable, and unevenly distributed across subjects, which is what a gate must respect.

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

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

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).

Two instantiations of one gate. Eq. (2) is the shared principle; it admits two instantiations that differ only in how c is obtained. In the *frozen* instantiation (R_{20}), c is computed once before training from a paired-intervention probe on a prior checkpoint and held fixed (App. G, Alg. 2). In the *online* instantiation (R_{ICA}), a certify token is updated in-loop from paired rollouts (Alg. 1). We treat the online instantiation as our main adaptive implementation and the frozen one as the control that isolates the certificate content from the update rule.

In-loop certify-token update. At a certificate event, the same minibatch is rolled out 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 $c(x; \theta_c)$ for subsequent steps. Certificate adaptation runs on a slow timescale relative to the policy update (App. B.3); the corresponding fixed point is what motivates using c as an amplification certificate (App. C.5). Both instantiations therefore share one principle and differ only in when the certificate is read.

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, read *diagnostically*: Δ_{TN} is the accuracy recovered on the rescued cell when the intervention removes TIR, and Δ_{NN} the never-TIR control, so a *larger* Δ_{TN} means more removable drag is present. $\hat{C}_\phi = \Delta_{TN} - \Delta_{NN}$ 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), which reduces concern about a purely suffix-specific artifact. After mitigation the same quantity is read in the opposite direction (Sec. 6).

Algorithm 1 ICA with certify token in DIAA paradigm

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

- 1: **for** step $t = 1, \dots, T$ **do**
- 2: Sample minibatch $\mathcal{B} \subset \mathcal{D}$; roll out under τ_{base}
- 3: **if** $t \bmod K_c = 0$ **then**
- 4: Roll out a probe minibatch under τ_{int}
- 5: Compute paired Δ_{acc} on ϕ -activating prompts $\rightarrow$ signal s_t
- 6: Update $\theta_c \leftarrow \theta_c + \eta_c \nabla_{\theta_c} s_t$
- 7: **end if**
- 8: Compute $\sigma_{\text{eff}}(x) = \sigma(x) \cdot c(x; \theta_c)$ for $x \in \mathcal{B}$
- 9: Update policy θ on $\mathcal{L}_{\text{ICA}}$ (Eq. (2))
- 10: **end for**

Run	Role	Acc	TIR	Δ_{TN}
R_{03}	GRPO baseline	44.4	9.2	+2.2
R_{14v4}	ungated DIAA	47.5	29.0	+10.6
R_{22}	hard-filtered RFT	44.6	33.0	+9.0
R_{21}	shuffled certificate	46.1	24.5	+8.1
R_{24}	length heuristic	46.4	23.0	+7.9
R_{20}	ICA, frozen c	48.7	13.8	+4.9
R_{ICA}	ICA, online c	49.3	11.2	+3.8

Table 3: MATH-500 at step 1200. For R_{03} and R_{14v4} the Δ_{TN} column is *diagnostic* (drag present); for the remaining rows it is *residual* drag left after the intervention, where lower is better. R_{22} removes flagged RFT samples outright instead of downweighting them; R_{21} rules out generic regularization; R_{24} rules out length-based confounds.

6 Experiments

Main result. All runs share the configuration of App. D and are evaluated on MATH-500 at step 1200. We accept ICA as a *weak win* when accuracy gains at least 1 pp over ungated R_{14v4} , the paired Δ_{TN} at least halves, and Counting & Probability does not regress by more than 2 pp. The subject-conditional fallback of App. G leaves Counting & Probability ungated by construction, since its certificate is non-positive there. Both certificate-gated configurations clear the accuracy and residual-drag criteria (Tab. 3), gaining +1.2 and +1.8 pp accuracy over R_{14v4} while more than halving residual drag. The controls separate certificate *content* from mere downweighting. A shuffled certificate and a length heuristic both reduce TIR and residual drag, yet land *below* R_{14v4} in accuracy, and hard-filtering the flagged RFT samples instead of downweighting them is worse still on all three columns. The evidence therefore supports intervention-certified gating as such. The frozen–online gap is small, and we do not treat it as establishing an advantage for online adaptation.

Consistency across runs. Repeating the three-arm comparison twice more under the same protocol preserves the ordering of Tab. 3 in every run; App. F gives the three-run mean $\pm$ SD and the per-run differences. These are matched consistency checks rather than a training-variance estimate, so we draw no significance conclusion from the frozen–online difference.

Held-out transfer. We evaluated the same three arms on the GSM8K (Cobbe et al., 2021) test split with the MATH-500-derived gates unchanged and no retuning. Both gated arms carry the joint improvement to a set that played no part in constructing the certificate (Tab. 4), trading off slightly against each other.

Certificate cost. Gating adds one τ_{int} probe minibatch every K_c steps and no additional model; App. F gives the full accounting. Certified gating thus buys lower emission and lower residual drag at a small, bounded generation cost.

Arm	Acc	TIR	Δ_{TN}
ungated DIAA	67.9	24.7	+12.4
ICA, frozen c	70.7	9.3	+6.4
ICA, online c	70.4	9.8	+6.1

Table 4: GSM8K test ($n=1319$), certificate frozen from MATH-500 and not retuned. Δ_{TN} is residual drag, where lower is better.

7 Discussion

Amplification reactivates pretrained channels.

A clean cross-model anchor (DeepSeek-Math-7B-Instruct teacher, 0% TIR) still trains a student emitting TIR at 45.8% with $\hat{C}_\phi = +8.1$ pp drag (R_{15b} , Tab. 1–2): post-training selects which pretrained tendency to amplify, not merely which trace to imitate, so data-side scrubbing cannot reach the ceiling of App. C.4. We use “behavioral channel” operationally, as a gap between RFT-level prevalence and post-training emission rate, not as a claim about circuits.

Instantiating the gate elsewhere. What generalizes is the interface rather than the TIR detector. For a candidate behavior with a measurable indicator ϕ , one constructs a paired intervention that preserves the task while perturbing that behavior, then estimates the induced change in task performance against a never-feature control. The resulting certificate sets $c(x)$ in $\sigma_{\text{eff}} = \sigma c$. Output-format shortcuts and stereotyped refusal patterns are natural candidates wherever task-preserving counterfactuals and measurable outcomes can be specified. Theorem A is stated for an arbitrary binary feature, but the demonstrated evidence covers one feature in one model family; cross-feature evaluation remains future work.

What the certificate does and does not decide.

The detector only flags a candidate form; the paired evidence, not the detector, decides whether ICA intervenes. Borderline outputs are classified by the fixed Python-fence marker rather than by executability, and c stays at 1 unless the certificate supports suppression. The conservative LCB composition (App. G) prefers false-negative gating, leaving a cargo-cult sample unweighted, over false-positive gating, downweighting a beneficial one. The certificate is computed on the same eval set as the headline metric; three pre-registered choices bound the leakage (App. G), and the GSM8K transfer of Tab. 4 evaluates the gate on a set that played no

part in its construction.

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 a lightweight SFT cold start for subsequent RL. RAFT (Dong et al., 2023) selects high-reward samples for fine-tuning, while STaR (Zelikman et al., 2022) iteratively fine-tunes on successful self-generated rationales. On-Policy Distillation (Lu and Thinking Machines Lab, 2025; Li et al., 2026c) densifies teacher signal; Hübotter et al. (2026) densify environment feedback through self-distillation. In the reinforcement-finetuning line, CARE-RFT (Li et al., 2026a) replaces reverse-KL regularization with a skew reverse KL, and Song et al. (2025) document refusal degradation and increased hallucination on unanswerable questions. None has a sample-level validity gate. ICA composes with any per-sample-weighted variant.

Prompt priming, shortcut learning, and TIR.

Castricato et al. (2024) study the Pink Elephant problem for avoidance instructions, which motivates our positive-framing τ_{int} , and Vrabcová et al. (2025) document model fragility to negation. 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 presupposes that the signal it amplifies is capability-bearing. Using DIAA as a probe, we 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. Intervention-certified gating turns that diagnosis into control. A frozen and an online certify gate both raise accuracy while cutting TIR prevalence and residual drag, on MATH-500 and on held-out GSM8K. By contrast, shuffled and length-based controls cut TIR and drag without gaining accuracy, and hard filtering is worse on all three. ICA attaches to per-sample-weighted methods without altering their loss family.

Limitations

ICA currently certifies one audited trace feature at a time; other amplified channels remain ungated unless separately discovered and certified. The paired intervention measures a composite prompt-suffix effect, so attribution is not isolated, and the rescued subset is endogenously selected. Our experiments use a single model family, Qwen2.5-1.5B base, so the single-model scope remains even with the held-out GSM8K transfer. The three runs of Sec. 6 are matched consistency checks rather than an estimate of training variance, so we draw no significance conclusion from the frozen–online difference. Our setting evaluates 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*.
- Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. [Training verifiers to solve math word problems](#). *arXiv preprint arXiv:2110.14168*.
- 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 Hübner, Frederike Lübeck, Lejs Behric, Anton Baumann, Marco Bagatella, Daniel Marta, Ido Hakimi, Idan Shenfeld, Thomas Kleine Büning, Carlos Guestrin, and Andreas Krause. 2026. [Reinforcement learning via self-distillation](#). In *ICML*. ArXiv:2601.20802.
- Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. [Efficient memory management for large language model serving with PagedAttention](#). In *Proceedings of the 29th Symposium on Operating Systems Principles (SOSP)*.
- Shuozhe Li, Jincheng Cao, Bodun Hu, Aryan Mokhtari, Leqi Liu, and Amy Zhang. 2026a. [CARE-RFT: Confidence-anchored reinforcement finetuning 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-an 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*.
- Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. 2024. [Let’s verify step by step](#). In *ICLR*. ArXiv:2305.20050.
- Heng Lin and Zhongwen Xu. 2025. [Understanding tool-integrated reasoning](#). *arXiv preprint arXiv:2508.19201*.
- Kevin Lu and Thinking Machines Lab. 2025. [On-policy distillation](#).
- Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. 2024. [DeepSeekMath: Pushing the limits of mathematical reasoning in open language models](#). *arXiv preprint arXiv:2402.03300*.
- 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.
- Tereza Vrabcová, Marek Kadlčík, Petr Sojka, Michal Štefánik, and Michal Spiegel. 2025. [Negation: A pink elephant in the large language models’ 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.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, and 22 others. 2024. [Qwen2.5 technical report](#). *arXiv preprint arXiv:2412.15115*.

Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah D. Goodman. 2022. [STaR: Self-taught reasoner 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 & RL for reasoning](#). In *NeurIPS*. ArXiv:2506.17211.

A Reproducibility

We describe the 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. The main training hyperparameters, 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 (Yang et al., 2024) (Apache 2.0). Eval sets: MATH-500 (Lightman et al., 2024), the 500-problem subset drawn from MATH (Hendrycks et al., 2021) (MIT), and the GSM8K (Cobbe et al., 2021) test split; both are used for research purposes consistent with their intended use. Decoding for all reported evaluations uses vLLM (Kwon et al., 2023) with temperature 0.6, max new tokens 2048, and no chat template.

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 $\pi_{\text{anchor}}(\cdot | q)$ (the source of supervised target traces) and the current policy $\pi_{\theta}(\cdot | q)$ (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

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

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

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

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:

$$\log \pi_{\text{anchor}}(x | q) \approx \log p_{\theta}(\text{rft} | q), \quad (5)$$

$$\log \pi_{\theta}(x | q) \approx \log p_{\theta}(\text{roll} | q). \quad (6)$$

The substitution is a per-sample approximation: π_{anchor} enters only through the policy’s likelihood of the observed anchor trace. No separate discriminator network is trained or queried.

Defining the policy–anchor log-likelihood gap

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

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

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

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 $\beta \rightarrow \infty$ 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). The result below is a *conditional* guarantee: it holds under (A1)–(A4), which we state as idealizations rather than verify for the realized nonconvex LoRA optimization. None of the empirical claims in Sec. 4 or Sec. 6 depends on it.

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

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

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

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) $b_k/a_k \rightarrow 0$ as $k \rightarrow \infty$ (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 $\theta^*(\bar{\sigma})$, and the mapping $\bar{\sigma} \mapsto \theta^*(\bar{\sigma})$ is Lipschitz.

(A4) in particular is an idealization: a nonconvex LoRA training objective need not have a unique stationary point for fixed $\bar{\sigma}$.

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 $(\theta^*, \bar{\sigma}^*)$ satisfying

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

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

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

Proof sketch. By (A2), θ_k tracks $\theta^*(\bar{\sigma}_k)$ asymptotically. The slow recursion in Eq. (11) thus reduces to a stochastic approximation for the fixed point of $\bar{\sigma} \mapsto \mathbb{E}[\sigma_{\text{DIAA}}(x, q; \theta^*(\bar{\sigma}))]$. (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

Consider the frozen instantiation, which 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} . Under the analogous assumptions (A1)–(A4) for the gated objective, the same conditional convergence argument applies. The fixed point shifts from Eq. (13) to

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

reflecting that ICA-gated samples contribute a discounted SFT pull but the convergence guarantee survives. This inheritance argument covers the frozen gate only. The online instantiation learns θ_c during training, so its gate is not parameter-independent and is not covered by the argument above; we treat its slow-timescale update (Sec. 5) as a design motivation rather than a proved guarantee.

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

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

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

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,

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

Eq. (18) is the minimizer of $\text{KL}(\sigma \cdot \pi_{\text{anchor}} \parallel \pi_{\theta})$ over a model class rich enough to express $\sigma \cdot \pi_{\text{anchor}}$. It is an SFT-side object: it abstracts away both the RL pull (the $(1-\sigma) \mathcal{L}_{\text{RL}}$ branch of Eq. (9)) and finite-sample drift; both are flagged in App. C.7.

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

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

Proof. Marginalizing Eq. (18),

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

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

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 ,

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

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. $\square$

C.4 Main result

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

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

The statement concerns the idealized SFT-side fixed point of Eq. (18), not the full RL-inclusive objective of Eq. (9); see C4 below.

Proof. Apply Lemma 2 (Eq. (20)) to π_{θ^*} and to π_{anchor} in turn and subtract:

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

using Lemma 1 (Eq. (19)) in the second equality and rearranging signs. $\square$

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 $\text{Cov}_{\pi_{\text{anchor}}}(\sigma, f)$. 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 Qwen2.5-1.5B base self-rollouts, and $\text{Cov}_{\pi_{\text{anchor}}}(\sigma, f) > 0$ 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. This is a directional prediction about the SFT-side fixed point; Sec. 4’s pathway decomposition is the empirical counterpart that also covers the RL branch.

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 $\text{Cov}_{\pi_{\text{anchor}}}(\sigma, f) > 0$ (the ceiling penalty is active), that gating preserves the sign of the covariance, $\mathbb{E}_{\pi_{\text{anchor}}}[\sigma s \mid f=1] > \mathbb{E}_{\pi_{\text{anchor}}}[\sigma s \mid f=0]$, and that the gate downweights feature-bearing samples on average:

$$\mathbb{E}_{\pi_{\text{anchor}}} [s \mid f=1] < \mathbb{E}_{\pi_{\text{anchor}}} [s \mid f=0]. \quad (22)$$

Then

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

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

$$\begin{aligned}\text{Cov}(\sigma, f) &= \mu(1 - \mu)(A - B), \\ \text{Cov}(\sigma s, f) &= \mu(1 - \mu)(s_1 A - s_0 B).\end{aligned}$$

The hypothesis $\text{Cov}(\sigma, f) > 0$ 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 $\text{Cov}(\sigma, f) - \text{Cov}(\sigma s, f) > 0$, so the strict reduction needs only $A > B$ and $s_1 < s_0$. Positivity of $\text{Cov}(\sigma s, f)$ is the remaining hypothesis $s_1 A > s_0 B$. $\square$

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. We state them as predictions; the quantitative agreement of P1 and P2 is not established by the experiments reported here.

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

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

Checking this requires an anchor on which the literal detector is non-degenerate. Our main R_{14v4} anchor is the cleaned RFT, which by construction has $p_f(\pi_{\text{anchor}})=0$ under the literal Python-fence detector, so the anchor-side covariance is degenerate there and the quantitative check must instead be run on the raw-RFT arm or with a non-literal detector. We leave that measurement to extended work and do not treat the magnitude predicted by P1 as empirically validated.

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 $\text{Cov}_{\pi_{\text{anchor}}}(\sigma, f) - \text{Cov}_{\pi_{\text{anchor}}}(\sigma w, f)$. 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 $\text{Cov}_{\pi_{\text{anchor}}}(\sigma, f') \approx 0$, Theorem A predicts no σ -amplification, i.e. $p_{f'}(\pi_{\theta^*}) \approx p_{f'}(\pi_{\text{anchor}})$. 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 base 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 $\text{Cov}(\sigma, f)$) to be robust, while the magnitude is approximate.

C2. σ self-consistency. $\sigma(x) = \text{sigmoid}(-\beta \Delta(x))$ 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, under its assumptions, that $\bar{\sigma}_k$ 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 α_f, α_{-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

$$\begin{aligned} & |\varepsilon_{\neg f}(1 - p_f) - \varepsilon_f p_f| \\ & \leq \max(|\varepsilon_f|, |\varepsilon_{\neg f}|) \cdot \max(p_f, 1 - p_f). \end{aligned} \quad (24)$$

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 = σ -RL-only, pure SFT = σ -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 base 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 omitted for space; the two effects that drive the analysis are summarized below.

Granularity dial. Define the intervention-persistent 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% persistence under τ_+/ τ_g , versus 22%/37% for per-prompt-aggregate R_{14v4} , a $1.5\text{--}1.7\times$ ratio. Finer-grained σ therefore leaves behavior that the intervention removes less often, which we use as an operational measure of entrenchment under the tested suffixes.

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 Additional experimental detail

Hard-filtered RFT (R_{22}). App. A lists a hard-filter mode that deletes flagged RFT samples rather than downweighting them. Trained with the same seed as the runs of Tab. 3, it reaches 44.6% accuracy, 33.0% TIR, and +9.0 pp residual Δ_{TN} ($n_{TN}=133$), with a never-TIR control of -2.6 pp ($n_{NN}=265$), so $\hat{C}_\phi = +11.6$ pp. Deleting the flagged samples therefore removes supervision without removing the amplified channel, which is consistent with the teacher-anchor result of Sec. 7.

Certificate overhead. The frozen instantiation requires one offline paired evaluation, whose marginal cost is a single additional τ_{int} pass when the τ_{base} outputs already exist; training itself adds no paired generation. The online instantiation adds one τ_{int} probe minibatch every K_c steps, since the ordinary rollout already supplies τ_{base} . Outside this extra generation, the certificate logic is aggregate paired statistics, the certify-token update, and a multiplication by $c(x)$, and invokes no additional model. Neither instantiation trains or queries a separate discriminator network.

Run-to-run consistency. The three-run summary of Sec. 6 combines the run reported in Tab. 3 with two further matched runs of all three arms under the same protocol, labelled 101 and 102. Descriptive mean $\pm$ sample SD over the three runs, for accuracy / TIR / residual Δ_{TN} ,

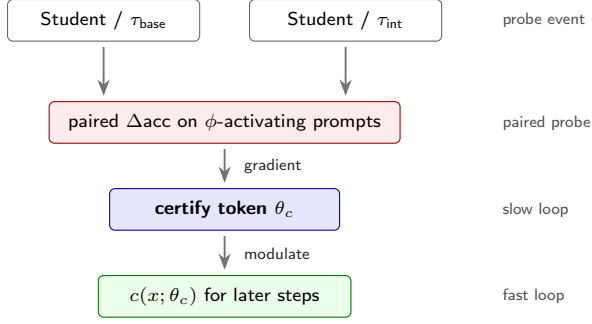


Figure 2: **Certify-token mechanism.** At a certificate event, paired rollouts update θ_c from Δacc on ϕ -activating prompts; the updated token gates subsequent steps via Eq. (2).

is 47.3 ± 0.3 / 29.4 ± 0.5 / $+10.9 \pm 0.4$ for ungated DIAA, 48.6 ± 0.3 / 14.3 ± 0.6 / $+5.2 \pm 0.4$ for frozen- c ICA, and 49.2 ± 0.3 / 12.0 ± 0.8 / $+4.1 \pm 0.3$ for online ICA. The paired online-minus-frozen differences are $(+0.6, -2.0, -1.1)$ pp under 101 and $(+0.7, -2.2, -1.2)$ pp under 102, and both reruns reproduce the submitted ordering. We report these as matched consistency checks rather than a training-variance estimate, and we do not interpret the small accuracy differences as evidence of online-over-static superiority.

G Frozen- c baselines and prompts

The online instantiation (Sec. 5) updates the certify token θ_c during training. The *frozen- c* instantiation computes c once before training from a paired-intervention probe on a prior checkpoint and holds it fixed; the same construction, with a deliberately corrupted or substituted certificate, yields the R_{21}/R_{24} controls of Tab. 3.

Feature detector ϕ . $\phi(x) = 1$ [“python appears in the text of x], a literal substring detector fixed before any ICA training run; alternative detectors (broader Python-fence regexes, neural classifiers, learned indicators) are admissible but not used here. The detector only nominates candidates: whether a nominated sample is actually downweighted is decided by the paired certificate below.

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 (T, T)/(T, N)/(N, T)/(N, N) cells by the pair $(\phi(y_{base}), \phi(y_{int}))$. We compute

- $\hat{C}_\phi = \text{LCB}(\Delta_{TN} - \Delta_{NN})$, 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 $(\hat{C}_\phi, p_{McN})$ are computed for each MATH subject with $n \geq 30$ paired prompts; otherwise the global estimate is used. With thresholds $\tau_C = \tau_p = 0.05$, the certificate is binary:

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

The thresholds 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 $\hat{C}_\phi \leq 0$: on Counting & Probability the certificate is non-positive, so c stays at 1.0 and those samples are never downweighted.

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

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

- 1: Evaluate π_0 on $\mathcal{M}$ under τ_{base} to obtain $\{y_{base}\}$
- 2: Evaluate π_0 on $\mathcal{M}$ under τ_{int} to obtain $\{y_{int}\}$
- 3: Partition $\mathcal{M}$ into cells by $(\phi(y_{base}), \phi(y_{int}))$
- 4: $\hat{C}_\phi \leftarrow \text{LCB}(\Delta_{TN} - \Delta_{NN})$, $p_{McN} \leftarrow \text{McNemar exact } p \text{ on } (T, N)$
- 5: **for** each subject s with $n_s \geq 30$ **do**
- 6: Compute subject-conditional $(\hat{C}_\phi^{(s)}, p_{McN}^{(s)})$
- 7: **end for**
- 8: **return** c assigning 0.3 when $\phi(x) \wedge \hat{C}_\phi^{(\text{subject})} > \tau_C \wedge p_{McN}^{(\text{subject})} \leq \tau_p$, 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 MATH-500 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. The GSM8K transfer of Tab. 4 additionally evaluates the gate on a set that played no part in its construction.