CREDIT WHERE REASONING IS DUE: LEARNING WHEN TO REASON VIA TREATMENT-EFFECT-AWARE CREDIT ASSIGNMENT

Anonymous authors
Paper under double-blind review

ABSTRACT

Reinforcement learning has made language models increasingly capable of reasoning, yet *when* to reason remains largely unaddressed. We show that standard group-relative RL cannot learn this decision when every rollout in a group reasons, as happens on-policy: the per-input effect of reasoning is absent from the policy gradient, regardless of the outcome reward. Forcing both modes into each group makes this counterfactual identifiable, but is not sufficient: the resulting signal must credit reasoning for its effect rather than for the output's quality, and it must remain reliable under noisy per-input comparisons. We address both by formulating selective reasoning as a treatment-effect learning problem and propose Treatment-Effect-Aware Credit Assignment (TECA), which combines forced 50/50 mode mixing with importance correction, a reward that scores both modes on the same scale, and a persistent empirical-Bayes estimate of whether reasoning beats direct generation on each input. That estimate gates reasoning-token credit and sets the target for the mode decision. On machine translation with Qwen3-8B, where reasoning barely moves mean quality yet widens its distribution, TECA matches or beats always-reasoning baselines on most language pairs while reasoning on selective inputs, and significantly outperforms the same objective without the gate. Learning when to reason therefore requires explicit treatment-effect credit, not only outcome-based RL or exposure to both modes.

1 INTRODUCTION

Reinforcement learning has taught language models how to reason (Wei et al., 2022; DeepSeek-AI et al., 2025), but not *when to*. Where reasoning always helps, as in verified math and code, reasoning on every input only costs tokens, and adaptive methods there learn how much to reason (Zhang et al., 2025; Huang et al., 2026; Xiang et al., 2025). In open-ended generation scored by a continuous metric, reasoning can improve one input and degrade another, so a capable policy should reason selectively. Learning that per-input decision is a credit-assignment problem. This paper shows why the standard objective cannot solve it, and what does.

Machine translation is a clean setting (Feng et al., 2025; Ye et al., 2026) for the question, and an unsettled one: reasoning has been reported to help on context-dependent text (Ataee et al., 2026) and to hurt when a model is pushed past its natural stopping point (Li et al., 2026). On Qwen3-8B both hold at once. Averaged over sentences, reasoning moves quality by at most $+0.021$ COMET-22 on any of three text tiers (figurative, conversational, and formal text). Yet on a given sentence it often produces the better translation: the best of twelve reasoning samples outscores the best of twelve direct ones on 48–77% of sources, while a single reasoning sample wins on only 33–49% (Figure 2). Reasoning widens the distribution of translation quality without moving its centre. A policy that emits one sample cannot pick from that upper tail, so the question it must answer is per input: on which sentences does one reasoning sample beat one direct sample more often than not?

The decision itself is a single token. A hybrid reasoning model's answer opens with a `<think>` tag already in place; if its first generated token is the closing tag, the block stays empty and the model translates directly, otherwise it reasons first. The policy therefore has one binary choice per input, and the reward should be enough to teach it: sample the input both ways and reinforce whichever

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

came out better. Group-relative methods such as GRPO (Shao et al., 2024) are the standard recipe for exactly this.

It does not work, because the comparison never happens. Whether reasoning helped a sentence is a counterfactual question: *would the translation have been better without it?* Answering it needs both translations in the same group, since GRPO scores each rollout against its group mean. We observe that the base Qwen3-8B policy assigns approximately 10^{-8} probability to skipping the think block, leaving virtually all sampled groups *mode-homogeneous*. Lemma 1 makes this exact. The per-input effect of reasoning enters the advantage weighted by the fraction of the group that did *not* reason, which is zero, and no choice of reward changes that (Corollary 1). The update is not noisy about when to reason; it is silent.

Forcing both modes into every group restores the missing comparison. Similar mode-mixing strategies are also used by existing adaptive-reasoning methods in verified domains (Zhang et al., 2025). However, it is not sufficient. Two further requirements follow. First, the reward must score both modes on the same scale, or the measured difference is partly a property of the metric rather than of the model. Second, the treatment effect is estimated from only a few stochastic rollouts, making the estimate noisy. The resulting gate should therefore be bounded and non-negative: uncertainty may reduce reasoning credit, but should not reverse the direction of the outcome-based update. Mixing solves identification; the two requirements govern how a noisy estimate is allowed to assign credit.

Treatment-Effect-Aware Credit Assignment (TECA) follows all three. Each source is sampled into a group with both modes forced present, with an importance correction for the forcing, and every candidate is scored by one reference-based metric, COMET-22. From those scores we estimate a per-source win rate: how often a reasoning sample outscores a direct one. One group estimates it noisily, so paired comparisons accumulate across training steps and are read through an empirical-Bayes posterior that shrinks toward the population rate while a source's own evidence is thin. The posterior is used twice. Its probability that reasoning wins scales the credit assigned to that source's reasoning tokens, floored so that a source on which reasoning is currently losing keeps the gradient that could improve it; and its mean is the target for the first-token decision (Section 3). Related work is discussed in Appendix A.

The gate carries the result. Against Mixed GRPO (ungated), which differs from TECA only in crediting every reasoning token in full, TECA wins on XCOMET-XL and COMET-22 on every language pair at 8B (paired bootstrap, $p < 0.05$). Removing the gate does not merely lose that gain: the ungated objective drives quality below the untrained model on two of three tiers and, at 4B, collapses into a policy that emits no valid translation. TECA reaches its quality while reasoning on 45% of inputs, 54–60% fewer effective reasoning tokens than the always-reason base model or GRPO, and out of distribution on WMT24++ it beats GRPO on every pair on XCOMET-XL and chrF++ while reasoning on 31–47% of inputs.

Contributions. (1) A decomposition of the group-relative advantage showing that the per-input effect of reasoning is unidentifiable from mode-homogeneous groups under any outcome reward (Section 2). (2) TECA: forced mode mixing, a common-scale reward, and an empirical-Bayes win-rate posterior that gates reasoning-token credit and targets the mode decision, with a condition under which gating beats ungated credit (Section 3). (3) Evidence on three language pairs and two model scales that the gate, not mixing, carries the quality result, and that ungated mode competition is actively harmful (Section 4). The experiments ask two questions: does selective reasoning win on quality and cost at once, and is the gain due to the gate (RQ1); and is the learned decision selective, and does it transfer across domain and scale (RQ2)?

2 Why Group-Relative RL Cannot Learn When to Reason

2.1 Reasoning as an Input-Dependent Treatment

Let x be a source sentence and $m \in \{0, 1\}$ the generation mode: $m{=}1$ reasons before translating, $m{=}0$ translates directly. Let τ be the reasoning trace generated for the input x and y be its corresponding translation. A hybrid policy factorizes as $\pi_\theta(\tau, y \mid x) = \pi_\theta(\tau \mid x)\, \pi_\theta(y \mid x, \tau)$ in think mode and sets $\tau = \varnothing$ in no-think mode. The mode is decided by the first generated token: $y_1 =$ `</think>` leaves the reasoning segment empty, any other token opens it. We write

$\pi(x) := \pi_\theta(y_1 \neq \texttt{</think>} \mid x)$ for the probability of reasoning on x, and $Y(m, x)$ for the translation produced under mode m. With Q a translation-quality metric, the per-input effect of reasoning is

$$\Delta(x) := \mathbb{E}[Q(x, Y(1, x))] - \mathbb{E}[Q(x, Y(0, x))]\,, \tag{1}$$

where the expectation is over the policy's sampling in each mode. Because the policy changes during training, $\Delta(x)$ changes with it: it is a property of the current policy, not a fixed property of the sentence. The deployed policy emits one translation, so what it needs to know is not the size of $\Delta(x)$ but whether one reasoning sample tends to beat one direct sample. We therefore track the *win-rate* and its margin over one-half,

$$\begin{aligned} p(x) &= \Pr\big[Q(x, Y(1, x)) > Q(x, Y(0, x))\big] + \tfrac{1}{2}\Pr\big[Q(x, Y(1, x)) = Q(x, Y(0, x))\big], \\ \delta(x) &= p(x) - \tfrac{1}{2}, \end{aligned} \tag{2}$$

so that $\delta(x) > 0$ means reasoning wins more often than it loses on x. Ties count as half a win here and in every estimator below. Appendix J explains why $\delta(x)$, not $\Delta(x)$, is the right target for a single-sample policy, and why it is also the stabler one under a learned metric whose scale drifts during training. Both quantities are so far only defined. The next subsection measures them on the base policy and finds that they disagree: $\Delta(x)$ is near zero, while $\delta(x)$ varies across sources in a way a selective policy could exploit.

2.2 Reasoning Changes the Spread of Quality, Not Its Level

We measure the effect of reasoning on the Qwen3-8B base model over three translation directions (English→Chinese, German, and Swahili) and three text tiers: figurative text (IdiomTranslate30), conversational dialogue (OpenSubtitles), and formal single-register text (FLORES-200). For each source we draw $K{=}12$ samples in each mode and score every sample against the reference with COMET-22 (Rei et al., 2022), the metric that later serves as the training reward. Per-tier effect sizes, tie rates, and generation-failure rates are logged in Table 4 (Appendix G).

Finding 1: the mean effect is near zero, and negative once reasoning's own failures are counted. Conditional on both modes emitting a well-formed translation, reasoning's mean effect never exceeds $+0.021$ COMET-22 on any tier. Counting its failures reverses the sign: reasoning samples fail to produce a translation on 8–18% of draws, against 0% for direct translation, and the unconditional effect is -0.05 to -0.09. Nearly every failure ($2378/2379$) is the model failing to close its reasoning segment before the token budget runs out (Appendix I).

Finding 2: reasoning changes the spread of quality, not its level. Figure 2(b) compares the two modes draw for draw. For each source we sort its twelve think scores and its twelve no-think scores from worst to best, then match them by position: best against best, median against median, worst against worst. On figurative text and FLORES, reasoning wins the best draw on 77% and 76% of sources but the worst draw on only 27% and 23%, and the median on under half; on OpenSubtitles it wins even the best draw on only 48%. Winning the top rank while losing the bottom one is the signature of a wider distribution around the same centre: a mode that were simply better would win at every rank. Panel (a) shows the same effect as a function of K: the chance that the best of K reasoning samples outscores the best of K direct samples rises from about 0.49 at $K{=}1$ to 0.77 at $K{=}12$, whereas a single reasoning sample beats a single direct one on only 0.33–0.49 of sources. Reasoning widens the distribution in both directions. A policy that emits one sample per input pays the lower tail without being able to select from the upper one, so what matters to it is not the mean effect but the single-sample win rate $p(x)$ of Eq. (2), and whether $\delta(x) > 0$ on a given source. $\Delta(x)$ and $\delta(x)$ need not even agree in sign, since reasoning has the heavier upper tail and the lower median; that is why the rest of the paper estimates $\delta(x)$.

Finding 3: short sources make reasoning redundant, not harmful. OpenSubtitles has the lowest single-sample win rate (0.33 against 0.49 on the other tiers, counting ties as non-wins as in Figure 2), and the tie column explains why: the two modes score *identically* on 16.5% of its comparisons, against 0.2% and 1.7% elsewhere. Its median source is five words, neither other tier contains a source that short, and within OpenSubtitles the win rate rises and the tie rate falls monotonically with length; at matched length the gap to the other tiers narrows (Appendix H). On a one-word source there is nothing to deliberate about, so both modes return the same translation and reasoning costs only the deliberation. Whether reasoning helps is therefore a property of the individual source and the next subsection shows that group-relative RL cannot see it.

2.3 Where the Effect Enters the Advantage

Consider a source x for which the current policy generates N think rollouts and M no-think rollouts. Let $\bar{Q}_1$ and $\bar{Q}_0$ be the sample-mean quality of each mode and $\Delta Q(x) = \bar{Q}_1 - \bar{Q}_0$ the group's estimate of $\Delta(x)$. Let A_i^{std} be the centred group-relative advantage of a think rollout computed over the N think rollouts alone, and A_i^{mix} the same advantage computed over the pooled group of $N + M$.

Lemma 1 (Advantage Decomposition under Mode Mixing)**.** *For a think rollout i,*

$$A_i^{\text{mix}} \;=\; A_i^{\text{std}} \;+\; \frac{M}{N+M}\,\Delta Q(x). \tag{3}$$

Proof. The pooled mean is $\bar{Q} = (N\bar{Q}_1 + M\bar{Q}_0)/(N+M)$, so $A_i^{\text{mix}} = Q_i - \bar{Q} = (Q_i - \bar{Q}_1) + (\bar{Q}_1 - \bar{Q})$ with $\bar{Q}_1 - \bar{Q} = \frac{M}{N+M}(\bar{Q}_1 - \bar{Q}_0)$. □

The effect of reasoning enters the advantage through the second term alone, weighted by the fraction of the group that did not reason. At $M = 0$ the term vanishes and the update is exactly the one obtained from think rollouts alone. Dividing by the pooled standard deviation, as the standardized advantage does, rescales both terms and changes nothing here (Appendix D).

Corollary 1 (Mode-Homogeneous Blindness)**.** *If every rollout in a group uses the same mode, the expected update does not depend on $\Delta(x)$: changing the quality of the unsampled mode changes neither the group, nor its advantages, nor the update. This holds for the centred and the standardized advantage, and for every outcome reward Q.*

Proof. Every rollout, every advantage, and the group standard deviation are computed on the sampled mode alone, while $\Delta(x)$ also depends on the unsampled mode's expected quality, which can vary freely without changing any of them. □

The problem is not that the reward is uninformative. The other mode is absent from the group, and no refinement of how outcomes are measured can recover information that was never sampled. On-policy this is the standard case: Qwen3-8B emits `</think>` first with probability of order 10^{-8}, so every group reasons and $M = 0$ for every source.

2.4 Mode Mixing Is Necessary but Not Sufficient

Lemma 1 names the first requirement: both modes must appear in the group for the same source. Once they do, their relative outcomes enter the pooled advantage through Eq. (3), and the cross-mode comparisons in the group estimate $p(x)$. Two problems remain.

First, the comparison is only valid if Q scores a candidate the same way regardless of which mode produced it. Verified rewards in math and code have this property trivially. A learned reference-free metric need not: we observe that COMETKiwi (Rei et al., 2023) scores no-think translations of templatic, low-ambiguity sources above think translations of the same sources, and a win rate built on such a score cannot separate the effect of reasoning from the metric's own mode dependence. Anchoring the score to a human reference, as COMET-22 does, removes that route (Appendix K).

Second, each source is represented by a handful of stochastic rollouts, so the group's estimate of $p(x)$ is noisy and can have the wrong sign even when both modes are present. Used as unrestricted credit, such an estimate can reinforce reasoning on sources where it hurts. Mixing solves identification; the remaining requirements govern how the identified but noisy effect may assign credit. A method that learns when to reason must therefore satisfy: **R1:** both modes appear in every group for the same source (Lemma 1); **R2:** the reward scores a candidate the same way whichever mode produced it; **R3:** the per-source estimate enters the update only through a bounded, non-negative factor, so that noise can shrink an update but never reverse it. The next section builds the method around R1–R3 and, once the gate is defined, states the condition under which gating beats ungated credit.

3 TREATMENT-EFFECT-AWARE CREDIT ASSIGNMENT (TECA)

Group-relative RL is memoryless: a group is sampled, scored, turned into an update, and discarded. TECA adds one piece of persistent state, a running count per source of how often reasoning has beaten not reasoning, and every other component exists to produce, maintain, or consume it. We follow one training step in order, marking where each of the requirements R1–R3 of Section 2.4 is met; Algorithm 1 (Appendix C) lists the step.

Notation. G rollouts receive gradient per source, N_x in think mode and $M_x = G - N_x$ in no-think, as in Lemma 1. Further no-think rollouts are scored but never backpropagated through. The persistent state is (W_x, n_x): accumulated wins and effective comparisons.

Step 1: sample the group (meets R1). Left to itself the policy never produces a no-think rollout, so the mode is forced. Each of the G slots is assigned think or no-think by a fair coin; if the coin returns a single-mode group (about 12.5% of the time at $G{=}4$), one slot is flipped, so both modes are present. A no-think slot has `</think>` spliced in as its first token; a think slot samples its first token from the policy restricted to every other token, so only the mode is forced, not the reasoning. The mode token's importance ratio is taken on the mode event against the proposal, $\rho_i^{y_1} = \pi_\theta(m_i \mid x)/\frac{1}{2}$. No-think candidates are then topped up to eight per source; the extra ones carry no reasoning segment, so they are cheap, and they are scored but never backpropagated through, sharpening the comparison without changing the group that the advantage is computed on. A rollout whose reasoning segment never closes has no translation and is dropped from everything that follows; if that leaves a group single-mode, it contributes no comparison that step.

Step 2: score, and compute the advantage (meets R2). Every candidate in both modes is scored against the reference with COMET-22, so a translation is scored the same way whichever mode produced it. The pooled group-relative advantage is standardized over the G rollouts with both modes in the pool,

$$A_i = \frac{Q_i - \bar{Q}}{\max(\hat{\sigma}_Q,\, \varepsilon)}, \qquad \bar{Q} = \frac{1}{G}\sum_{j=1}^{G} Q_j, \tag{4}$$

where Q_i is the COMET-22 score less a fixed penalty for a malformed output, $\hat{\sigma}_Q$ is the group standard deviation, and ε a floor guarding the degenerate case; both statistics are taken over the rollouts that produced a translation. Pooling both modes into $\bar{Q}$ is what puts the mode contrast of Lemma 1 into A_i.

Step 3: update the per-source evidence. Write $\mathcal{T}_x$ for the think candidates and $\mathcal{N}_x$ for all no-think candidates, auxiliary ones included. Their graded win fraction over all cross-pairs,

$$u_x = \frac{1}{|\mathcal{T}_x||\mathcal{N}_x|}\sum_{i\in\mathcal{T}_x}\sum_{j\in\mathcal{N}_x}\Big(\mathbf{1}[Q_i > Q_j] + \tfrac{1}{2}\mathbf{1}[Q_i = Q_j]\Big), \tag{5}$$

is the Mann–Whitney estimate (Mann & Whitney, 1947) of $p(x)$, computed on the unpenalized scores. The pairs reuse the same rollouts, so the evidence is incremented by an effective sample size rather than by the number of pairs,

$$(W_x,\, n_x) \;\leftarrow\; \gamma\,(W_x,\, n_x) \;+\; n_{\text{eff}}\,(u_x,\, 1), \qquad n_{\text{eff}} = \min\big(|\mathcal{T}_x|,\, |\mathcal{N}_x|\big), \tag{6}$$

so that correlated comparisons cannot make the posterior more confident than the evidence supports; widening the no-think side sharpens u_x without inflating n_{eff}. The discount $\gamma < 1$ is applied per visit and lets the state track a win rate that moves as the policy trains. At our two-epoch schedule each source is visited about twice, so n_x reaches only about 4 and the posterior below sits near even weight between the prior and the source's own evidence (Appendix E).

Step 4: read the posterior, twice. A single group supplies a handful of comparisons, so the counts are read through a posterior rather than at face value. Treating W_x as a pseudo-count with $W_x \mid p(x) \sim \text{Binomial}(n_x, p(x))$ and a population prior $p(x) \sim \text{Beta}(a, b)$, the posterior is conjugate, and the gate is the posterior probability that reasoning wins more often than it loses,

$$w_x = \Pr\big(\delta(x) > 0 \mid W_x, n_x\big) = 1 - I_{1/2}\big(a + W_x,\; b + n_x - W_x\big), \tag{7}$$

with $I_{1/2}$ the regularized incomplete beta function. Its sharpness is the posterior's concentration, which grows with n_x; the only free choice is the prior's concentration $\kappa = a + b$, which sets how much evidence a source needs before its own comparisons outweigh the population rate (Appendix F). The gate is floored,

$$\tilde{w}_x = w_{\min} + (1 - w_{\min})\, w_x, \qquad w_{\min} = 0.25, \tag{8}$$

so that credit on a source where reasoning is currently losing falls toward the floor rather than to zero and the reasoning mode stays trainable while the estimate settles. The same posterior is read a second time: its mean $\hat{p}_x = \mathbb{E}[p(x) \mid W_x, n_x]$ estimates the win rate itself rather than its sign, and becomes the target for the mode decision in Step 5.

Step 5: the loss. Writing $\mathcal{C}(\rho, a) = \min\big(\rho a,\ \mathrm{clip}(\rho, 1-\epsilon, 1+\epsilon)\, a\big)$ for the PPO-clipped surrogate and $\rho_i^{\mathrm{ans}}, \rho_i^{\mathrm{rsn}}, \rho_i^{y_1}$ for the probability ratios on the answer span, the reasoning span, and the mode token, the per-rollout loss has three surrogates and two regularizers,

$$\begin{aligned} \mathcal{L}_i &= -\,\mathcal{C}\big(\rho_i^{\mathrm{ans}},\ A_i\big) \;-\; \mathbf{1}[m_i = 1]\,\mathcal{C}\big(\rho_i^{\mathrm{rsn}},\ \tilde{w}_x A_i\big) \;-\; \mathcal{C}\big(\rho_i^{y_1},\ A_i\big) \;+\; \lambda\,\mathcal{L}_{\mathrm{bal}}(x) \;+\; \beta\,\mathrm{KL}_i, \\ \mathcal{L}_{\mathrm{bal}}(x) &= -\,\big[\hat{p}_x \log \pi(x) + (1 - \hat{p}_x)\log(1 - \pi(x))\big], \end{aligned} \tag{9}$$

where KL_i is the log-ratio of π_θ to the frozen base policy summed over the generated tokens other than y_1, the light regularizer ($\beta = 0.001$) standard in group-relative RL. The table lists which tokens each term trains and where the posterior enters: The answer span is trained on the pooled advantage in both modes, since translation quality is trainable however it was produced. The reasoning span is trained on the gated advantage; the next paragraph says why the gate multiplies. The mode token lies outside both spans and has its own surrogate on the realized mode event against the $\frac{1}{2}$ proposal. Its expected gradient carries a factor $\pi(1-\pi)$ and is near zero at the $\pi \approx 1$ corner the base policy starts from (Appendix L); the balance term supplies gradient there, with logit-gradient $\pi - \hat{p}_x$, order one at the corner and zero once the think probability reaches the source's estimated win rate. Its target is the source's posterior mean rather than a constant $\frac{1}{2}$, so each source is pulled toward its own evidence. Advantage rescaling and the periodic refit of the prior are in Appendix D.

Table 1: Credit assignment in the TECA loss (Eq. (9)). The win-rate posterior affects only the reasoning span, through the floored gate $\tilde{w}_x$, and the mode token, through the balance target $\hat{p}_x$.

Tokens	Present in	Credit	Posterior enters?
Answer span	both modes	A_i	no
Reasoning span	think rollouts	$\tilde{w}_x A_i$	yes: floored $\Pr(\delta > 0)$
Mode token y_1, surrogate	both modes	A_i, ratio $\pi_\theta(m_i \mid x)/\frac{1}{2}$	no
Mode token y_1, balance	both modes	pull $\pi(x)$ toward $\hat{p}_x$	yes: posterior mean

Why the gate multiplies rather than adds (meets R3). The per-source estimate could enter reasoning-token credit as a bonus added to the advantage or as a factor multiplying it. A bonus lets a noisy estimate reverse an update; a factor in $[0, 1]$ can only shrink one, which is what R3 asks. The following result says when shrinking pays.

Proposition 1 (Selective Attenuation)**.** *Write $\bar{A}(x) = \mathbb{E}[A \mid x]$ and consider reasoning-token credit of the form $g_c(x) = c(x)\bar{A}(x)$ with $c(x) \in [0, 1]$, where c is computed from comparisons accumulated independently of the rollouts receiving the credit. Take as target $g^\star(x) = \bar{A}(x)\,\mathbf{1}[\delta(x) > 0]$, which keeps full credit where reasoning wins more often than it loses and none where it loses more often. Then*

$$\mathbb{E}\big[(g_c - g^\star)^2\big] = \mathbb{E}\big[\bar{A}^2(1-c)^2\mathbf{1}_{\delta>0}\big] + \mathbb{E}\big[\bar{A}^2 c^2 \mathbf{1}_{\delta<0}\big], \tag{10}$$

and gating with $c = w_x$ improves on ungated credit $c \equiv 1$ if and only if

$$\mathbb{E}\big[\bar{A}^2(1-w)^2\mathbf{1}_{\delta>0}\big] < \mathbb{E}\big[\bar{A}^2(1-w^2)\mathbf{1}_{\delta<0}\big]. \tag{11}$$

As the posterior concentrates ($n_x \to \infty$ with consistent counts), $w_x \to \mathbf{1}[\delta(x) > 0]$ wherever $\delta(x) \neq 0$, the left side vanishes, and the inequality holds strictly whenever $\Pr[\delta(x) < 0,\ \bar{A}(x) \neq 0] > 0$.

Proof. On $\{\delta > 0\}$, $g^\star = \bar{A}$ and the pointwise error of g_c is $\bar{A}^2(1-c)^2$; on $\{\delta < 0\}$, $g^\star = 0$ and it is $\bar{A}^2 c^2$. Summing gives Eq. (10); $c \equiv 1$ leaves $\mathbb{E}[\bar{A}^2\mathbf{1}_{\delta<0}]$, and comparing gives Eq. (11). The limit is concentration of the Beta posterior at $p(x)$. □

Table 2: The same 8B systems evaluated two ways. *Left*: in-distribution, broken down by text tier, pooled across the three language pairs. *Right*: out-of-distribution on WMT24++ (news, social, speech, literary; zero overlap with training), broken down by language pair, pooled across tiers. $K{=}8$ sampled, temperature 0.7. Best result per column in bold.

	In-distribution, by tier						*WMT24++*, by language pair					
	Figurative		OpenSubtitles		FLORES		en→zh		en→de		en→sw	
Method	XC	chrF++	XC	chrF++	XC	chrF++	XC	chrF++	XC	chrF++	XC	chrF++
Base (no-think)	0.605	21.1	0.765	33.9	0.674	37.0	0.773	**30.1**	0.827	49.8	0.280	21.6
Base (think)	0.624	20.9	0.767	33.6	0.684	37.4	0.779	29.0	0.872	49.9	0.282	22.6
GRPO (always think)	**0.661**	21.4	0.780	29.3	**0.706**	36.7	0.790	25.5	0.881	46.5	0.283	23.7
Mixed GRPO (ungated)	0.636	22.0	0.720	32.5	0.670	36.8	0.728	25.8	0.869	48.5	0.242	24.6
TECA	0.658	**22.7**	**0.781**	**33.9**	0.702	**38.4**	**0.818**	29.9	**0.893**	**50.4**	**0.296**	**26.2**

The first term of Eq. (10) is the cost of withholding credit where reasoning helps; the second is the cost of paying it where reasoning hurts. Ungated credit pays the second in full. Three caveats separate the proposition from the algorithm. At finite n_x the gate is a trade, and Eq. (11) is the condition under which it pays; at our horizon $n_x \approx 4$, so the gate is deliberately conservative. The floor of Eq. (8) keeps the second term from vanishing, a price paid to keep the reasoning mode trainable. And a step's comparisons are added to (W_x, n_x) before its gate is read, so the independence the proposition assumes holds only approximately; Appendix E bounds the overlap. We gate the mode decision rather than crediting spans within the trace because credit for a span's causal effect on the answer is maximized by a span that simply states the answer (Appendix N).

4 Experiments

4.1 Setup

We train Qwen3-8B and Qwen3-4B (Yang et al., 2025) in the zero-SFT R1-Zero convention (DeepSeek-AI et al., 2025) on three English-source pairs (Chinese, German, Swahili), mixing a figurative pool (IdiomTranslate30 (Tang, 2022; Tang et al., 2024)) and a general pool (OpenSubtitles (Lison & Tiedemann, 2016), FLORES-200 (NLLB Team et al., 2022)) at $60\%/40\%$, $2{,}500$ steps, groups of $G{=}4$ forced $50/50$ with no-think candidates topped up to eight per source. We evaluate on a frozen 9-bucket in-distribution test set and an out-of-distribution axis (Tables 2 and 3) under XCOMET-XL (Guerreiro et al., 2024), COMET-22 (Rei et al., 2022), and chrF++ (Popović, 2017), the last corpus-level as sacrebleu (Post, 2018) computes it, macro-averaged over buckets, with paired bootstrap significance (Koehn, 2004) ($p < 0.05$) throughout; think rate and reasoning length are averaged over $K{=}8$ sampled completions rather than a single greedy decode, since think rate is discontinuous near the mode-indifference point while quality is not. Full data composition, evaluation protocol, and metric details are in Appendix B.

4.2 Compared Systems

Trained baselines share backbone, data, and optimization with TECA and differ in exactly one structural choice each. GRPO is mode-homogeneous and always-think. MT-R1-Zero keeps that setup with a rule-metric mixed reward (Feng et al., 2025) in place of ours. Mixed GRPO (ungated) forces $50/50$ mode-mixed groups with pooled, ungated credit ($w \equiv 1$), as in AdaptThink's cold-start scaffolding (Zhang et al., 2025). Qwen3-32B, evaluated in both modes, is the strongest external reference.

4.3 Does Selective Reasoning Win on Both Axes at Once? (RQ1)

We do not claim reasoning raises average quality; it does not. The claim is that gating reasoning-token credit wins on quality while spending far less of it. Weighting each system's think rate by its mean reasoning length gives the effective cost plotted in Figure 1: 281 tokens for TECA against 609 for GRPO and 705 for Base (think), **a 54–60% reduction against either**. Against Mixed GRPO (ungated), the comparison that isolates the credit mechanism, **TECA wins significantly on both neural metrics on every pair at 8B** (paired bootstrap, $p < 0.05$): XCOMET-XL by 1.1% to 10.7% and COMET-22 by 1.9% to 5.0% relative (Table 3). Its chrF++ gains are larger and less even, 6.0%

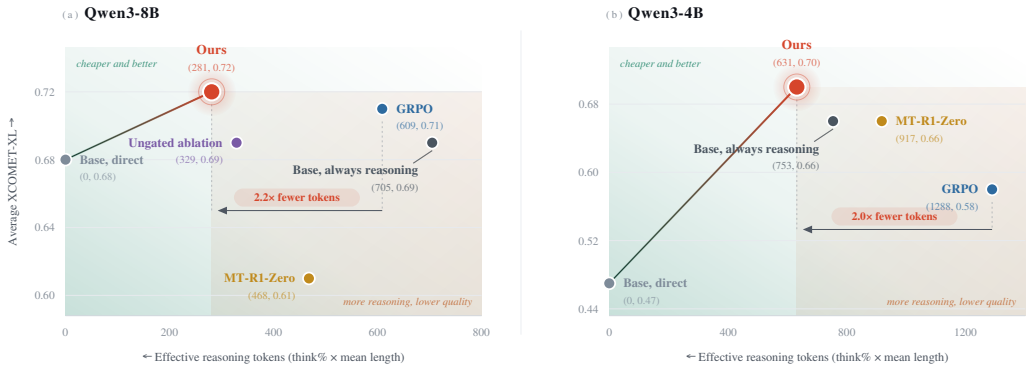


Figure 1: Quality against effective reasoning cost (think rate × mean reasoning tokens) at *(a)* 8B and *(b)* 4B. The shaded block is the region our policy dominates: more reasoning spent and lower average quality. Every other trained system falls inside it.

to 44.7%, and significant on en-de and en-sw ($p = .001$, $p < .001$) but not on en-zh ($p = .14$) (Table 3). Against GRPO, mode-homogeneous and always-think as Corollary 1 predicts, TECA wins on en-zh and en-de but loses on en-sw COMET-22 and XCOMET-XL specifically ($p = .0004$, $p = .001$), while still winning chrF++ there at under half the reasoning tokens. We report this plainly: on this pair, always reasoning still buys a small quality edge on two of three metrics.

Where the gain comes from. The left half of Table 2 breaks that comparison down by tier, and two things follow. First, **ungated mode-competition credit is actively harmful**. It drives quality *below* the untrained base model on two of three tiers: 0.720 on OpenSubtitles against the base brackets (the untrained model's two forced-mode rows) of 0.765 and 0.767, and 0.670 on FLORES against 0.674 and 0.684. Mixing the modes is not enough on its own: competing them on pooled reward, with every reasoning token credited in full, learns something worse than the model it started from. At 4B the same ungated run does not merely degrade but collapses: by step 400 its no-think probe produces no valid output, and the final checkpoint emits no valid translation on any input (Appendix O); it therefore has no row in Table 3.

Second, the gate's repair is concentrated where reasoning loses. **The gate's gain over the ungated ablation is largest on the tier where reasoning wins least often**: +8.5% XCOMET-XL on OpenSubtitles, whose single-sample win rate in Figure 2 is 0.325, against +3.5% and +4.8% on figurative text and FLORES (0.485, 0.493). Finding 3 traces that tier's low rate to short sources on which reasoning is redundant, so it is where ungated credit pays most for reasoning that buys nothing. This is what Proposition 1 predicts: ungated credit does its damage on inputs where reasoning tends to lose, so withholding it pays off most where those inputs concentrate. Where they concentrate was measurable before training began.

Against GRPO the picture is different and worth stating plainly. On XCOMET-XL the two are within 0.004 on every tier; our gains over it are on chrF++ (+6.1%, +15.7%, +4.6%), at 45% think rate against its 100%. In distribution the two are therefore close on the neural metric. They are not close off it: **on WMT24++ TECA beats GRPO on every pair on XCOMET-XL and chrF++** (Section 4.4), the setting where an always-reason policy has least to fall back on.

4.4 Is the Mechanism the One Designed, and Does It Transfer? (RQ2)

The trained policy reasons on 45% of inputs, against 100% for both Base (think) and GRPO, at quality within 0.004 XCOMET-XL of GRPO on every tier and above both base brackets throughout (Table 2, left). Three checks ask whether that reduction is selective, whether it holds outside the training distribution, and how it responds at a second scale.

The reduction is selective. Sorting the 1,650 test sources by the base model's own think-versus-no-think gap, the policy reasons on 49.5% of the top third against 43.3% of the bottom third

Table 3: Main results on the test set; Avg averages each metric over the three pairs. Think% is the fraction of inputs the policy reasons on; Len the mean reasoning tokens; best trained-row result per column in bold. The 4B Mixed GRPO (ungated) run collapsed and has no row (Appendix O).

	en→zh			en→de			en→sw			Avg				
Method	XC	C22	chrF++	XC	C22	chrF++	XC	C22	chrF++	XC	C22	chrF++	Think%	Len
Qwen3-32B (think)	0.88	0.83	20.9	0.95	0.82	48.6	0.39	0.59	32.0	0.74	0.74	33.8	99	643
Qwen3-32B (no-think)	0.90	0.84	21.4	0.94	0.83	50.1	0.40	0.60	32.2	0.75	0.76	34.6	0	0
Qwen3-8B														
Base (no-think)	0.89	0.83	19.9	0.90	0.79	47.5	0.27	0.47	8.1	0.68	0.69	25.2	0	0
Base (think)	0.88	0.82	17.6	0.93	0.80	46.7	0.26	0.44	21.9	0.69	0.68	28.7	100	705
GRPO (always think)	0.89	0.82	17.5	0.94	0.81	45.2	**0.32**	**0.57**	22.0	0.71	0.73	28.1	100	609
MT-R1-Zero	0.88	0.82	15.0	0.90	0.77	47.6	0.05	0.08	4.7	0.61	0.56	22.4	100	468
Mixed GRPO (ungated)	0.86	0.80	18.4	0.94	0.80	45.0	0.28	0.54	17.0	0.69	0.72	26.8	46	715
TECA	**0.91**	**0.84**	**19.5**	**0.95**	**0.82**	**48.1**	0.31	0.55	**24.6**	**0.72**	**0.74**	**30.7**	45	625
Qwen3-4B														
Base (no-think)	0.52	0.66	13.2	0.71	0.63	40.9	0.17	0.34	3.0	0.47	0.54	19.0	0	0
Base (think)	0.85	0.81	16.3	0.90	0.77	43.3	0.22	0.39	18.4	0.66	0.65	26.0	100	753
GRPO (always think)	0.88	0.82	**18.7**	0.79	0.68	39.5	0.07	0.11	5.1	0.58	0.54	21.1	100	1288
MT-R1-Zero	0.87	0.81	17.4	0.91	0.77	42.0	0.22	0.38	16.3	0.66	0.65	25.3	100	917
TECA	**0.90**	**0.83**	17.9	**0.94**	**0.80**	**44.3**	**0.26**	**0.46**	**16.9**	**0.70**	**0.70**	**26.4**	98	644

(COMET-22; $r = 0.125$, $p = 0.012$ under a permutation test shuffling think rates within each tier-language cell). The effect is concentrated in the upper tail, which is what the credit floor of Eq. (8) is designed to produce.

It survives a change of domain. The right half of Table 2 evaluates on WMT24++ (Deutsch et al., 2025), news, social, speech, and literary text with zero overlap with training. **TECA exceeds both base brackets and GRPO on every pair on XCOMET-XL and chrF++, with one exception**: en-zh chrF++, where it trails the stronger base bracket by 0.2. It does so while reasoning on 31–47% of inputs against GRPO's 100%, rather than reverting to always-think. Quality generalizes, and the mode decision does not collapse under a full distribution shift.

It adapts to the backbone. The same recipe, with identical hyperparameters, settles at 45% think at 8B and 92–100% across buckets at 4B. That is what the treatment-effect framing asks for, not a failure to generalize. The quantity the gate estimates is the per-input effect of reasoning, and **that effect is far larger on the smaller backbone**: the 4B base model's forced-think and forced-no-think brackets differ by 0.19 average XCOMET-XL against 0.010 at 8B, more than an order of magnitude. The learned think rate moves with it. A policy that reasoned on 45% of inputs at 4B would be leaving most of that effect unclaimed.

The gate is still doing work at that rate: the 4B policy beats both always-think baselines by 6–8% relative on XCOMET-XL and COMET-22 on **half the reasoning tokens of GRPO** (Figure 1b). Within 4B it leaves little room for selectivity to show, so the adaptive claim is across scales, not within this one, and two scales cannot separate a mechanism from a coincidence.

5 Conclusion

A group of rollouts that all reason contains no evidence about whether reasoning was worth it, and the same decomposition that shows this forces the remedy: mix the modes inside the group, score both modes with a reward that does not know which produced what, and couple the estimate to reasoning-token credit through a bounded, non-negative gate. Machine translation shows why that coupling is the part that matters. Without the gate, a policy in this regime does not merely fail to improve: it falls below the model it started from on two of three tiers, and at 4B collapses into a fixed point that emits no translation at all. We expect this structure wherever a reasoning treatment effect is not assumed non-negative: dialogue, summarization, code under soft quality criteria, and agentic settings where deliberation trades against latency. One lesson carries with it: reasoning there is better modelled as a variance-increasing intervention than as a mode that is simply better or worse, which changes what a selective policy should optimize.

Ethics Statement

This work studies when a translation model should reason before producing output, using public parallel-text corpora (IdiomTranslate30, OpenSubtitles, FLORES-200, WMT24++) under their respective licenses for research use. We do not collect new human data or involve human subjects. The method is trained and evaluated against automatic reference-based metrics (XCOMET-XL, COMET-22, chrF++); as with any RL objective built on a learned or automatic metric, the policy can in principle overfit to metric-specific artifacts rather than genuine translation quality, which is why we report a metric not used in training (XCOMET-XL) as our primary quality claim. We are not aware of a specific dual-use or harm pathway distinct from those already associated with general-purpose machine translation systems.

Limitations

Our experiments cover three English-source language pairs (Chinese, German, Swahili) and two model scales (Qwen3-4B, Qwen3-8B); we do not test non-English-source pairs or scales beyond 8B for the full method. The win-rate estimator requires accumulating paired comparisons per source over training, so the method's advantage depends on a training distribution with enough steps and repeated exposure to similar inputs for the posterior gate to concentrate; its behavior in low-data or single-pass fine-tuning regimes is untested. The reasoning-drafting audit of Corollary 2 was not logged during the reported runs (Audit 1, Appendix P), so we cannot rule out some self-dictation degeneracy in these particular checkpoints. The no-think probe tracking of Audit 2 finds no evidence of the related no-think sabotage failure mode at 8B. The blindness result (Corollary 1) and its remedy are stated for a treatment with two discrete modes (reason or not); extending the analysis to continuous or multi-way reasoning-budget decisions is left to future work.

Reproducibility Statement

The method is specified as a per-step procedure in Section 3 and Algorithm 1, with every component defined by an equation in that section. Data composition, the frozen test-set construction, decoding settings, metric versions and aggregation, and the full hyperparameter list are given in Appendix B; chrF++ is corpus-level as sacrebleu computes it, macro-averaged over buckets, and think rate and reasoning length are averaged over $K{=}8$ sampled completions. Proofs of Lemma 1 and Corollary 1 are given in Section 2, Proposition 1 in Section 3, and Proposition 2 and Corollary 2 in Appendix N, each stated with its assumptions. The failure modes we encountered, including the two whose fixes are load-bearing for reproducing the reported runs, are documented in Appendix P. Training and evaluation code will be released on acceptance.

AI Use Statement

We used generative AI tools for writing assistance, specifically for copy-editing prose and for suggesting revisions to the organization of the manuscript, and as a coding assistant for analysis and plotting scripts used to produce the tables and figures. All AI-assisted text was reviewed and edited by the authors, all reported numbers were generated from the experimental artifacts and checked against the analysis code, and any AI-assisted code was read and tested by the authors. We take responsibility for the final content of this work, including all text, claims, and artifacts.

References

Sayantan Adak, Pratyush Chatterjee, Somnath Banerjee, Rima Hazra, Somak Aditya, and Animesh Mukherjee. AURA: Affordance-understanding and risk-aware alignment technique for large language models. *arXiv preprint arXiv:2508.06124*, 2025. URL https://arxiv.org/abs/2508.06124.

Shabnam Ataee, Hugo Huart, and Andrei Popescu-Belis. Chain-of-thought reasoning improves context-aware translation with large language models. In *Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC)*, pp. 3725–3741, Palma de Mallorca, Spain, May

2026. ELRA Language Resource Association. doi: 10.63317/37kz9rawf9fn. URL https://aclanthology.org/2026.lrec-1.298/. arXiv:2510.18077.

DeepSeek-AI, Daya Guo, et al. DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning. *Nature*, 645(8081):633–638, 2025. doi: 10.1038/s41586-025-09422-z. URL https://www.nature.com/articles/s41586-025-09422-z. arXiv:2501.12948.

Daniel Deutsch, Eleftheria Briakou, Isaac Rayburn Caswell, Mara Finkelstein, Rebecca Galor, Juraj Juraska, Geza Kovacs, Alison Lui, Ricardo Rei, Jason Riesa, Shruti Rijhwani, Parker Riley, Elizabeth Salesky, Firas Trabelsi, Stephanie Winkler, Biao Zhang, and Markus Freitag. WMT24++: Expanding the language coverage of WMT24 to 55 languages & dialects. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 12257–12284, Vienna, Austria, July 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.findings-acl.634. URL https://aclanthology.org/2025.findings-acl.634/. arXiv:2502.12404.

Zhaopeng Feng, Shaosheng Cao, Jiahan Ren, Jiayuan Su, Ruizhe Chen, Yan Zhang, Jian Wu, and Zuozhu Liu. MT-R1-Zero: Advancing LLM-based machine translation via R1-Zero-like reinforcement learning. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 18685–18702, Suzhou, China, November 2025. Association for Computational Linguistics. URL https://aclanthology.org/2025.findings-emnlp.1015/. arXiv:2504.10160.

Nuno M. Guerreiro, Ricardo Rei, Daan van Stigt, Luisa Coheur, Pierre Colombo, and André F. T. Martins. xCOMET: Transparent machine translation evaluation through fine-grained error detection. *Transactions of the Association for Computational Linguistics*, 12:979–995, 2024. URL https://aclanthology.org/2024.tacl-1.54/.

Shijue Huang, Hongru Wang, Wanjun Zhong, Zhaochen Su, Jiazhan Feng, Bowen Cao, and Yi R. Fung. AdaCtrl: Towards adaptive and controllable reasoning via difficulty-aware budgeting. *Transactions on Machine Learning Research*, 2026. URL https://openreview.net/forum?id=4J2Ako20V4. arXiv:2505.18822.

Philipp Koehn. Statistical significance tests for machine translation evaluation. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pp. 388–395, Barcelona, Spain, 2004. Association for Computational Linguistics. URL https://aclanthology.org/W04-3250/.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with PagedAttention. In *Proceedings of the 29th Symposium on Operating Systems Principles (SOSP)*, 2023.

Zihao Li, Shaoxiong Ji, and Jörg Tiedemann. Test-time scaling of reasoning models for machine translation. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2902–2917, Rabat, Morocco, March 2026. Association for Computational Linguistics. doi: 10.18653/v1/2026.eacl-long.133. URL https://aclanthology.org/2026.eacl-long.133/.

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In *International Conference on Learning Representations (ICLR)*, 2024. URL https://openreview.net/forum?id=v8L0pN6EOi. arXiv:2305.20050.

Pierre Lison and Jörg Tiedemann. OpenSubtitles2016: Extracting large parallel corpora from movie and TV subtitles. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC)*, pp. 923–929, Portorož, Slovenia, May 2016. European Language Resources Association (ELRA). URL https://aclanthology.org/L16-1147/.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations (ICLR)*, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

H. B. Mann and D. R. Whitney. On a test of whether one of two random variables is stochastically larger than the other. *The Annals of Mathematical Statistics*, 18(1):50–60, 1947. ISSN 0003-4851. doi: 10.1214/aoms/1177730491.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, et al. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*, 2022. URL https://arxiv.org/abs/2207.04672.

Maja Popović. chrF++: words helping character n-grams. In *Proceedings of the Second Conference on Machine Translation (WMT)*, pp. 612–618, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. URL https://aclanthology.org/W17-4770/.

Matt Post. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pp. 186–191, Brussels, Belgium, 2018. Association for Computational Linguistics. URL https://aclanthology.org/W18-6319/.

Sara Rajaee, Sebastian Vincent, Alexandre Berard, Marzieh Fadaee, Kelly Marchisio, and Tom Kocmi. Unlocking reasoning capability on machine translation in large language models. *arXiv preprint arXiv:2602.14763*, 2026. URL https://arxiv.org/abs/2602.14763.

Jeff Rasley, Samyam Rajbhandari, Olatunji Ruwase, and Yuxiong He. DeepSpeed: System optimizations enable training deep learning models with over 100 billion parameters. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 3505–3506, 2020.

Ricardo Rei, José G. C. de Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André F. T. Martins. COMET-22: Unbabel-IST 2022 submission for the metrics shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pp. 578–585, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. URL https://aclanthology.org/2022.wmt-1.52/.

Ricardo Rei, Nuno M. Guerreiro, José Pombal, Daan van Stigt, Marcos Treviso, Luisa Coheur, José G. C. de Souza, and André Martins. Scaling up CometKiwi: Unbabel-IST 2023 submission for the quality estimation shared task. In *Proceedings of the Eighth Conference on Machine Translation (WMT)*, pp. 841–848, Singapore, 2023. Association for Computational Linguistics. URL https://aclanthology.org/2023.wmt-1.73/.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017. URL https://arxiv.org/abs/1707.06347.

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

Kenan Tang. PETCI: A parallel English translation dataset of Chinese idioms. *arXiv preprint arXiv:2202.09509*, 2022. URL https://arxiv.org/abs/2202.09509.

Kenan Tang, Peiyang Song, Yao Qin, and Xifeng Yan. Creative and context-aware translation of East Asian idioms with GPT-4. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 9285–9305, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.544. URL https://aclanthology.org/2024.findings-emnlp.544/.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel R. Bowman. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. *Advances in Neural Information Processing Systems (NeurIPS)*, 36, 2023. URL https://arxiv.org/abs/2305.04388.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems (NeurIPS)*, 35, 2022. URL https://arxiv.org/abs/2201.11903.

Violet Xiang, Chase Blagden, Rafael Rafailov, Nathan Lile, Sang Truong, Chelsea Finn, and Nick Haber. Just enough thinking: Efficient reasoning with adaptive length penalties reinforcement learning. *arXiv preprint arXiv:2506.05256*, 2025. URL https://arxiv.org/abs/2506.05256.

An Yang et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025. URL https://arxiv.org/abs/2505.09388.

Yongshi Ye, Biao Fu, Chongxuan Huang, Yidong Chen, and Xiaodong Shi. Translation with thought: Difficulty-adaptive reasoning via reinforcement learning for multi-domain machine translation. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 30339–30370, San Diego, California, USA, July 2026. Association for Computational Linguistics. doi: 10.18653/v1/2026.acl-long.1400. URL https://aclanthology.org/2026.acl-long.1400/.

Armel Zebaze, Rachel Bawden, and Benoît Sagot. LLM reasoning for machine translation: Synthetic data generation over thinking tokens. *arXiv preprint arXiv:2510.11919*, 2025. URL https://arxiv.org/abs/2510.11919.

Jiajie Zhang, Nianyi Lin, Lei Hou, Ling Feng, and Juanzi Li. AdaptThink: Reasoning models can learn when to think. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 3716–3730, Suzhou, China, November 2025. Association for Computational Linguistics. URL https://aclanthology.org/2025.emnlp-main.184/. arXiv:2505.13417.

A Related Work

Adaptive and hybrid reasoning. Methods that let a model choose how much to reason train the choice with RL in verified domains. AdaptThink (Zhang et al., 2025) forces both modes into the group and biases the policy toward not thinking subject to an accuracy constraint. AdaCtrl (Huang et al., 2026) and Just Enough Thinking (Xiang et al., 2025) allocate a reasoning budget from estimated problem difficulty, the former also exposing the budget to the user through length tags. In these domains reasoning's expected effect on correctness is non-negative, so the learned object is a budget, how much to reason, and the question of whether reasoning hurts a given input does not arise. Our ablation reproduces the mixed-group scaffolding these methods share and shows that, once the effect can be negative, mixing alone drives quality below the untrained model.

Adaptive reasoning for translation. Closest to our setting, Translation with Thought (Ye et al., 2026) modulates between intuitive and deliberate translation by estimated difficulty: it fine-tunes on difficulty-aware reasoning traces distilled from DeepSeek-R1, then optimizes a reward that combines translation quality with reasoning efficiency, cutting reasoning tokens by 32–60% across fifteen benchmarks. The two approaches differ in what is learned and from what. TwT learns how much to reason on a source by trading quality against length, starting from distilled traces; TECA learns whether reasoning wins on a source from the model's own paired rollouts, with no distilled traces or supervised stage, and shows that the standard group-relative advantage cannot supply that information on its own (Corollary 1). The two are complementary: a budget can be learned on top of a decision to reason at all.

Does reasoning help translation? The evidence points both ways. Chain-of-thought helps on context-dependent text (Ataee et al., 2026) and hurts when a model is pushed past its natural stopping point (Li et al., 2026). Rajaee et al. (2026) find that enabling explicit reasoning in reasoning-oriented models consistently lowers translation quality, attribute it to traces that are linear and never revise, and repair it by training on structured traces that draft, check adequacy and fluency, and revise. Zebaze et al. (2025) find that thinking tokens alone do not help and that gains appear only when the

intermediate tokens contain translation attempts. Section 2.2 shows both outcomes on one model as a change in the spread of quality rather than its level. These works change what the trace contains; we change whether a trace is produced, and the two are orthogonal.

Reinforcement learning for translation. MT-R1-Zero (Feng et al., 2025) optimizes translation quality under a group-relative reward in mode-homogeneous groups, without asking whether the reasoning that preceded a good translation caused it; we use it as a baseline. Learned metrics used as rewards can be optimized for their own artifacts, which is why every quality claim in this paper rests on a metric not used in training.

Credit assignment for reasoning tokens. Group-relative methods (Shao et al., 2024; DeepSeek-AI et al., 2025) assign one outcome-based advantage to every token of a rollout, and Lemma 1 states what that advantage cannot contain. Process reward models score individual steps (Lightman et al., 2024; Adak et al., 2025), and traces can misrepresent the computation behind an answer (Turpin et al., 2023). Crediting a span by its causal effect on the answer is maximized by a span that states the answer (Appendix N), which is why TECA credits the mode decision rather than the trace.

B Full Setup Detail

Models. Two base backbones, zero-SFT R1-Zero convention (no system prompt, a free-form reasoning segment closed by `</think>`, the translation as the plain text after it): Qwen3-8B, on which all baselines and ablations run, and Qwen3-4B as a scale check, on which we train TECA, GRPO (always think), MT-R1-Zero, and Mixed GRPO (ungated), the last of which collapsed (Appendix O).

Training data. Training mixes two pools at a fixed $60\%/40\%$ ratio, figurative to general, held constant for the full run with no curriculum schedule. At 24 groups sampled per step this averages 14.4 figurative and 9.6 general groups per step, and each pool cycles through its full set exactly twice over the 2,500-step run. Within each pool, the three language pairs are mixed in exact 1:1:1 proportion. The figurative pool is IdiomTranslate30 (Tang, 2022; Tang et al., 2024) only, 6,000 pairs per language (18,000 total). The general pool combines OpenSubtitles (3,300 pairs per language, conversational dialogue) and FLORES-200 (700 pairs per language, formal single-register text), 4,000 pairs per language, 12,000 total. FLORES-200 training and test sources are disjoint subsets of its dev split; devtest is not used. Groups of $G{=}4$ rollouts at forced $50/50$ mode mixing; no-think candidates are topped up to eight per source with auxiliary rollouts that are scored but never backpropagated through.

Evaluation data. *In-distribution*: a frozen test set of $1{,}650$ sources in 9 buckets (figurative, conversational, and FLORES suites $\times$ 3 pairs; 200, 150 and 200 sources per pair), none of which appears in training, evaluated exactly once; a disjoint frozen validation set serves checkpoint selection and diagnostics. *Out-of-distribution domain*: WMT24++ (Deutsch et al., 2025) (news, social, speech, and literary text), all three pairs, with zero overlap with the training data.

Metrics. Three metrics per language pair: XCOMET-XL (XC; `Unbabel/XCOMET-XL`), reference-based, used nowhere in training, which carries all quality claims; COMET-22 (C22; `Unbabel/wmt22-comet-da`), reference-based, the training reward metric, reported disclosed as such; and chrF++, independent of the COMET model family and the training reward, computed with sacrebleu (Post, 2018). A reference-free metric's gap to the reference-based metrics on think-versus-no-think comparisons is a diagnostic, not one of the three reported quantities. Empty or invalid hypotheses are floor-scored rather than passed to the metric: XCOMET-XL scores an empty string well above many genuine low-quality translations, so scoring it at face value would understate whichever system fails more often. Significance: paired bootstrap resampling (Koehn, 2004), $p < 0.05$, on all system-versus-ours comparisons.

Selectivity reporting. Think rate and mean reasoning tokens are averaged over $K{=}8$ sampled completions at temperature 0.7 rather than read from a single greedy decode. Think rate is discontinuous in the argmax path near the mode-indifference point $\pi \approx \frac{1}{2}$, while quality is not, so greedy decoding can show think rate swinging sharply between near-zero and near-total across adjacent

checkpoints with no corresponding change in quality, an artifact we verified directly during training and avoid by reporting sampled means throughout.

Optimization and compute. We optimize with AdamW (Loshchilov & Hutter, 2019) at learning rate 5×10^{-7} under DeepSpeed (Rasley et al., 2020), with a light KL penalty of coefficient 0.001 to the frozen base policy on all generated tokens except y_1, for $2{,}500$ steps. The PPO clip is $\epsilon = 0.2$ and the advantage's standard-deviation floor is $\varepsilon = 10^{-8}$. Both modes are sampled at temperature 0.9 (top-p 0.95) with a $1{,}280$-token generation cap; evaluation decodes with a $2{,}048$-token cap. The posterior gate uses concentration $\kappa = 3.15$, so that a source's own comparisons outweigh the population prior once n_x exceeds about three, which at our group size is roughly two visits. It further uses a per-source count decay $\gamma = 0.99$ applied per visit, a prior refit every 100 steps, and the credit floor $w_{\min} = 0.25$ of Eq. (8); the y_1 balance term (Eq. (9)) is weighted at 0.01. The training reward is reference-based COMET-22 (`wmt22-comet-da`); the reference-free diagnostic is COMETKiwi (Rei et al., 2023) (`wmt23-cometkiwi-da-xl`), which also scores any reference-less item. Paired-bootstrap significance uses $10{,}000$ resamples. Training runs on four H100 80GB GPUs (three policy shards plus one vLLM (Kwon et al., 2023) rollout server).

C Training Procedure in Pseudocode

Algorithm 1 TECA (one step)

Persistent state: per-source (W_x, n_x); population prior $\mathrm{Beta}(a, b)$.
Sample a batch of sources x.
for each x **do**
  Assign each of G slots think or no-think by a fair coin; flip one slot if the draw is single-mode.
  No-think slots: splice `</think>` as y_1. Think slots: sample y_1 from the policy restricted to other tokens. Generate.
  Top up no-think candidates to eight per source (scored only, no gradient).
  Score every candidate with COMET-22 against the reference; drop rollouts with no translation.
  $A_i \leftarrow$ pooled standardized advantage over the valid group rollouts (Eq. (4)).
  $u_x \leftarrow$ graded win fraction over all think $\times$ no-think cross-pairs, ties as $\frac{1}{2}$ (Eq. (5)).
  $(W_x, n_x) \leftarrow \gamma\,(W_x, n_x) + n_{\text{eff}}\,(u_x,\ 1), \quad n_{\text{eff}} = \min(|\mathcal{T}_x|,\ |\mathcal{N}_x|)$.
  $w_x \leftarrow 1 - I_{1/2}(a + W_x,\ b + n_x - W_x); \quad \tilde{w}_x \leftarrow w_{\min} + (1 - w_{\min})\,w_x; \quad \hat{p}_x \leftarrow (a + W_x)/(a + b + n_x)$.
end for
Loss (Eq. (9)). Answer tokens (both modes): A_i. Reasoning tokens (think only): $\tilde{w}_x A_i$.
  Mode token y_1: clipped surrogate with A_i on the realized mode event (proposal $\frac{1}{2}$) + balance term toward $\hat{p}_x$.
  KL to the frozen base policy on all generated tokens except y_1.
Periodically refit $\mathrm{Beta}(a, b)$ by moment-matching the accumulated per-source rates.

D Advantage Rescaling and Per-Step Bookkeeping

Two remarks on Eq. (4). Lemma 1 is stated for the centred advantage, whereas Eq. (4) also divides by $\hat{\sigma}_Q$, as is standard in group-relative RL. The division rescales both terms of the decomposition rather than removing either, so the residual carrying $\Delta Q(x)$ survives and the $M = 0$ conclusion is unchanged. It does introduce a second-order effect worth naming: by the law of total variance, pooling two modes whose means differ makes $\hat{\sigma}_Q$ larger than the within-mode variation alone, so mixing supplies the contrast and mildly damps every advantage in the group at the same time.

Bookkeeping outside the step. The $\mathrm{Beta}(a, b)$ population prior of Eq. (7) is refit periodically by moment-matching its mean to the accumulated per-source rates, while its concentration $\kappa = a + b$ is held fixed. The refit mean is what makes the $n_x = 0$ fallback a measured base rate rather than a chosen default; the fixed κ is what sets how much evidence a source needs before its own

comparisons outweigh that base rate, since prior and data enter the posterior in the ratio $\kappa : n_x$. Separately, free-form generation occasionally violates the expected format: a stray fragment before the answer, or a reasoning segment that fails to close. A completion that still contains a translation stays in the loss, scored on its quality less the format penalty, and the sequence that receives gradient must match the sequence generated, anomaly included; a completion with no translation at all is dropped from the group (Step 2). Crediting the penalty to a cleaned-up reconstruction the policy never generated removes the only gradient that could correct the anomaly, turning a rare formatting failure into an absorbing one (Appendix P, Audit 3).

E THE GATE AT THIS TRAINING HORIZON

At the two-epoch schedule of Section 4.1 each source is drawn about twice, so n_x reaches roughly 4 and the posterior sits at approximately even weight between the fitted prior and the source's own evidence. The gate is correspondingly conservative at this horizon, and γ is close to inert. Both the accumulation and the discount are built for longer schedules, where n_x grows, the prior's share falls, and the gate sharpens toward the indicator that Proposition 1 describes in the limit.

The posterior is updated with a step's comparisons before it is read (Algorithm 1), so the gate applied to a step's reasoning tokens is not fully independent of the rollouts receiving the credit, as Proposition 1 assumes. The dependence is bounded. Each visit adds $n_{\text{eff}} = \min(|\mathcal{T}_x|, |\mathcal{N}_x|) \approx 2$ comparisons against a prior of weight $\kappa = 3.15$, so the current step supplies about 39% of the posterior's weight on a source's first visit and 28% on its second, a share that keeps falling as n_x grows and vanishes in the limit the proposition considers.

F WHAT THE POSTERIOR FORM SETTLES

The posterior gate of Eq. (7) settles three choices a point estimate would leave to a hyperparameter: it is bounded in $[0, 1]$ and cannot flip an update's sign; at $n_x{=}0$ it falls back to the measured population base rate rather than an assumed default; and its sharpness is the posterior concentration, which grows with n_x. The one free choice is the prior concentration κ, which sets how fast a source's own evidence takes over from the population rate (Appendix D).

G BASE-MODEL EFFECT SIZES BY TIER

Table 4: Base-model reasoning versus direct translation, pooled across three language pairs (COMET-22 (Rei et al., 2022), $K = 12$ paired samples per sentence). Δ_{cond} conditions on both modes emitting a well-formed output; Δ_{uncond} floor-scores the failures in the rightmost column instead, nearly all of them reasoning segments running past the token budget before closing (Finding 1). Win-rate and best-of-K appear as a curve in Figure 2.

Tier	Δ_{cond}	Δ_{uncond}	ties	invalid
Figurative	$+0.020$	-0.070	0.2%	17.7%
OpenSubtitles	$+0.000$	-0.050	16.5%	8.1%
FLORES	$+0.021$	-0.086	1.7%	18.2%

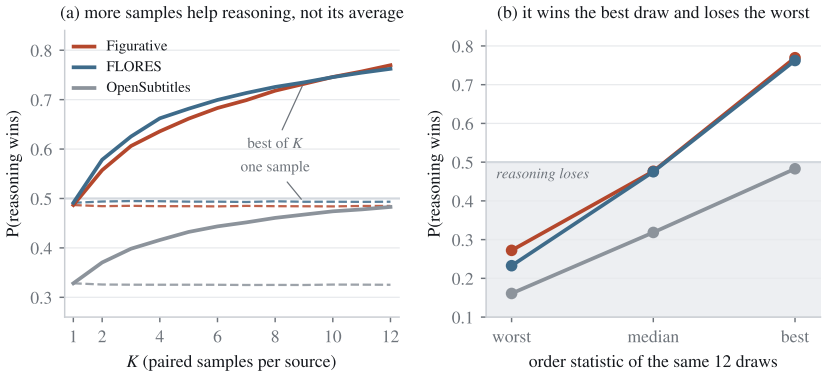


Figure 2: Reasoning widens translation quality without raising its mean (Qwen3-8B, COMET-22, K=12 samples per mode per source, three language pairs pooled; ties count as non-wins). *(a)* Solid: probability that the best of K reasoning samples outscores the best of K direct samples. Dashed: probability that one reasoning sample outscores one direct sample, which does not depend on K and is drawn for reference. *(b)* At K=12, each source's draws are sorted within each mode and compared at matched rank: reasoning wins the best draw and loses the worst on two of three tiers.

H Source Length Explains the OpenSubtitles Tier

Finding 3 reports that OpenSubtitles is the weakest tier for reasoning in every language pair, and that the ordering is not one of difficulty. The property that does predict it is source length.

The three tiers are drawn from different genres and barely overlap in length. OpenSubtitles is subtitle dialogue: its median source is 5 words, 52% of its sources are five words or shorter and 31% are three words or shorter. Figurative text and FLORES have median lengths of 20 and 19 words, and *neither contains a single source of five words or fewer*.

Within OpenSubtitles, reasoning's win rate rises with length and its tie rate falls, monotonically across every bin:

source length	n	win rate	tie rate
1–3 words	141	0.246	0.336
4–5 words	93	0.423	0.139
6–10 words	150	0.479	0.084
11+ words	66	0.506	0.022

Across sources, length correlates $+0.28$ with the win rate and -0.37 with the tie rate. The shortest sources are proper nouns and interjections, and on these the two modes agree almost perfectly: `"Telemachus."`, `"Okay."` and `"-Yes."` each have tie rates of 0.92 or above. A one-word source admits essentially one correct translation, so deliberating cannot improve on producing it directly, and under the convention of Figure 2 an identical output is scored as a non-win.

Comparing tiers at matched length narrows the gap. Restricting all three tiers to sources of at least ten words, OpenSubtitles' win rate is 0.487 (95% bootstrap CI $[0.427, 0.549]$, $n = 83$) against 0.556 for figurative text and 0.535 for FLORES; at fifteen words and above it is 0.570 ($[0.449, 0.689]$, $n = 24$), nominally the highest of the three. The intervals overlap at both thresholds. We report this with the caveat that restricting OpenSubtitles to long sources leaves few of them, so the matched-length comparison is underpowered at the upper thresholds; the within-tier gradient above, which uses all 450 sources, is the stronger evidence.

We therefore read the tier effect as a length effect rather than a genre effect. Reasoning does not fail on conversational text because the text is conversational. It fails on sources short enough that there is nothing to deliberate about, and subtitle corpora contain a great many of them.

I The Failure Rate of Finding 1 Is Measured at the Training Cap

The samples behind Finding 1 were generated with the same $1{,}280$-token generation cap that training uses (Appendix B), so the rate at which reasoning runs past the cap without closing is measured under the conditions the policy is trained under, not under a tighter budget imposed for the study. Of the $2{,}379$ reasoning samples that produced no translation, $2{,}378$ reached this cap without closing the reasoning segment.

J Why the Gate Reads a Win-Rate and Not a Mean Effect

We gate on $\delta(x)$ rather than on $\Delta(x)$, and the two are not interchangeable here. They would agree in sign if the two quality distributions were stochastically ordered, but Figure 2 shows they are not: reasoning's distribution has the heavier upper tail and the lower median, so the two cross. Which statistic is correct therefore depends on how the policy is deployed. $\Delta(x)$ is the right target for a policy whose output is averaged over many samples, since it weights the tail that averaging preserves. $\delta(x)$ is the right target for a policy that emits one sample and cannot select among draws, which is ours: what matters then is how often a single reasoning sample beats a single direct one, not by how much the best of them would win. A second reason is practical. Because Q is a learned metric whose scale drifts as the policy improves, the magnitude of $\Delta(x)$ is not comparable across training iterations, whereas $\delta(x)$ is invariant to any monotone transformation of Q. Figure 2a is the first measurement of this quantity, and $\delta(x)$ is the signal the method models throughout training.

K The Estimator's Reward Must Be Mode-Invariant

Mode mixing returns $\Delta(x)$ to the gradient only as the sample-mean $\Delta Q(x)$ of Eq. (3), and that stand-in is faithful only if Q scores every candidate the same way regardless of which mode produced it; otherwise the measured $\Delta Q(x)$ mixes the true effect with an artifact of the metric.

Definition 1. *A quality functional Q is* mode-invariant *if it is computable from the unordered set of candidate answers alone, independent of which generation mode produced each candidate.*

Verified correctness rewards, as used in math and code reinforcement learning with verifiable rewards, are mode-invariant trivially: correctness of an answer does not depend on how it was produced. Absolute, learned translation quality metrics need not be: we observe a policy for which mode-forced no-think translations score *higher* on COMETKiwi, a standard reference-free neural metric, than mode-forced think translations on templatic, low-ambiguity text. Whatever its cause, any $\Delta(x)$ estimate built on a non-mode-invariant Q is confounded by the reward's own mode-dependence rather than purely by the treatment effect of interest. We therefore anchor the reward to a human reference: training scores every candidate with a reference-based metric, tying the score to an external target rather than to the learned output distribution where the mode-dependent gap appears. The graded win indicator between paired candidates, one for a win and one half for a tie, then yields the win-rate $p(x)$ of Eq. (2), so the reward and the estimand the method tracks are built from the same scores.

L The Mode Gradient Under Forced Mixing

With the first token drawn from a known proposal, the mode update treats y_1 as the binary event it is: for a rollout forced into mode m, the importance ratio is $\pi_\theta(m \mid x)/\frac{1}{2}$ on the realized mode event, clipped as in PPO (Schulman et al., 2017). With both modes present in every group, the coin that assigned each rollout to a mode can be integrated out. Let $\bar{A}_1$ and $\bar{A}_0$ be the mean advantages of the think and no-think rollouts in the group. Taking the conditional expectation over the coin, the expected gradient of the surrogate objective J_{y_1} with respect to the logit ℓ of $\pi = \pi(x)$ is

$$\mathbb{E}\big[\nabla_\ell J_{y_1}\big] = \big(\bar{A}_1 - \bar{A}_0\big)\,\pi(1-\pi), \tag{12}$$

in which the importance ratio and the $\frac{1}{2}$ proposal weight cancel exactly, so ascent raises π when the think rollouts carry the higher advantage and lowers it when they do not.

Two remarks. First, with the standardized quality advantage of Eq. (4), which is what TECA uses, $\bar{A}_1 - \bar{A}_0 = \Delta Q(x)/\hat{\sigma}_Q$: the surrogate moves the mode probability by the group's mean-quality gap, the statistic Appendix J argues is the wrong target for a single-sample policy. If each candidate were instead scored by the fraction of opposite-mode candidates it beats, the difference would equal $2\hat{\delta}(x)$, the group's win-rate margin, and Eq. (12) would read $2\pi(1-\pi)\hat{\delta}(x)$. We do not use that scoring: in TECA the win-rate posterior reaches the mode decision through the balance term of Eq. (9), whose target is the posterior mean $\hat{p}_x$. Second, Eq. (12) is a Rao–Blackwellized, lower-variance form of the sampled estimator; training on it directly destabilized training by removing the PPO clip's trust region along with the importance ratio (Appendix M), so the method uses the sampled, clipped estimator throughout. Either way the expected update carries the factor $\pi(1-\pi)$, which vanishes at the corners the policy starts near, which is why the balance term, whose logit-gradient is $\pi - \hat{p}_x$, is paired with it (Section 3).

M The Marginalized Mode-Gradient Estimator

Appendix L derives the expected mode gradient in closed form (Eq. (12)) by integrating out the coin that assigns each rollout to the think or no-think mode. This is a Rao–Blackwellization of the sampled, importance-corrected estimator the method actually uses.

By the Rao–Blackwell theorem, conditioning out the mode-assignment coin leaves the marginalized estimator of Eq. (12) unbiased for the expected y_1 policy gradient with variance no larger than the sampled, importance-corrected estimator, and strictly smaller whenever $\pi \in (0, 1)$, since the coin's importance ratio contributes variance and, by the cancellation in Eq. (12), no bias.

The variance reduction argues for training on this marginalized form directly rather than on the sampled, clipped estimator the method actually uses. We tried this and it destabilized training. Removing the importance ratio also removes the PPO clip that acts on it, and the clip was doing more than de-noising the mode-assignment coin: at this decision point it was the only trust region on the mode update at all. Without it, the low-variance marginalized gradient pushed the mode probability π from one corner toward the other in slow relaxation cycles rather than settling, and the reasoning mode degenerated over the course of this drift as think-mode rollouts became increasingly rare on one side of each cycle. The sampled, clipped estimator run under otherwise identical conditions showed no such instability.

The lesson is not that variance reduction is undesirable here, but that variance was not the binding constraint on this particular gradient; the trust region was, and the two are coupled in the sampled estimator in a way that the marginalized form breaks. A marginalized estimator with its own, separately-defined trust region, clipping each mode's policy ratio against the previous checkpoint rather than relying on the now-absent importance ratio to supply one, is a natural next step we did not pursue in this work.

N Why the Gate Targets the Decision, Not the Trace

TECA (Section 3) gates reasoning credit on the per-input win-rate margin $\delta(x)$, a property of the mode *decision*. A natural-seeming alternative is to credit reasoning by *what the trace contributes*, rewarding a trace in proportion to its effect on the output. This section shows why that alternative is unsound: any reward that credits reasoning by its causal effect on the answer is maximized by a policy that stops deliberating and starts *drafting the answer inside the trace*. Translation then hands us a free detector for exactly this degeneracy, should it arise under any objective.[1]

Proposition 2 (Self-Dictation Dominates Necessity Credit)**.** *Let $\tau = s_1 \cdots s_n$ and define, for span s_k, the necessity reward $R_{\mathrm{nec}}(s_k) = Q(x, y) - \mathbb{E}_{y' \sim \pi_\theta(\cdot|x, \tau_{-k})}[Q(x, y')]$, the drop in answer quality when s_k is ablated. Among spans that induce the same output distribution under presence, a* self-dictating *span s_k^{sd} that literally states the intended answer attains $R_{\mathrm{nec}}(s_k^{\mathrm{sd}}) \geq R_{\mathrm{nec}}(s_k^{\mathrm{gen}})$ for every genuinely deliberative span s_k^{gen}, with weakly lower variance.*

[1] A shallow variant is immediate and we set it aside: a claim-consistency reward with the convention that a trace making no checkable claim scores maximally is maximized by a policy whose trace makes no claims at all. It is defeated by any extractor that does not reward non-committal text; the consequential version below cannot be.

Proof. With the output distribution under presence fixed, both spans share the $Q(x, y)$ term, so R_{nec} differs only in the ablation fallback $\mathbb{E}[Q(x, y')]$; a larger reward means a worse fallback. A deliberative span states a fact the model must re-derive under ablation (fallback q_{gen}); a self-dictating span carries the literal target, so under ablation the model falls back to the same re-derivation, giving $q_{\text{sd}} \leq q_{\text{gen}}$ (it can ignore the copy and re-derive, and does strictly worse whenever the copy would have succeeded where inference fails). Hence $R_{\text{nec}}(s_k^{\text{sd}}) \geq R_{\text{nec}}(s_k^{\text{gen}})$; the variance claim follows since copying is lower-entropy than inference. □

The proposition is not about a flawed implementation of necessity credit; it is about the credit *target*. Necessity credit rewards the span whose removal most degrades the output, and by construction the content that most degrades the output when removed is the content that *is* the output. No refinement of how the causal effect is measured escapes this; only abandoning causal-effect-on-output as the target does. This is why TECA gates on the per-input win-rate margin $\delta(x)$, a property of the mode decision, rather than on any within-trace attribution, a property of the trace content.

Corollary 2 (Cross-Lingual Detectability of Self-Dictation)**.** *Let x be in source language L_{src} and the answer in $L_{\text{tgt}} \neq L_{\text{src}}$ with disjoint scripts. Define the target-script fraction $\sigma(\tau) = |\{c \in \tau : c \in \text{Unicode}(L_{\text{tgt}})\}|/|\tau|$. A self-dictating span of Proposition 2 contains the answer, hence a substring in L_{tgt}, so $\mathbb{E}[\sigma(\tau^{\text{sd}})]$ is bounded away from zero, whereas genuine deliberation conducted in L_{src} has $\sigma(\tau^{\text{gen}}) \approx 0$. Thus $\sigma(\tau)$ separates the two, is model-free and zero-cost, and is* structurally unavailable *in monolingual reasoning domains, where deliberation and answer share a script and no such marker exists.*

We recommend $\sigma(\tau)$ as a standing monitor; the audit we could run on these checkpoints is reported as Audit 1 in Appendix P.

O The Ungated Collapse in Full

The clearest evidence that reasoning-token credit needs gating comes from removing it and watching what the objective does on its own.

At 4B, the ungated ablation collapses. The no-think probe holds 32 of 50 valid outputs at step 100 and reaches 0 of 50 by step 400, and never exceeds 1 of 50 over the remaining 2,100 steps. This is the failure Audit 2 of Appendix P is built to detect: a policy can raise its measured $\delta(x)$ by degrading the mode it is measured against, and with every reasoning token credited in full there is nothing opposing that gradient. The end state is degenerate. Re-serving the final checkpoint and probing it directly, it emits an unterminated sequence of newlines inside the reasoning block on every input, producing zero valid translations at temperature 0.0, 0.7, and 1.0 alike. That is a fixed point of the policy, not a decoding artifact, and it is why we report a collapse here instead of a row of scores.

At 8B the same configuration survives but degrades late: **format validity falls from** 86% **to** 43% **and target-script leakage rises from** 2% **to** 56% **over the run.** TECA under otherwise identical settings holds leakage in $[0, 7\%]$ and format validity in $[81, 100\%]$ across all 2500 steps.

The severity therefore tracks backbone capacity in the direction Proposition 1 implies. Ungated credit carries a wrong-way gradient wherever reasoning does not help, the 4B base model depends on reasoning far more than the 8B one, and the smaller backbone has correspondingly more to lose from crediting reasoning indiscriminately. Two scales are not enough to establish this as a law, but the direction is consistent and the mechanism is one the theory names in advance.

GRPO at 4B degrades differently and for an unrelated reason: it remains mode-homogeneous throughout and fails by producing no parseable translation on 81% of Swahili test items, against 5% for TECA. Its 4B averages in Table 3 should be read with that in mind.

P Failure-Mode Audits

Audit 1: the reasoning-drafting audit. Corollary 2 gives a cheap, zero-cost detector for the self-dictation degeneracy on any pair with disjoint source and target scripts, and English→Chinese in

our data qualifies. The detector was not run against these training runs: we did not retain per-step reasoning-trace text, so the trajectory cannot be reported retroactively. This is a logging gap, not a null result. We did not observe symptoms consistent with self-dictation in spot checks of final-checkpoint outputs (no target-script leakage in sampled reasoning segments), but this is not the systematic, per-step audit the detector is designed to run, and we do not present it as one. We recommend $\sigma(\tau)$ be logged as a standing monitor in any reproduction of this method.

Audit 2: the no-think mode is not sabotaged. A win-rate reward can be gamed from either side: a policy can raise $\delta(x)$ by degrading its own no-think mode rather than improving reasoning. We track no-think quality on a fixed probe set throughout training. At 8B, TECA shows no sabotage: probe quality is stable across all checkpoints sampled, with no degrading trend. At 4B, Mixed GRPO (ungated) shows exactly this failure: probe validity collapses from 32 of 50 outputs valid to 0 of 50 by step 400, concurrent with the broader collapse reported in Appendix O. This degeneracy is specific to relative, mode-competition credit and distinct from self-dictation; Audit 2 is designed to catch it, and it did.

Audit 3: credit leakage through cleaned reconstruction. Two collapses in early training runs traced to the same cause, reported together because the fix generalizes beyond translation. In one, a stray fragment appears between the closed reasoning segment and the answer; in the other, the reasoning segment fails to close and the completion degenerates into repetition. In both, a parser that strips the anomaly before scoring produces a clean string for the reward while the format penalty is credited to a training sequence *reconstructed* to look clean, a sequence the policy never generated. The penalty's gradient then trains a boundary absent from the sampled trajectory, and, because the degenerate tokens themselves never receive negative credit, the state is absorbing: nothing in the loss opposes further drift once it is entered. The fix has two parts: reconstruct the training sequence exactly as generated, anomaly included, so the penalty's gradient reaches the tokens that produced it; and keep every completion that still contains a translation in the loss, anomaly and format penalty included, rather than filtering it out as invalid, so a gradient opposing the drift exists at all (a completion with no translation receives no gradient; Step 2). We report both fixes together as a general caution for relative-credit training on free-form output with a format requirement, not a translation-specific finding.