TRUST-REGION CONSTRAINT IMPROVES CONTINUAL LEARNING OF SELF-DISTILLATION FINE-TUNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Continually adapting large language models to new domains risks erasing capabilities acquired during pretraining. Self-Distillation Fine-Tuning (SDFT) mitigates this problem by using a demonstration-conditioned version of the model as an on-policy teacher. We show, however, that although this teacher is substantially closer to the student than one-hot supervision on average, its teacher–policy divergence remains strongly heavy-tailed across model families and adaptation domains. A controlled down-weighting intervention further shows that high-divergence teachers contribute disproportionately to forgetting. Motivated by this finding, we introduce Trust-Region Constrained Self-Distillation Fine-Tuning (TRSD), which constructs a sample-wise trust-region-constrained teacher distribution that remains faithful to the demonstration-conditioned teacher while limiting its divergence from the current student. Formulated over the f-divergence family, TRSD induces distinct teacher-distribution geometries for forward KL, reverse KL, and squared Hellinger distance without requiring an additional model forward pass beyond SDFT. Across tool use, scientific question answering, and medical reasoning, TRSD improves instruction-following over SDFT, generally without requiring weaker task accuracy. These gains persist under sequential adaptation and broad hyperparameter sweeps. Our results establish teacher calibration as an important principle for continual self-distillation. Anonymized code is available at `https://anonymous.4open.science/r/TRSD-B218/`.

1 INTRODUCTION

Foundation models are increasingly adapted to new domains, knowledge, and skills after pretraining, making continual learning an important requirement for their deployment (De Lange et al., 2021; Wang et al., 2023). However, learning from new data can overwrite previously acquired capabilities. Supervised fine-tuning (SFT), the standard approach for learning from expert demonstrations, directly trains the model toward externally provided responses. Because these responses form concentrated targets that may lie far from the model's current distribution, SFT can induce substantial model changes and severe catastrophic forgetting (Yang et al., 2024b; Shenfeld et al., 2025).

Self-Distillation Fine-Tuning (SDFT) (Shenfeld et al., 2026) provides an approach to alleviating this problem. Instead of training directly on the original expert response, SDFT trains the model to match the teacher distribution on self-sampled on-policy completions, where the teacher distribution is generated by the student model itself, conditioned on the question and a ground-truth demonstration. Compared with the one-hot target used by SFT, the resulting teacher is expressed through the model's own distribution and remains substantially closer to its current policy. This enables SDFT to acquire the new task while preserving more of the model's existing capabilities.

In this work, we take a closer look at the teacher constructed by SDFT. Although the divergence between the teacher and student distribution is considerably smaller than that induced by the one-hot SFT target on average, we find that this divergence remains strongly heavy-tailed across training demonstrations. Most demonstrations produce a teacher close to the current policy, but a small fraction induces exceptionally large shifts and contributes disproportionately to the overall divergence. We further find that suppressing these high-divergence examples reduces forgetting without degrading new-task performance. These results link SDFT's residual forgetting to the long tail of its teacher–policy divergence.

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.

This finding further motivates us to adaptively regulate the teacher distribution. We introduce *Trust-Region Constrained Self-Distillation Fine-Tuning* (TRSD), which preserves the learning signal introduced by conditioning on the expert demonstration while constraining how far the resulting teacher may deviate from the current policy. Consequently, teacher distributions that are already close to the student remain largely unchanged, whereas excessive teacher shifts are adaptively reduced. We further extend the formulation to f-divergence. Different choices of divergences induce distinct trust-region teacher distributions together with corresponding local budget calibrations, yielding a unified family of constrained self-distillation objectives.

Under both single-task and sequential adaptation on tool use, scientific question answering, and medical reasoning, we evaluate whether TRSD acquires the new skill and whether the model's original instruction-following remains afterward. Across Qwen2.5-7B-Instruct (Yang et al., 2024a) and InternLM2.5-7B-Chat (Cai et al., 2024), in all single-task comparisons, TRSD improves instruction-following over SDFT, with comparable or better task accuracy. Under sequential adaptation, after the final Medical stage, Qwen2.5 shows stronger prior-task retention while incoming Medical accuracy stays within 0.7 points or improves by up to 3.6 points, with InternLM2.5 exhibiting the same overall trend.

Our main contributions are listed as follows: (1) We identify a heavy-tailed divergence between teacher distribution and student model in SDFT across different model families and adaptation domains, and show through a controlled intervention that suppressing its tail reduces forgetting without compromising new-task learning. (2) We introduce Trust-Region Constrained Self-Distillation (TRSD), which adaptively calibrates the teacher distribution through a divergence-constrained projection, directly limiting excessive teacher–policy shifts while preserving informative conditioning signals. (3) We instantiate our approach with forward-KL, reverse-KL, and squared-Hellinger divergences. The experiments demonstrate consistent gains over SDFT objectives in both single-task and sequential adaptation, with robustness across a broad hyperparameter sweep.

2 PRELIMINARIES

We briefly review the adaptation objectives and notation used throughout the paper. Let π_{θ_0} denote the pretrained reference model and π_θ the model being adapted, initialized at $\theta = \theta_0$. Each example $(x, c) \sim \mathcal{D}$ consists of an input x and an expert demonstration c.

Supervised fine-tuning. Supervised fine-tuning (SFT) learns directly from the expert demonstration:

$$\mathcal{L}_{\mathrm{SFT}}(\theta) = -\mathbb{E}_{(x,c)\sim\mathcal{D}}\left[\sum_{t=1}^{|c|}\log \pi_\theta(c_t \mid x, c_{<t})\right] = \mathbb{E}_{(x,c)\sim\mathcal{D}}\left[\sum_{t=1}^{|c|} D_{\mathrm{KL}}(\delta_{c_t} \,\|\, \pi_\theta(\cdot \mid x, c_{<t}))\right]. \tag{1}$$

At each demonstrated prefix $(x, c_{<t})$, this objective can be regarded to converge π_θ to a one-hot target distribution δ_{c_t}, even when multiple continuations may be valid.

Self-Distillation Fine-Tuning. Self-Distillation Fine-Tuning (SDFT) (Shenfeld et al., 2026) replaces the one-hot supervision used in SFT with a soft target distribution induced by the model itself. Specifically, the teacher is conditioned on both the input and the expert demonstration, using a prompt such as *"<Question> This is an example for a response to the question: <Demonstration>. Now answer with a response of your own, including the thinking process:"*. This defines a demonstration-conditioned target distribution $\pi_{\theta_0}(v \mid x, c, y_{<t})$, which captures the information in the demonstration through the model's own predictive distribution rather than through a single one-hot target. SDFT then matches the current policy to this distribution along trajectories sampled from the policy itself:

$$\mathcal{L}_{\mathrm{SDFT}}(\theta) = \mathbb{E}_{(x,c)\sim\mathcal{D}\ y\sim\pi_\theta(\cdot|x)}\left[\frac{1}{|y|}\sum_{t=1}^{|y|} D_{\mathrm{KL}}\left(\pi_{\theta_0}(v \mid x, c, y_{<t}) \| \pi_\theta(v \mid x, y_{<t})\right)\right]. \tag{2}$$

Thus, compared with SFT, SDFT provides a softer target distribution and trains on trajectories closer to the model's own distribution, which can reduce forgetting during continual fine-tuning.

3 REVISITING THE TEACHER DISTRIBUTION IN SDFT

In this section, we revisit the demonstration-conditioned teacher in SDFT. We first provide a local analysis linking target divergence to forgetting, explaining why the in-distribution teacher of SDFT is less destructive than the one-hot target of SFT (§3.1). We then show that this advantage conceals a pronounced heavy tail in the per-example teacher–policy divergence (§3.2). Finally, a controlled down-weighting study shows that suppressing this tail reduces forgetting without degrading new-task performance (§3.3), motivating us to reconsider how the SDFT teacher itself is constructed.

3.1 LOW-DIVERGENCE TARGETS PRESERVE PRETRAINED CAPABILITIES

SDFT has been shown to reduce forgetting without compromising, or even improving substantially, new-task accuracy relative to SFT. At initialization, the teacher distribution is obtained by conditioning the base model on the demonstration c, whereas SFT trains toward an external one-hot target. The SDFT teacher therefore reshapes probability mass already represented by the model rather than concentrating all supervision on a single external sequence. This suggests that the distance between the fine-tuning target and the base policy can be a key determinant of forgetting. The following result formalizes this relationship in the local regime.

Lemma 3.1 (Forgetting is controlled by target divergence, See Proofs in Appendix A.1)**.** *Let π_{θ_0} be the base model and $p(\cdot \mid x)$ the target distribution on new-task inputs. Define the drift on old-capability inputs as $\Phi(\theta) := \mathbb{E}_{x\sim\mu_{\mathrm{old}}} D_{\mathrm{KL}}(\pi_{\theta_0}(\cdot \mid x)\|\pi_\theta(\cdot \mid x))$, which also controls changes in bounded old-capability metrics through Pinsker's inequality. Define the mean target divergence as $\epsilon := \mathbb{E}_{x\sim\mu_{\mathrm{new}}} D_{\mathrm{KL}}(p(\cdot \mid x)\|\pi_{\theta_0}(\cdot \mid x))$. Let F_{old} and F_{new} be the Fisher information matrices induced by μ_{old} and μ_{new} at θ_0. Under the regularity conditions, suppose fine-tuning reaches the target through a sufficiently small update $\Delta\theta = \theta - \theta_0$ in $\mathcal{R} := \mathrm{range}(F_{\mathrm{new}})$. Then*

$$\Phi(\theta) \le \kappa\,\epsilon + O(\epsilon^{3/2}), \qquad \kappa := \max_{v\in\mathcal{R}\setminus\{0\}} \frac{v^\top F_{\mathrm{old}} v}{v^\top F_{\mathrm{new}} v}. \tag{3}$$

The bound factorizes forgetting into a task-geometry constant κ, which is determined by the alignment between old- and new-task curvature, and the target divergence ϵ from the base model, the factor an algorithm controls. At fixed geometry, forgetting scales *linearly* in ϵ, so it is the choice of fine-tuning target that determines how much prior capability survives. This is exactly where SDFT and SFT part ways. SFT regresses toward the one-hot δ_c, which sits far outside π_{θ_0} and yields a large ϵ, incurring proportionally large forgetting. The SDFT teacher instead reshapes the probability mass already represented by the base model. Without forcing all probability onto a single demonstrated answer as SFT does, the SDFT target stays much closer to the base model's original distribution. Fine-tuning against such a low-divergence, on-policy target is thus what lets SDFT preserve pretrained capabilities, and it directly motivates the experiments in Sec. 3.2: despite this overall tendency, can SDFT still produce target distributions that deviate substantially from the original model?

3.2 SDFT EXHIBITS A LONG-TAILED KL-DIVERGENCE DISTRIBUTION

The previous subsection identifies the target divergence ϵ as a key quantity governing forgetting. This raises a natural empirical question: how large is this divergence under SDFT, and does its low-divergence tendency hold uniformly across training examples? To answer this, we examine the per-example target divergence for SDFT and SFT using Qwen2.5-7B-Instruct (Yang et al., 2024a); corresponding experiments on InternLM2.5-7B-Chat (Cai et al., 2024) are provided in Appendix C. Following SDFT (Shenfeld et al., 2026), we consider three adaptation domains: Tool Use, Science Q&A, and Medical, and we use IFEval to evaluate instruction-following retention after training. The detailed setting can be found in §5.1.

Figure 1 compares the empirical distributions of the target divergence induced by SDFT and SFT. Across both model families and all three domains, SDFT shifts the distribution toward substantially smaller values than SFT, confirming its target-proximity advantage. Nevertheless, each SDFT distribution retains a pronounced right tail, with a small fraction of demonstrations inducing much larger shifts than the majority. The recurrence of this pattern across models and domains suggests that it is not specific to a single setting. This recurring pattern motivates us to examine whether suppressing the influence of high-divergence demonstrations can further reduce forgetting without compromising new-task learning.

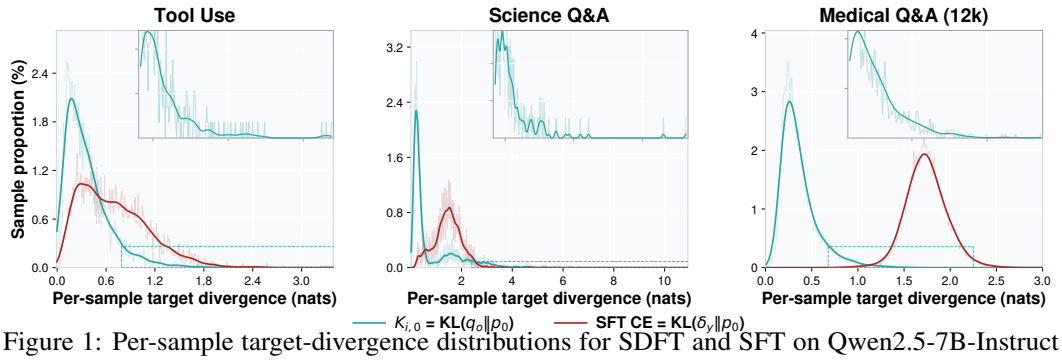

Figure 1: Per-sample target-divergence distributions for SDFT and SFT on Qwen2.5-7B-Instruct across three domains. Solid curves show Gaussian-smoothed proportions, faint curves show raw bins, and insets magnify the right tails. Corresponding InternLM2.5-7B-Chat results appear in Appendix Figure 5.

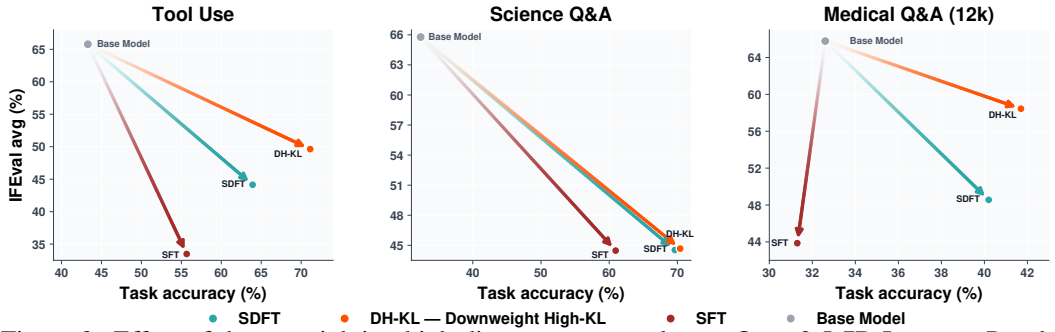

Figure 2: Effect of down-weighting high-divergence examples on Qwen2.5-7B-Instruct. Panels report new-task accuracy and retained IFEval for SFT, SDFT, and high-KL down-weighting (DH-KL) across three domains. Corresponding InternLM2.5-7B-Chat results appear in Appendix Figure 6.

3.3 Downweighting High-KL Training Samples Improves Continual Learning

To test whether the high-divergence tail contributes disproportionately to forgetting, we conduct a simple controlled intervention that *down-weights* the SDFT losses of demonstrations with large $\delta(x, c)$. This intervention preserves the supervision provided by the majority of demonstrations while reducing the influence of extreme teacher shifts. If the tail is an important source of residual forgetting, this targeted suppression should improve retention without compromising new-task learning.

Concretely, we rank demonstrations within each batch using their current teacher–policy divergence, denoted by $\delta_\theta(x, c)$, and assign a reduced weight $\gamma \in [0, 1)$ to the high-divergence tail with $\delta_\theta(x, c) > \tau_{\mathcal{B}}$,

where $\tau_{\mathcal{B}}$ selects an upper divergence quantile across batch $\mathcal{B}$. This intervention leaves the majority of SDFT supervision unchanged while selectively reducing the influence of extreme teacher shifts. We compare it with vanilla SDFT under otherwise identical settings and evaluate both new-task learning and old-capability retention.

Figure 2 shows a consistent pattern across models and domains. Relative to vanilla SDFT, the selected down-weighting configurations improve retained IFEval performance in every case while preserving or increasing new-task accuracy. The retention gains are largest on Tool Use and Medical, with average improvements of 4.03 and 5.51 percentage points across the two model families, respectively. Science Q&A shows a smaller but consistent gain of 0.85 points. Down-weighting also mitigates forgetting on IFEval, improving retained performance by an average of 3.41 points.

These results support our hypothesis that high-divergence demonstrations contribute disproportionately to SDFT's residual forgetting and that suppressing their influence need not compromise new-task learning.

While this prototype verifies the importance of the high-divergence tail, down-weighting remains a coarse intervention: all demonstrations above the threshold receive the same suppression strength, regardless of the extent of their teacher-policy deviation. This motivates our approach in Sec. 4.

4 TRSD: TRUST-REGION CONSTRAINED SELF-DISTILLATION FINE-TUNING

In this section, we propose **T**rust-**R**egion constrained **S**elf-**D**istillation (**TRSD**), which constructs a *trust-region-constrained teacher distribution* with a sample-wise coefficient β adaptively controlling how much of the SDFT teacher is retained. We first formulate TRSD under forward KL and derive the resulting teacher distribution and calibration coefficient in §4.1. We then extend the construction to f-divergences in §4.2, followed by the online sample-wise calibration and three implemented geometries in §4.3.

4.1 TEACHER DISTRIBUTION WITH A TRUST-REGION CONSTRAINT

To go beyond suppressing all samples with high-divergence by a uniform coefficient, we require the target distribution itself to adapt: when it is far from the current student, the adjustment should be more substantial; when it lies within a prescribed range, meaning that the teacher already deviates little from the student distribution, no adjustment is needed. This motivates us to construct a new teacher distribution $q^\star$ with a trust-region constraint (Schulman et al., 2015), which can be obtained by

$$q^\star = \arg\min_q D_{\mathrm{KL}}(q\|\pi_{\theta_0}(\cdot \mid x, c, y_{<t})) \quad \text{s.t.} \quad D_{\mathrm{KL}}(q\|\pi_\theta(\cdot \mid x, y_{<t})) \le C. \tag{4}$$

The objective preserves fidelity to the demonstration-conditioned SDFT teacher, while the constraint keeps the resulting target within the KL neighborhood around the current student defined by C. According to Lemma 3.1, the constrained KL divergence between the teacher distribution and the student can regulate forgetting. Thus, this formulation adaptively calibrates the target-to-student divergence to the prescribed range while retaining the fidelity of the teacher supervision signal. Next, we solve the exact formulation of q^* in Eq. (4) by introducing a Lagrange multiplier for its trust-region constraint.

Lemma 4.1 (Teacher distribution under a forward-KL trust region. Proofs are in Appendix D.2.)**.** *Assume that π_θ and π_{θ_0} have full support under the formulation in Eq. 4. The optimal teacher distribution is*

$$q^\star_{\beta_C}(v) = Z^{-1}_{\beta_C}\Big[\pi_\theta(v \mid x, y_{<t})\Big]^{1-\beta_C}\Big[\pi_{\theta_0}(v \mid x, c, y_{<t})\Big]^{\beta_C}, \tag{5}$$

where Z_β is the normalization constant and $\beta_C = \max\{\beta \in [0,1] : D_{\mathrm{KL}}(q_\beta\|\pi_\theta(\cdot \mid x, y_{<t})) \le C\}$. For a small trust-region budget, β_C can be approximated by

$$\beta_C \approx \min\{1, \sqrt{\frac{2C}{s_{\mathrm{KL}}}}\}, \qquad s_{\mathrm{KL}} = \mathrm{Var}_{v\sim\pi_\theta}\left[\log \frac{\pi_{\theta_0}(v \mid x, c, y_{<t})}{\pi_\theta(v \mid x, y_{<t})}\right]. \tag{6}$$

Thus, the solution geometrically interpolates between the current policy ($\beta = 0$) and the original SDFT teacher ($\beta = 1$), while respecting the trust-region budget.

The lemma solves the teacher distribution in the optimization problem in Eq. (4) with the calibration of a scalar β. In Eq. (5), β controls how much the teacher distribution shifts from the original SDFT teacher: $\beta = 1$ keeps the teacher unchanged, while a smaller β moves the target closer to the current student. Equation (6) determines this coefficient from the divergence budget C and the local curvature s_{KL}. A large s_{KL} means that the teacher strongly upweights some tokens while downweighting others, so moving toward it causes the KL divergence to grow rapidly and therefore requires a smaller β. Conversely, a small s_{KL} indicates that the teacher only mildly reshapes the current distribution, allowing a larger interpolation step under the same divergence budget.

TRSD follows the same on-policy training procedure as SDFT: the student samples a completion $y \sim \pi_\theta(\cdot \mid x)$, and the TRSD teacher distribution $q^\star_{\widehat{\beta}_C}$ is constructed at each position using Eqs. (5) and (6). Replacing the original SDFT teacher with this distribution gives

$$\mathcal{L}^{\mathrm{KL}}_{\mathrm{TRSD}}(\theta) = \mathbb{E}_{\substack{(x,c)\sim\mathcal{D} \\ y\sim\pi_\theta(\cdot|x)}}\left[\frac{1}{|y|}\sum_{t=1}^{|y|} D_{\mathrm{KL}}\big(\mathrm{sg}[q^\star_{\beta_C}]\|\pi_\theta(\cdot \mid x, y_{<t})\big)\right], \tag{7}$$

where $\mathrm{sg}[\cdot]$ denotes the stop-gradient operator. Notably, TRSD requires no additional model forward passes beyond SDFT: $q^\star_{\beta_C}$ is obtained directly from the already-computed student and teacher distributions through a lightweight reweighting step. TRSD therefore introduces the trust-region constraint with negligible additional computational overhead while preserving the original SDFT training pipeline.

4.2 Extension to f-Divergence

Although KL divergence is a standard choice for distillation, alternative divergences offer additional flexibility to induce different optimization behaviors (Gu et al., 2024; Wen et al., 2023).

We capture these choices with the f-divergence family.

Definition 4.1 (f-divergence class)**.** *Let a and b be probability distributions over $\mathcal{V}$ with $a \ll b$. For any convex function $f : (0, \infty) \to \mathbb{R}$ satisfying $f(1) = 0$, the corresponding f-divergence is $\mathcal{D}_f(a\|b) := \sum_{v\in\mathcal{V}} b(v)\, f\left(\frac{a(v)}{b(v)}\right)$. It is nonnegative and equals zero when $a = b$.*

Given that the f-divergence (Gu et al., 2024; Wen et al., 2023) between the teacher distribution and the student distribution also controls forgetting, as proved in Lemma A.2 in the appendix, we extend the formulation of the TRSD teacher distribution to f-divergence as follows,

$$q_f^\star = \arg\min_q \mathcal{D}_f(q\|\pi_{\theta_0}(\cdot \mid x, c, y_{<t})) \quad \text{s.t.} \quad \mathcal{D}_f(q\|\pi_\theta(\cdot \mid x, y_{<t})) \le C. \tag{8}$$

Similarly, the objective preserves fidelity to the SDFT teacher, while the f-divergence constraint regulates forgetting with C. Next, we derive the formulation of $q_f^\star$ in the above optimization problem.

Lemma 4.2 (Teacher distribution under an f-divergence trust region. Proofs are in Appendix D.1)**.** *Suppose the current student and the demonstration-conditioned teacher have full support. The optimal teacher distribution of Eq. 8 can be formulated as*

$$q^\star_{f,\beta}(v) = G^{-1}_{f,\beta,v}(-\eta), \tag{9}$$

where $G_{f,\beta,v}(q(v)) := \beta f'(q(v)/\pi_{\theta_0}(v \mid x, c, y_{<t})) + (1-\beta) f'(q(v)/\pi_\theta(v \mid x, y_{<t}))$, η enforces the normalization of $q^\star_{f,\beta}$, $q(v)$ denotes the probability assigned to vocabulary item v by the optimization variable q in Eq. (8), and $G^{-1}_{f,\beta,v}$ denotes the functional inverse with respect to $q(v)$. The coefficient selected by the trust-region constraint is $\beta_{f,C} = \max\{\beta \in [0, 1] : \mathcal{D}_f(q^\star_{f,\beta}\|\pi_\theta(\cdot \mid x, y_{<t})) \le C\}$. For a small trust-region budget, $\beta_{f,C}$ can be approximated by

$$\beta_{f,C} \approx \min\left\{1, \sqrt{\frac{2C}{s_f}}\right\}, \qquad s_f := \frac{\mathrm{d}^2}{\mathrm{d}\beta^2}\mathcal{D}_f\big(q^\star_{f,\beta}\|\pi_\theta(\cdot \mid x, y_{<t})\big)\Big|_{\beta=0}. \tag{10}$$

Lemma 4.2 derives the TRSD teacher distribution and its locally calibrated coefficient for the f-divergence formulation. The complete KKT and local-curvature derivations are provided in Appendix D.1.

f-divergence training objective. Using the teacher distribution from Lemma 4.2 in place of the original SDFT teacher gives

$$\mathcal{L}^f_{\mathrm{TRSD}}(\theta) = \mathbb{E}_{\substack{(x,c)\sim\mathcal{D} \\ y\sim\pi_\theta(\cdot|x)}}\left[\frac{1}{|y|}\sum_{t=1}^{|y|}\mathcal{D}_f\Big(\mathrm{sg}\Big[q^\star_{f,\beta_{f,C}}\Big]\,\|\pi_\theta(\cdot \mid x, y_{<t})\Big)\right]. \tag{11}$$

As in the KL case, $\mathrm{sg}[\cdot]$ holds the TRSD teacher distribution fixed during the student update. When $\beta_{f,C} = 1$, this distribution equals the original demonstration-conditioned teacher, recovering the corresponding SDFT objective.

4.3 Practical Implementation

Sample-wise calibration. A fixed absolute budget C can correspond to different degrees of adjustment across tasks, divergence geometries, and training stages. We therefore determine the budget online from the current batch. Specifically, we average the local curvature s_f over each on-policy completion and use ρ to scale the batch mean into an online budget. The complete formulation is provided in Appendix Eq. (37). Substituting this online peer-relative budget into Eq. (10) calibrates each sample against the current batch: a sample with larger curvature reaches the budget at a smaller β, whereas one with lower curvature retains more of the original SDFT teacher. We further clip the resulting coefficient to $[\beta_{\min}, \beta_{\max}]$; Appendix D.5 provides the complete online rule.

Divergence instantiations. We instantiate TRSD with forward KL, reverse KL, and squared Hellinger distance. For each choice, substituting its generator into Eq. (9) determines the TRSD teacher distribution $q^{\star}_{f,\beta,t}$, while the corresponding local curvature provides $s_{f,i}$ used for the online calibration described above. Table 1 summarizes both quantities. Forward KL follows a geometric path, reverse KL an arithmetic path, and squared Hellinger a path in square-root probability space. Their local curvatures quantify how rapidly each path departs from the current student; Appendices D.2–D.4 define these quantities and derive the corresponding paths. For reverse KL, the exact curvature is Pearson χ^2, while the implemented peer calibration uses the bounded relative-Pearson surrogate in Eq. (33) (Yamada et al., 2011). This substitution stabilizes the coefficient without changing the resulting teacher path or training objective.

Table 1: TRSD teacher geometries and local curvatures.

Divergence / generator	TRSD teacher $q^{\star}_{f,\beta}$	Local curvature s_f
Forward KL $f(u) = u \log u$	$Z_\beta^{-1}(p_\theta^t)^{1-\beta}(q_0^t)^\beta$	$\mathrm{Var}_{v\sim p_\theta^t}\left[\log \frac{q_0^t(v)}{p_\theta^t(v)}\right]$
Reverse KL $f(u) = -\log u$	$(1-\beta)p_\theta^t + \beta q_0^t$	$\sum_{v\in\mathcal{V}} \frac{(q_0^t(v) - p_\theta^t(v))^2}{p_\theta^t(v)}$
Squared Hellinger $f(u) = \frac{1}{2}(\sqrt{u}-1)^2$	$Z_\beta^{-1}\left((1-\beta)\sqrt{p_\theta^t} + \beta\sqrt{q_0^t}\right)^2$	$1 - \left(\sum_{v\in\mathcal{V}} \sqrt{p_\theta^t(v) q_0^t(v)}\right)^2$

$p_\theta^t := \pi_\theta(\cdot \mid x, y_{<t})$ is the current student, $q_0^t := \pi_{\theta_0}(\cdot \mid x, c, y_{<t})$ is the demonstration-conditioned teacher, and Z_β is the normalization constant.

5 Experiments

5.1 Experimental Settings

We evaluate continual skill acquisition on three domains following SDFT (Shenfeld et al., 2026): Science Q&A using the Chemistry L-3 subset of SciKnowEval (Feng et al., 2024), Tool Use using ToolAlpaca (Tang et al., 2023), and Medical using the stage-1 HuatuoGPT-o1 corpus (Chen et al., 2024) for training and held-out Medical Verifiable Problems for evaluation. Skill acquisition is measured by task accuracy, while retention of general instruction-following ability is evaluated on IFEval (Zhou et al., 2023) using the average of its four standard metrics. We experiment with Qwen2.5-7B-Instruct (Yang et al., 2024a) and InternLM2.5-7B-Chat (Cai et al., 2024), comparing TRSD against SFT, DFT (Wu et al., 2025), and SDFT (Shenfeld et al., 2026). To isolate the effect of trust-region teacher construction, both SDFT and TRSD are instantiated with forward KL, reverse KL, and squared Hellinger distance. We evaluate both single-task adaptation, where each domain is learned independently from the same initialization, and sequential continual learning over Tool Use, Science Q&A, and Medical, evaluating the current task, previously learned tasks, and IFEval after each stage. Appendix E provides the training procedure in Algorithm 1 and the full configurations.

5.2 Results and Analyses

Single-task adaptation. Table 2 compares new-skill acquisition with the retention of general instruction-following capabilities. On Qwen2.5-7B-Instruct, TRSD improves both task accuracy and aggregate IFEval over SDFT across all nine matched domain–geometry comparisons, while also outperforming SFT and DFT. Because each matched pair shares the same divergence objective, these improvements isolate the effect of constraining the original demonstration-conditioned teacher through TRSD. For example, under forward KL on Medical, TRSD raises IFEval from 48.56 to 60.91 while increasing Medical accuracy from 40.20 to 40.90; under squared Hellinger on Tool Use, it improves task accuracy from 62.89 to 72.16 and IFEval from 53.14 to 58.08. The consistency of these gains across all three divergence geometries supports the general f-divergence formulation, rather than attributing the benefit to a particular divergence instantiation. InternLM2.5-7B-Chat shows the same overall pattern: TRSD improves IFEval in all nine matched comparisons and task accuracy in eight. The only task-accuracy decrease occurs under reverse KL on Medical, where a 1.3-point reduction is accompanied by a 3.9-point IFEval improvement. Complete InternLM configuration details are provided in Appendix E.3.

Sequential continual learning. Figure 3 reports Qwen2.5-7B-Instruct comparisons with matched SDFT throughout sequential Tool Use → Science Q&A → Medical adaptation. SFT and DFT

Table 2: Single-task skill acquisition and instruction-following retention on Qwen2.5-7B-Instruct and InternLM2.5-7B-Chat. All values are percentages. Base denotes the initial model.

Geometry	Method	Tool Use		Science Q&A		Medical	
		Acc.	IFEval	Acc.	IFEval	Acc.	IFEval
Qwen2.5-7B-Instruct							
	Base	43.30	65.78	32.35	65.78	32.60	65.78
	SFT	55.67	33.48	60.95	44.47	31.30	43.86
	DFT	56.70	45.34	61.70	49.72	32.60	48.28
Forward KL	SDFT	63.92	44.14	65.88	45.60	40.20	48.56
	TRSD (ours)	**65.98**	**57.81**	**68.24**	**53.06**	**40.90**	**60.91**
Reverse KL	SDFT	64.95	43.80	67.06	49.82	38.60	49.01
	TRSD (ours)	**67.01**	**55.62**	**67.65**	**55.88**	**42.30**	**54.73**
Squared Hellinger	SDFT	62.89	53.14	65.88	49.72	37.60	55.73
	TRSD (ours)	**72.16**	**58.08**	**68.07**	**53.40**	**40.20**	**62.77**
InternLM2.5-7B-Chat							
	Base	31.96	57.56	27.02	57.56	27.40	57.56
	SFT	55.67	32.03	33.53	46.12	28.30	44.97
	DFT	39.18	31.71	55.42	41.50	25.40	48.64
Forward KL	SDFT	48.45	46.59	50.69	45.95	30.30	46.58
	TRSD (ours)	**49.48**	**50.25**	**56.80**	**47.73**	**30.90**	**51.65**
Reverse KL	SDFT	39.18	46.39	53.85	47.27	30.20	47.29
	TRSD (ours)	**47.42**	**54.82**	**57.00**	**51.22**	**28.90**	**51.14**
Squared Hellinger	SDFT	43.30	49.80	48.72	46.14	30.70	46.09
	TRSD (ours)	**47.42**	**53.44**	**53.85**	**48.33**	**30.90**	**50.92**

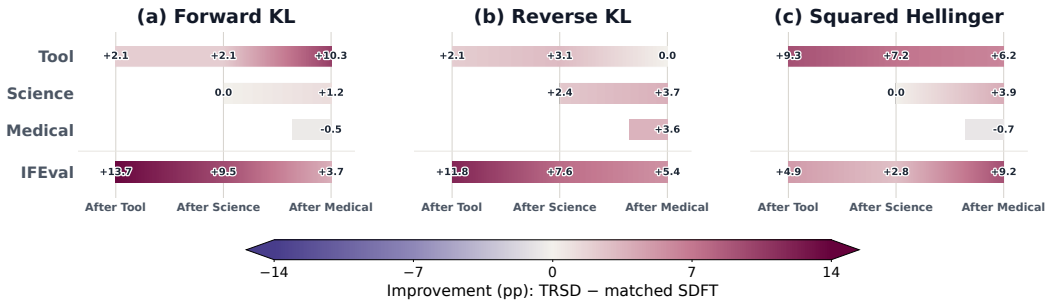

Figure 3: TRSD improvements over matched SDFT during sequential Tool Use → Science Q&A → Medical adaptation on Qwen2.5-7B-Instruct. Panels report percentage-point differences for all skills introduced by each stage and for IFEval. Colors are clipped at ± 14 points.

reach 66.86 and 65.09 Science accuracy but retain only 0.99 and 23.08 after Medical, whereas all three TRSD variants retain at least 67.26 while attaining at least 37.20 Medical accuracy (Figure 7). Generation analysis attributes this collapse to answer-format and scientific-decision interference; trajectories, examples, and per-divergence comparisons appear in Appendices E.7.1, E.7.2, and E.7. Against matched SDFT, after Science training all three TRSD variants improve Tool Use retention and IFEval by up to 7.2 and 9.5 points, while matching or improving Science by up to 2.4 points. After Medical training, none loses Tool Use relative to matched SDFT, and gains in Science retention and IFEval reach 3.9 and 9.2 points; incoming Medical remains within 0.7 points or improves by up to 3.6 points. InternLM2.5-7B-Chat corroborates this pattern: after Science training, all variants improve IFEval by at least 2.8 points, while two of three improve Tool retention, and two improve Science acquisition, by up to 14.4 and 7.7 points, respectively. After Medical training, all variants improve prior Science, incoming Medical, and IFEval; notably, squared-Hellinger SDFT collapses to 0.00% Science accuracy while TRSD retains 51.87% (Figure 9). Thus, the retention gains generally do not require weaker acquisition of the incoming task; complete stage-wise grids are reported in Appendices E.7.3 and E.8.

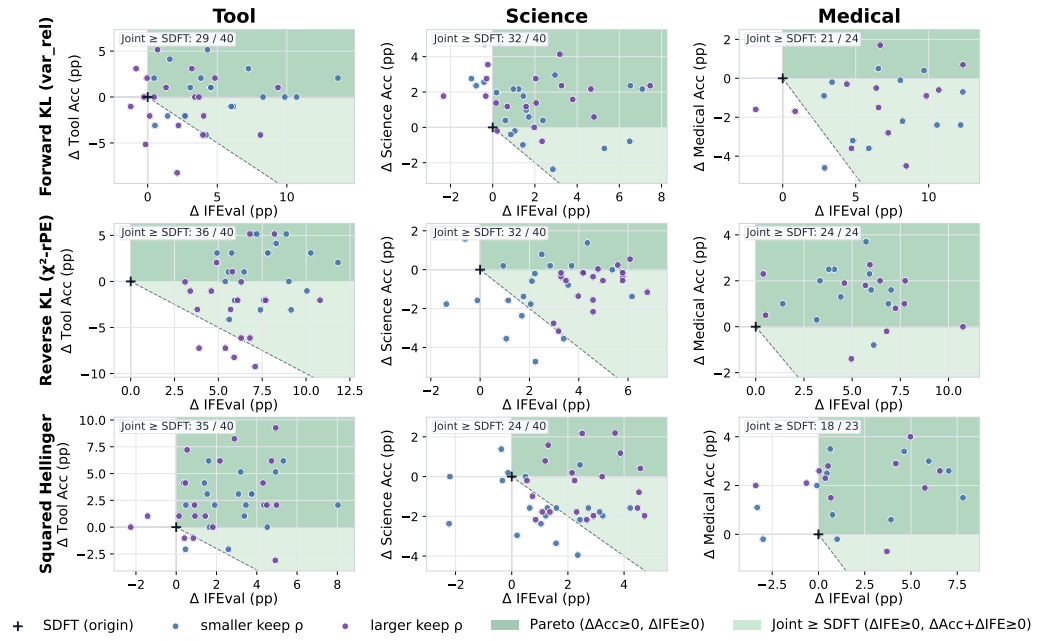

Figure 4: Single-task robustness of TRSD. Dark shading marks $\Delta\mathrm{Acc} \geq 0$ and $\Delta\mathrm{IFEval} \geq 0$; the combined shaded region marks $\Delta\mathrm{IFEval} \geq 0$ and $\Delta J = \frac{1}{2}(\Delta\mathrm{Acc} + \Delta\mathrm{IFEval}) \geq 0$.

Ablation: Robustness to TRSD teacher-calibration hyperparameters. On Qwen2.5-7B-Instruct, we examine whether the gains of TRSD depend on isolated hyperparameter choices by sweeping two peer-relative budget scales and the associated clipping grids for every divergence–domain pair. Figure 4 reports each configuration relative to its matched SDFT baseline in task accuracy and aggregate IFEval. Across the complete sweep, 251 of 311 configurations (80.7%) improve the retention-aligned joint score over matched SDFT, and 157 configurations (50.5%) improve both task accuracy and IFEval simultaneously. The shaded region covers all nine divergence–domain pairs, showing that the gains are not confined to some specific configurations. Strict simultaneous improvements are also observed in every pair, indicating that this behavior is not particular to a single objective or task. Complete results for all configurations are provided in Appendix E.9.

6 Related Work

TRSD lies at the intersection of continual adaptation and forgetting (De Lange et al., 2021; Wang et al., 2023), on-policy contextual self-distillation (Shenfeld et al., 2026; Zhao et al., 2026; Hübotter et al., 2026; Ye et al., 2026), and trust-region optimization (Schulman et al., 2015). Motivated by heavy-tailed self-teacher shifts, TRSD differs from parameter regularization (Kirkpatrick et al., 2017; Li & Hoiem, 2018), on-policy replay (Chen et al., 2026), and fixed teacher interpolation (Xie et al., 2026) by directly calibrating the demonstration-conditioned teacher through a sample-wise f-divergence constraint. Appendix F provides expanded discussion and positioning.

7 Conclusion

We show that the demonstration-conditioned SDFT teacher, despite being closer to the student than one-hot supervision on average, exhibits a heavy-tailed teacher–policy divergence whose tail contributes disproportionately to forgetting. TRSD addresses this failure mode through a sample-wise trust-region-constrained teacher distribution. Its unified f-divergence formulation supports forward KL, reverse KL, and squared Hellinger without an auxiliary teacher, replay buffer, or additional model forward pass beyond SDFT. Across single-task and sequential adaptation, TRSD improves retention over matched SDFT while maintaining or strengthening new-skill acquisition, establishing teacher calibration against the evolving student as a useful principle for continual self-distillation.

Reproducibility Statement

Complete derivations, dataset-processing details, training configurations, hyperparameter selections, and full experimental results are provided in the appendix. Anonymized code is available at `https://anonymous.4open.science/r/TRSD-B218/`.

AI Use Statement

AI tools were used solely to assist with code review and language editing. All figures and reported numerical results were produced from genuine experimental data; no synthetic or AI-generated data were used.

References

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

Zheng Cai, Maosong Cao, Haojiong Chen, Kai Chen, Keyu Chen, Xin Chen, Xun Chen, Zehui Chen, Zhi Chen, Pei Chu, Xiaoyi Dong, Haodong Duan, Qi Fan, Zhaoye Fei, Yang Gao, Jiaye Ge, Chenya Gu, Yuzhe Gu, Tao Gui, Aijia Guo, Qipeng Guo, Conghui He, Yingfan Hu, Ting Huang, Tao Jiang, Penglong Jiao, Zhenjiang Jin, Zhikai Lei, Jiaxing Li, Jingwen Li, Linyang Li, Shuaibin Li, Wei Li, Yining Li, Hongwei Liu, Jiangning Liu, Jiawei Hong, Kaiwen Liu, Kuikun Liu, Xiaoran Liu, Chengqi Lv, Haijun Lv, Kai Lv, Li Ma, Runyuan Ma, Zerun Ma, Wenchang Ning, Linke Ouyang, Jiantao Qiu, Yuan Qu, Fukai Shang, Yunfan Shao, Demin Song, Zifan Song, Zhihao Sui, Peng Sun, Yu Sun, Huanze Tang, Bin Wang, Guoteng Wang, Jiaqi Wang, Jiayu Wang, Rui Wang, Yudong Wang, Ziyi Wang, Xingjian Wei, Qizhen Weng, Fan Wu, Yingtong Xiong, Chao Xu, Ruiliang Xu, Hang Yan, Yirong Yan, Xiaogui Yang, Haochen Ye, Huaiyuan Ying, Jia Yu, Jing Yu, Yuhang Zang, Chuyu Zhang, Li Zhang, Pan Zhang, Peng Zhang, Ruijie Zhang, Shuo Zhang, Songyang Zhang, Wenjian Zhang, Wenwei Zhang, Xingcheng Zhang, Xinyue Zhang, Hui Zhao, Qian Zhao, Xiaomeng Zhao, Fengzhe Zhou, Zaida Zhou, Jingming Zhuo, Yicheng Zou, Xipeng Qiu, Yu Qiao, and Dahua Lin. Internlm2 technical report, 2024.

Howard Chen, Noam Razin, Karthik Narasimhan, and Danqi Chen. Retaining by doing: The role of on-policy data in mitigating forgetting. *arXiv preprint arXiv:2510.18874*, 2025.

Junying Chen, Zhenyang Cai, Ke Ji, Xidong Wang, Wanlong Liu, Rongsheng Wang, Jianye Hou, and Benyou Wang. Huatuogpt-o1, towards medical complex reasoning with llms. *arXiv preprint arXiv:2412.18925*, 2024.

Yan Chen, Taojie Zhu, Meng Zhang, Xin Chen, Jiaqi Huang, Dongyang Xu, and Yizhi Wang. On-policy replay for continual supervised fine-tuning. *arXiv preprint arXiv:2605.29495*, 2026.

Matthias De Lange, Rahaf Aljundi, Marc Masana, Sarah Parisot, Xu Jia, Aleš Leonardis, Gregory Slabaugh, and Tinne Tuytelaars. A continual learning survey: Defying forgetting in classification tasks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(7):3366–3385, 2021.

Kehua Feng, Xinyi Shen, Weijie Wang, Xiang Zhuang, Yuqi Tang, Qiang Zhang, and Keyan Ding. Sciknoweval: Evaluating multi-level scientific knowledge of large language models. *arXiv preprint arXiv:2406.09098*, 2024.

Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. Minillm: Knowledge distillation of large language models. In *International Conference on Learning Representations*, 2024.

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

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114 (13):3521–3526, 2017.

Song Lai, Haohan Zhao, Rong Feng, Changyi Ma, Wenzhuo Liu, Hongbo Zhao, Xi Lin, Dong Yi, Qingfu Zhang, Hongbin Liu, Gaofeng Meng, and Fei Zhu. Reinforcement fine-tuning naturally mitigates forgetting in continual post-training, 2026. URL https://arxiv.org/abs/2507.05386.

Zhizhong Li and Derek Hoiem. Learning without forgetting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(12):2935–2947, 2018.

Stéphane Ross, Geoffrey J. Gordon, and J. Andrew Bagnell. A reduction of imitation learning and structured prediction to no-regret online learning. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pp. 627–635, 2011.

John Schulman, Sergey Levine, Pieter Abbeel, Michael Jordan, and Philipp Moritz. Trust region policy optimization. In *Proceedings of the 32nd International Conference on Machine Learning*, pp. 1889–1897, 2015.

Idan Shenfeld, Jyothish Pari, and Pulkit Agrawal. RL's razor: Why online reinforcement learning forgets less. *arXiv preprint arXiv:2509.04259*, 2025.

Idan Shenfeld, Mehul Damani, Jonas Hübotter, and Pulkit Agrawal. Self-distillation enables continual learning. *arXiv preprint arXiv:2601.19897*, 2026.

Charlie Snell, Dan Klein, and Ruiqi Zhong. Learning by distilling context. *arXiv preprint arXiv:2209.15189*, 2022.

Qiaoyu Tang, Ziliang Deng, Hongyu Lin, Xianpei Han, Qiao Liang, Boxi Cao, and Le Sun. Toolalpaca: Generalized tool learning for language models with 3000 simulated cases. *arXiv preprint arXiv:2306.05301*, 2023.

Xiao Wang, Yuansen Zhang, Tianze Chen, Songyang Gao, Senjie Jin, Xianjun Yang, Zhiheng Xi, Rui Zheng, Yicheng Zou, Tao Gui, Qi Zhang, and Xuanjing Huang. TRACE: A comprehensive benchmark for continual learning in large language models. *arXiv preprint arXiv:2310.06762*, 2023.

Yuqiao Wen, Zichao Li, Wenyu Du, and Lili Mou. f-divergence minimization for sequence-level knowledge distillation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, pp. 10817–10834, 2023. doi: 10.18653/v1/2023.acl-long.605.

Yongliang Wu, Yizhou Zhou, Ziheng Zhou, Yingzhe Peng, Xinyu Ye, Xinting Hu, Wenbo Zhu, Lu Qi, Ming-Hsuan Yang, and Xu Yang. On the generalization of sft: A reinforcement learning perspective with reward rectification. *arXiv preprint arXiv:2508.05629*, 2025.

Zhengpeng Xie, Li Lyna Zhang, Zeke Xie, and Mao Yang. Trust region policy distillation. *arXiv preprint arXiv:2607.04751*, 2026.

Makoto Yamada, Taiji Suzuki, Takafumi Kanamori, Hirotaka Hachiya, and Masashi Sugiyama. Relative density-ratio estimation for robust distribution comparison. In *Advances in Neural Information Processing Systems*, volume 24, 2011.

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, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024a.

Zhaorui Yang, Tianyu Pang, Haozhe Feng, Han Wang, Wei Chen, Minfeng Zhu, and Qian Liu. Self-distillation bridges distribution gap in language model fine-tuning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, pp. 1028–1043, 2024b.

Tianzhu Ye, Li Dong, Xun Wu, Shaohan Huang, and Furu Wei. On-policy context distillation for language models. *arXiv preprint arXiv:2602.12275*, 2026.

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

Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. Instruction-following evaluation for large language models. *arXiv preprint arXiv:2311.07911*, 2023.

A Detailed Theoretical Analyses

A.1 Forgetting is bounded by the target-to-base KL

We give the detailed proof of Lemma 3.1 and then extend the result to smooth f-divergences. For compactness, x below indexes a prediction event. In token-level SDFT, its distribution μ_{new} includes the sampled example (x, c), a frozen on-policy prefix $y_{<t}$, and the uniform token-position weighting $1/|y|$ from Eq. (2); $\pi_\theta(\cdot \mid x)$ denotes the student conditional at that event. Let $p(\cdot \mid x)$ be the target distribution toward which fine-tuning drives the student. At initialization, the SDFT target is the demonstration-conditioned teacher, so

$$p(\cdot \mid x) \;=\; \pi_{\theta_0}(\cdot \mid x, c), \qquad \text{so} \qquad D_{\mathrm{KL}}\big(p(\cdot \mid x) \,\|\, \pi_{\theta_0}(\cdot \mid x)\big) = \delta(x, c)\big|_{\theta_0}, \tag{12}$$

where $\delta(x, c)$ is the per-sample forward KL underlying Eq. (2).

Quantities. Old-capability prediction contexts are drawn from μ_{old}, while new-task prediction events follow μ_{new} as defined above. *Forgetting* is the drift of the trained model from its initialization on old contexts,

$$\Phi(\theta) := \mathbb{E}_{x\sim\mu_{\text{old}}}\big[D_{\mathrm{KL}}(\pi_{\theta_0}(\cdot \mid x) \,\|\, \pi_\theta(\cdot \mid x))\big], \tag{13}$$

and the mean *target divergence* is the target-to-base KL on new inputs,

$$\epsilon := \mathbb{E}_{x\sim\mu_{\text{new}}}\big[D_{\mathrm{KL}}(p(\cdot \mid x) \,\|\, \pi_{\theta_0}(\cdot \mid x))\big] \overset{(12)}{=} \mathbb{E}_{(x,c)\sim\mathcal{D}}\big[\delta(x, c)\big|_{\theta_0}\big]. \tag{14}$$

By Pinsker, any old-task metric g with $\|g\|_\infty \le B$ obeys $|\mathbb{E}_{\mu_{\text{old}}}[\mathbb{E}_{\pi_{\theta_0}} g - \mathbb{E}_{\pi_\theta} g]| \le B\sqrt{2\,\Phi(\theta)}$, so bounding Φ bounds observable degradation.

Assumptions. *(A1) Local smoothness.* The model has common support near θ_0, $\log \pi_\theta$ is three times continuously differentiable, and its third-order derivatives admit an integrable uniform bound on the neighborhood considered. Define the domain Fisher matrices

$$F_\bullet = \mathbb{E}_{x\sim\mu_\bullet}\,\mathbb{E}_{y\sim\pi_{\theta_0}(\cdot|x)}\big[\nabla_\theta \log \pi_{\theta_0}(y \mid x)\, \nabla_\theta \log \pi_{\theta_0}(y \mid x)^\top\big], \quad \bullet \in \{\text{old}, \text{new}\}.$$

(A2) Local target matching. The target is locally realizable and reached by fine-tuning: for a sufficiently small $\Delta\theta := \theta - \theta_0$, $p(\cdot \mid x) = \pi_{\theta_0+\Delta\theta}(\cdot \mid x)$ on μ_{new}. *(A3) Nondegenerate update geometry.* The update lies in $\mathcal{R} := \mathrm{range}(F_{\text{new}})$, and F_{new} is positive definite on $\mathcal{R}$.

Because the Fisher information is the local Hessian of KL, (A1) gives

$$\mathbb{E}_{x\sim\mu_\bullet}\big[D_{\mathrm{KL}}(\pi_{\theta_0} \,\|\, \pi_\theta)\big] = \tfrac{1}{2}\,\Delta\theta^\top F_\bullet\, \Delta\theta + O(\|\Delta\theta\|^3), \tag{15}$$

and reversing the KL arguments gives the same quadratic term.

Lemma A.1 (Target divergence sizes the update)**.** *Under* (A1)–(A3)*,*

$$\tfrac{1}{2}\,\Delta\theta^\top F_{\text{new}}\, \Delta\theta = \epsilon + O(\epsilon^{3/2}).$$

Proof. By (A2), the target-to-base divergence in Eq. (14) is the reverse orientation of the local model displacement. Equation (15) therefore gives $\epsilon = \frac{1}{2}\Delta\theta^\top F_{\text{new}}\Delta\theta + O(\|\Delta\theta\|^3)$. Assumption (A3) implies $\|\Delta\theta\| = O(\sqrt{\epsilon})$, converting the remainder to $O(\epsilon^{3/2})$. □

Proof of Lemma 3.1. By Eq. (15) on μ_{old}, $\Phi(\theta) = \frac{1}{2}\Delta\theta^\top F_{\text{old}}\Delta\theta + O(\|\Delta\theta\|^3)$. Since $\Delta\theta \in \mathcal{R}$ by (A3), the generalized Rayleigh quotient gives $\Delta\theta^\top F_{\text{old}}\Delta\theta \le \kappa\, \Delta\theta^\top F_{\text{new}}\Delta\theta$. Substituting Lemma A.1 yields Eq. (3). □

Lemma A.2 (f-divergence extension)**.** *Let $\mathcal{D}_f$ be the f-divergence used in Sec. 4, with f three times continuously differentiable near 1 and $c_f := f''(1) > 0$. Under* (A1)–(A3)*, define*

$$\epsilon_f := \mathbb{E}_{x\sim\mu_{\text{new}}}\left[\mathcal{D}_f(p(\cdot \mid x) \,\|\, \pi_{\theta_0}(\cdot \mid x))\right]. \tag{16}$$

Then

$$\Phi(\theta) \le \frac{\kappa}{c_f}\,\epsilon_f + O(\epsilon_f^{3/2}). \tag{17}$$

Proof. Adding an affine term to f does not change $\mathcal{D}_f$, so we may take $f'(1) = 0$. For distributions q near an anchor a, Taylor expansion of the generator around 1 gives

$$\mathcal{D}_f(q\|a) = \frac{c_f}{2}\sum_v \frac{(q(v)-a(v))^2}{a(v)} + o\big(\|q-a\|_2^2\big)\,. \tag{18}$$

Applying this shared local geometry to a parametric model gives

$$\mathbb{E}_{x\sim\mu_\bullet}\left[\mathcal{D}_f(\pi_{\theta_0+\Delta\theta}(\cdot \mid x)\|\pi_{\theta_0}(\cdot \mid x))\right] = \frac{c_f}{2}\,\Delta\theta^\top F_\bullet \Delta\theta + O(\|\Delta\theta\|^3). \tag{19}$$

Using (A2) on the new-task contexts yields $\epsilon_f = \frac{c_f}{2}\Delta\theta^\top F_{\text{new}}\Delta\theta + O(\|\Delta\theta\|^3)$. Meanwhile, Eq. (15) and the generalized Rayleigh quotient give $\Phi(\theta) \le \frac{\kappa}{2}\Delta\theta^\top F_{\text{new}}\Delta\theta + O(\|\Delta\theta\|^3)$. Combining these expressions and using (A3) to write $\|\Delta\theta\| = O(\sqrt{\epsilon_f})$ proves Eq. (17). □

For the forward- and reverse-KL generators, $c_f = 1$; for the squared-Hellinger generator in Table 1, $c_f = 1/4$. The factor $1/c_f$ reflects the arbitrary scale of a divergence generator: normalizing the generator so that $f''(1) = 1$ recovers the same leading coefficient κ for every smooth f-divergence.

Corollary A.1 (Bounded-metric form)**.** *For any old-task metric with $\|g\|_\infty \le B$, $|\Delta\text{perf}| \le B\sqrt{2\kappa\,\epsilon + O(\epsilon^{3/2})}$.*

Target-proximity reading and the SFT contrast. For SDFT, ϵ is exactly the expected teacher-to-base forward KL at initialization. For a one-hot SFT target, the analogous target-to-base KL is the negative log-likelihood of the demonstrated token or sequence. There is no universal ordering between these quantities without assumptions on the data and teacher. Empirically, Figure 1 shows that the SDFT targets are closer to the base policy in the settings than SFT, though it still has a long heavy tail. Lemma 3.1 explains why this observed reduction tightens the local forgetting bound.

Scope and link to the long tail. Eq. (3) is a second-order trust-region result: it characterizes the local $O(\epsilon)$ upper bound on forgetting at fixed geometry κ under small, locally target-matching updates. Because the bound depends on the *mean* target-to-base KL $\epsilon = \mathbb{E}[\delta(x,c)|_{\theta_0}]$, a heavy right tail of high-δ samples (Sec. 3.2) can contribute disproportionately to ϵ. The down-weighting intervention of Sec. 3.3 reduces the optimization influence of precisely these samples, while TRSD directly constrains the corresponding target divergence.

B Details of High-Divergence Down-weighting

The diagnostic intervention in Sec. 3.3 uses the same teacher–student divergence optimized by the forward-KL SDFT objective. For each example $i = (x_i, c_i)$ in a rollout batch $\mathcal{B}$, let y_i denote its on-policy completion and define

$$K_i(\theta) := \frac{1}{|y_i|}\sum_{t=1}^{|y_i|} D_{\text{KL}}\big(q_{0,i}^t \,\|\, p_{\theta,i}^t\big)\,, \tag{20}$$

where $q_{0,i}^t$ is the demonstration-conditioned teacher and $p_{\theta,i}^t$ is the current student distribution along the same completion. At initialization, $K_i(\theta_0)$ is the token-level empirical counterpart of the target divergence defined in Sec. 3.2; during training, recomputing $K_i(\theta)$ measures the divergence of the teacher from the current student. The ranking is computed jointly over the rollout batch, and the resulting weights remain fixed for the corresponding optimization update.

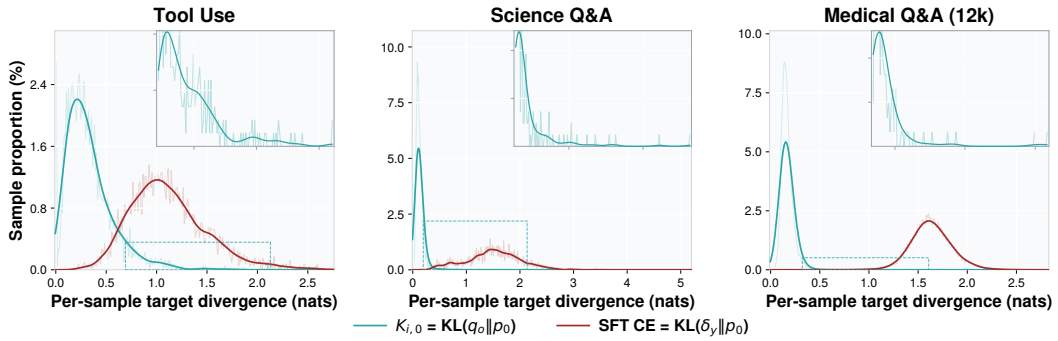

Figure 5: Per-sample forward-KL distributions for InternLM2.5-7B-Chat across three adaptation domains. The corresponding Qwen2.5-7B-Instruct results are shown in Figure 1.

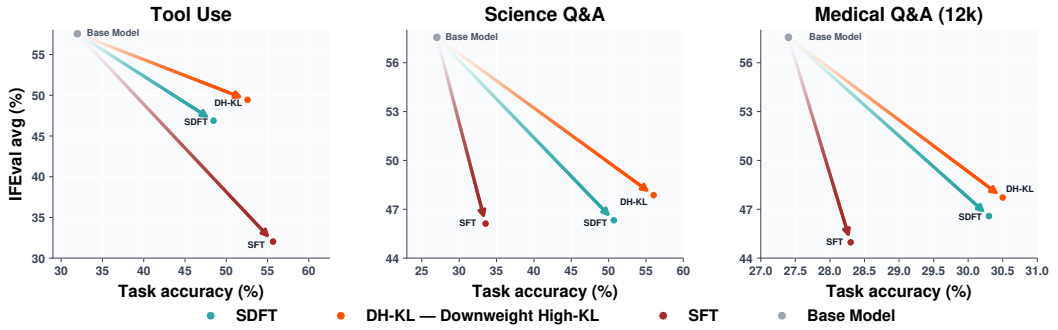

Figure 6: Effect of down-weighting high-divergence training examples for InternLM2.5-7B-Chat. The corresponding Qwen2.5-7B-Instruct results are shown in Figure 2.

Let $\eta \in (0, 1]$ be the fraction of examples retained at full weight and $k = \max\{1, \mathrm{round}(\eta|\mathcal{B}|)\}$. We rank the examples in ascending order of $K_i(\theta)$ within $\mathcal{B}$ and assign

$$w_i = \begin{cases} 1, & \mathrm{rank}_{\mathcal{B}}(K_i(\theta)) \le k, \\ \gamma, & \mathrm{rank}_{\mathcal{B}}(K_i(\theta)) > k, \end{cases} \qquad \mathcal{L}_{\mathcal{B}}(\theta) = \frac{1}{|\mathcal{B}|} \sum_{i \in \mathcal{B}} w_i \, \ell_{\mathrm{SDFT}}(\theta; i), \tag{21}$$

Thus, $\gamma = 0$ gives hard filtering, $\gamma \in (0, 1)$ softly down-weights the selected tail, and $\eta = 1$ recovers standard SDFT.

The batch-relative rule avoids imposing a single absolute KL threshold across tasks, models, and training stages, whose divergence scales can differ substantially. More importantly, it provides a controlled test of the heavy-tail hypothesis: supervision from the low-divergence majority is left unchanged, while only the contribution of the highest-divergence examples is reduced.

C Additional Motivation Results on InternLM2.5

We run the same two motivation studies from Sec. 3 on InternLM2.5-7B-Chat. Figure 5 reproduces the heavy-tailed teacher–policy divergence analysis, while Figure 6 reports the corresponding high-divergence down-weighting results. Together with the Qwen2.5 results in the main text, these experiments show that the motivating pattern is not specific to a single model family.

D Budget-to-Interpolation Maps and Adaptive Calibration

We give the exact and approximate maps from a trust-region budget C to the interpolation coefficient β for the geometric, arithmetic, and square-root teacher paths induced by forward KL, reverse KL,

and squared Hellinger distance, respectively. We use the main-text notation $p_\theta^t := \pi_\theta(\cdot \mid x, y_{<t})$ for the current student and $q_0^t := \pi_{\theta_0}(\cdot \mid x, c, y_{<t})$ for the demonstration-conditioned teacher. The coefficient β is the weight on the teacher (β=1 $\Rightarrow$ teacher, β=0 $\Rightarrow$ current-student anchor). All three constructions use the same divergence in the teacher-fidelity objective and the trust-region constraint.

D.1 General Trust-Region Characterization and Local Calibration

We prove Lemma 4.2, with q denoting the trust-region-constrained teacher distribution. The Lagrangian of Eq. (8) is

$$\mathcal{L}(q, \lambda, \nu) = \sum_v q_0^t(v) f\left(\frac{q(v)}{q_0^t(v)}\right) + \lambda \left[\sum_v p_\theta^t(v) f\left(\frac{q(v)}{p_\theta^t(v)}\right) - C\right] + \nu \left(\sum_v q(v) - 1\right), \quad (22)$$

where $\lambda \geq 0$ enforces the divergence budget and ν enforces normalization. Under the full-support and interior-solution conditions of the lemma, stationarity with respect to each $q(v)$ gives

$$f'\left(\frac{q^\star(v)}{q_0^t(v)}\right) + \lambda f'\left(\frac{q^\star(v)}{p_\theta^t(v)}\right) + \nu = 0. \quad (23)$$

Defining $\beta := 1/(1+\lambda)$ and $\eta := \nu/(1+\lambda)$ yields

$$\beta f'\left(\frac{q^\star(v)}{q_0^t(v)}\right) + (1-\beta) f'\left(\frac{q^\star(v)}{p_\theta^t(v)}\right) + \eta = 0. \quad (24)$$

For fixed $p_\theta^t(v)$, $q_0^t(v)$, and β, define

$$G_{f,\beta,v}(z) := \beta f'\left(\frac{z}{q_0^t(v)}\right) + (1-\beta) f'\left(\frac{z}{p_\theta^t(v)}\right).$$

Strict convexity makes $G_{f,\beta,v}$ strictly increasing in z. Hence the path has the implicit representation

$$q_{f,\beta}^\star(v) = G_{f,\beta,v}^{-1}(-\eta), \qquad \sum_v q_{f,\beta}^\star(v) = 1, \quad (25)$$

where the scalar η is fixed by normalization. Strict convexity of the fidelity objective also makes the optimizer unique. At $\beta = 1$, stationarity recovers $q_{f,1}^\star = q_0^t$; at the limiting endpoint $\beta = 0$, it recovers $q_{f,0}^\star = p_\theta^t$.

For $\beta \in (0, 1)$, the same distribution uniquely minimizes the scalarized objective

$$\beta\, \mathcal{D}_f(q \| q_0^t) + (1-\beta)\, \mathcal{D}_f(q \| p_\theta^t).$$

Comparing the optimality inequalities at any $0 < \beta_1 < \beta_2 < 1$ shows that

$$\mathcal{D}_f(q_{f,\beta_1}^\star \| p_\theta^t) \leq \mathcal{D}_f(q_{f,\beta_2}^\star \| p_\theta^t).$$

Thus the displacement from the student is continuous and nondecreasing along the path under the regularity conditions of the lemma.

Complementary slackness gives

$$\lambda \left[\mathcal{D}_f\left(q_{f,\beta}^\star \| p_\theta^t\right) - C\right] = 0. \quad (26)$$

If $\mathcal{D}_f(q_0^t \| p_\theta^t) \leq C$, the fidelity minimizer q_0^t is feasible, so $\lambda = 0$ and $\beta^\star = 1$. Otherwise the constraint is active, and the solution is the largest point along the divergence-induced teacher path that satisfies the budget, as described in Sec. 4.

Local expansion. Let $g_f^t(\beta) := \mathcal{D}_f(q_{f,\beta}^\star \| p_\theta^t)$. Since $q_{f,0}^\star = p_\theta^t$, we have $g_f^t(0) = 0$. The gradient of $\mathcal{D}_f(q \| p_\theta^t)$ at $q = p_\theta^t$ is constant across vocabulary entries, while the tangent to the probability simplex satisfies $\sum_v \frac{d}{d\beta} q_{f,\beta}^\star(v) = 0$; therefore $g_f^{t\prime}(0) = 0$. When the induced path is twice differentiable at the student endpoint, Taylor expansion gives

$$g_f^t(\beta) = \frac{1}{2} s_f^t \beta^2 + o(\beta^2), \qquad s_f^t := g_f^{t\prime\prime}(0), \quad (27)$$

which gives the inverse-square-root dependence used by the practical calibration in Eq. (10); its peer-relative online form is given in Eq. (37).

D.2 Forward KL-selection: geometric path

For forward KL, $f(u) = u \log u$ and $f'(u) = 1 + \log u$. Substituting this generator into Eq. (24) gives

$$\begin{aligned} 0 &= \beta\left[1 + \log \frac{q_\beta(v)}{q_0^t(v)}\right] + (1-\beta)\left[1 + \log \frac{q_\beta(v)}{p_\theta^t(v)}\right] + \eta \\ &= \log q_\beta(v) - \beta \log q_0^t(v) - (1-\beta)\log p_\theta^t(v) + 1 + \eta. \end{aligned} \tag{28}$$

Therefore

$$q_\beta(v) = e^{-1-\eta} p_\theta^t(v)^{1-\beta} q_0^t(v)^\beta.$$

Normalization sets $e^{-1-\eta} = Z_\beta^{-1}$, where $Z_\beta = \sum_v p_\theta^t(v)^{1-\beta} q_0^t(v)^\beta$, yielding the geometric path $q_\beta(v) = Z_\beta^{-1} p_\theta^t(v)^{1-\beta} q_0^t(v)^\beta$ reported in Table 1.

Exact map (Chernoff log-partition). With $r_t(v) := q_0^t(v)/p_\theta^t(v)$ and the cumulant generating function $\psi_t(\beta) := \log \sum_v p_\theta^t(v)\, r_t(v)^\beta = \log Z_\beta$, the target is $q_\beta(v) = p_\theta^t(v) r_t(v)^\beta / e^{\psi_t(\beta)}$ and the forward-KL budget has the closed form

$$D_{\mathrm{KL}}(q_\beta \| p_\theta^t) = \beta\psi_t'(\beta) - \psi_t(\beta), \qquad \psi_t(0) = 0,\ \ \psi_t'(1) = D_t,\ \ \psi_t''(0) = \mathrm{Var}_{p_\theta^t}\big(\log r_t\big), \tag{29}$$

increasing monotonically from 0 (β=0) to $D_t := D_{\mathrm{KL}}(q_0^t \| p_\theta^t)$ (β=1). The exact choice is the largest admissible tilt $\beta_C = \max\{\beta \in [0,1] : \beta\psi_t'(\beta) - \psi_t(\beta) \le C\}$. This needs ψ_t over $[0,1]$, i.e. the full token distribution, not merely the scalar $D_t = \psi_t'(1)$.

Local curvature near the anchor. Let $g_t(\beta) := \beta\psi_t'(\beta) - \psi_t(\beta)$. Direct differentiation gives

$$g_t'(\beta) = \beta\psi_t''(\beta), \qquad g_t''(\beta) = \psi_t''(\beta) + \beta\psi_t'''(\beta), \qquad g_t''(0) = \psi_t''(0).$$

Thus the $\psi_t'''(0)$ term does not contribute to the local curvature because it is multiplied by $\beta = 0$. Expanding (29) at β=0 (the endpoint $q_\beta = p_\theta^t$) gives

$$D_{\mathrm{KL}}(q_\beta \| p_\theta^t) = \tfrac{1}{2} V_t\, \beta^2 + O(\beta^3), \qquad V_t := \psi_t''(0) = \mathrm{Var}_{p_\theta^t}\big(\log(q_0^t/p_\theta^t)\big). \tag{30}$$

The coefficient $V_t = \psi_t''(0)$ is the exact local curvature. Retaining V_i per sample and imposing $\frac{1}{2} V_i \beta^2 \le C$ gives $\beta \le \sqrt{2C/V_i}$.

D.3 Reverse KL-selection: arithmetic path

For reverse KL, $f(u) = -\log u$ and $f'(u) = -1/u$. Substituting this generator into Eq. (24) gives

$$0 = -\beta \frac{q_0^t(v)}{q_\beta(v)} - (1-\beta)\frac{p_\theta^t(v)}{q_\beta(v)} + \eta, \qquad \text{so} \qquad q_\beta(v) = \frac{\beta q_0^t(v) + (1-\beta) p_\theta^t(v)}{\eta}.$$

Because both p_θ^t and q_0^t are normalized, imposing $\sum_v q_\beta(v) = 1$ sets $\eta = 1$. Thus the reverse-KL problem $\min_q D_{\mathrm{KL}}(q_0^t \| q)$ subject to $D_{\mathrm{KL}}(p_\theta^t \| q) \le C$ yields the arithmetic path

$$q_\beta(v) = \beta\, q_0^t(v) + (1-\beta)\, p_\theta^t(v), \qquad \beta = 1/(1+\lambda). \tag{31}$$

Let $d_t(v) := q_0^t(v) - p_\theta^t(v)$. The constraint along this path is

$$G_t(\beta) := D_{\mathrm{KL}}(p_\theta^t \| q_\beta) = \sum_v p_\theta^t(v) \log \frac{p_\theta^t(v)}{p_\theta^t(v) + \beta d_t(v)}.$$

The exact coefficient is the largest $\beta \in [0,1]$ satisfying $G_t(\beta) \le C$.

Local curvature near the anchor. Differentiating the constraint gives

$$G_t'(\beta) = -\sum_v \frac{p_\theta^t(v) d_t(v)}{p_\theta^t(v) + \beta d_t(v)}, \qquad G_t''(\beta) = \sum_v \frac{p_\theta^t(v) d_t(v)^2}{\big(p_\theta^t(v) + \beta d_t(v)\big)^2}.$$

Because $\sum_v d_t(v) = 0$, we have $G_t(0) = G_t'(0) = 0$ and

$$G_t''(0) = \sum_v \frac{d_t(v)^2}{p_\theta^t(v)} =: \chi_t^2.$$

Taylor expansion at $\beta = 0$ therefore gives

$$D_{\mathrm{KL}}(p_\theta^t \| q_\beta) = \tfrac{1}{2}\,\chi_t^2\,\beta^2 + O(\beta^3), \qquad \chi_t^2 = \mathbb{E}_{p_\theta^t}\big[(q_0^t/p_\theta^t - 1)^2\big] \geq 0. \tag{32}$$

Imposing $\frac{1}{2}\chi_t^2\beta^2 \leq C$ gives $\beta \leq \sqrt{2C/\chi_t^2}$. Unlike the forward-KL construction, the reverse-KL rule retains χ_t^2 rather than introducing a further KL-based approximation.

Stable peer calibration via relative Pearson. Although the Pearson coefficient in Eq. (32) is the exact local curvature, its density ratio q_0^t/p_θ^t is unbounded and can be dominated by tokens to which the student assigns vanishing probability. Following relative density-ratio estimation (Yamada et al., 2011), we therefore replace the peer statistic used for adaptive calibration with the direction-matched relative-Pearson surrogate

$$m_\zeta^t(v) := \zeta q_0^t(v) + (1-\zeta)p_\theta^t(v), \qquad \widetilde{\chi}^2_{\zeta,t} := \mathbb{E}_{p_\theta^t}\left[\left(\frac{q_0^t}{m_\zeta^t} - 1\right)^2\right], \qquad \zeta \in (0,1). \tag{33}$$

The relative ratio is bounded by $1/\zeta$, and $\widetilde{\chi}^2_{\zeta,t} \to \chi_t^2$ as $\zeta \to 0$. This substitution stabilizes peer-relative calibration while leaving the reverse-KL teacher path and distillation objective unchanged. The local identity in Eq. (32) remains governed by the bare Pearson curvature, while the bounded surrogate enters only the adaptive calibration.

D.4 Squared-Hellinger selection: square-root path

For

$$\mathcal{D}_f(q\|r) = H^2(q,r) := \frac{1}{2}\sum_v \big(\sqrt{q(v)} - \sqrt{r(v)}\big)^2, \qquad f(u) = \frac{1}{2}(\sqrt{u}-1)^2,$$

we have $f'(u) = \frac{1}{2}(1 - u^{-1/2})$. Substitution into Eq. (24) yields

$$\begin{aligned} 0 &= \frac{\beta}{2}\left(1 - \sqrt{\frac{q_0^t(v)}{q_\beta(v)}}\right) + \frac{1-\beta}{2}\left(1 - \sqrt{\frac{p_\theta^t(v)}{q_\beta(v)}}\right) + \eta \\ &= \frac{1}{2} + \eta - \frac{\beta\sqrt{q_0^t(v)} + (1-\beta)\sqrt{p_\theta^t(v)}}{2\sqrt{q_\beta(v)}}. \end{aligned} \tag{34}$$

Hence

$$\sqrt{q_\beta(v)} = \frac{(1-\beta)\sqrt{p_\theta^t(v)} + \beta\sqrt{q_0^t(v)}}{1+2\eta}.$$

Writing $u_\beta^t(v) := (1-\beta)\sqrt{p_\theta^t(v)} + \beta\sqrt{q_0^t(v)}$, normalization gives $(1+2\eta)^2 = \sum_w u_\beta^t(w)^2$. Therefore the constrained problem $\min_q H^2(q, q_0^t)$ subject to $H^2(q, p_\theta^t) \leq C$ has the normalized square-root path

$$q_\beta(v) = \frac{u_\beta^t(v)^2}{\sum_w u_\beta^t(w)^2}. \tag{35}$$

Thus, squared Hellinger interpolates arithmetically in square-root probability space, between the log-linear forward-KL path and the probability-space reverse-KL path. The exact coefficient is the largest $\beta \in [0,1]$ satisfying $H^2(q_\beta, p_\theta^t) \leq C$.

Local curvature near the anchor. Let $\Delta_t(v) := \sqrt{q_0^t(v)} - \sqrt{p_\theta^t(v)}$ and $c_t := \sum_v \sqrt{p_\theta^t(v)}\,\Delta_t(v)$, and $b_t := \sum_v \Delta_t(v)^2$. Since $\sum_v u_\beta^t(v)^2 = 1 + 2c_t\beta + b_t\beta^2$, the constraint along the normalized path has the exact form

$$J_t(\beta) := H^2(q_\beta, p_\theta^t) = 1 - \frac{1 + c_t\beta}{\sqrt{1 + 2c_t\beta + b_t\beta^2}}.$$

Direct differentiation at the anchor gives

$$J_t'(\beta) = \frac{(b_t - c_t^2)\beta}{(1 + 2c_t\beta + b_t\beta^2)^{3/2}}, \qquad J_t''(0) = b_t - c_t^2 =: S_H^t = 1 - \left(\sum_v \sqrt{p_\theta^t(v) q_0^t(v)}\right)^2.$$

In particular, $J_t(0) = J_t'(0) = 0$. Taylor expansion at $\beta = 0$ therefore gives

$$H^2(q_\beta, p_\theta^t) = \frac{1}{2} S_H^t \beta^2 + O(\beta^3), \tag{36}$$
$$S_H^t = \sum_v \left(\Delta_t(v) - c_t \sqrt{p_\theta^t(v)}\right)^2 \geq 0.$$

The Hellinger curvature S_H^t therefore occupies the same role as V_t on the geometric path and χ_t^2 on the arithmetic path. Imposing the local budget yields $\beta \leq \sqrt{2C/S_H^t}$.

D.5 Online Peer-Relative Calibration

All three paths share the same calibration template. At each update, every sample i in the current rollout batch contributes a token-averaged peer statistic s_i. The exact local curvatures are $s{=}V$ for forward KL, $s{=}\chi^2$ for reverse KL, and $s{=}S_H$ for squared Hellinger distance. For reverse KL, the unstable bare χ^2 is replaced in the peer-relative rule by $\widetilde{\chi}_\zeta^2$ from Eq. (33).

Absolute budget:

$$\beta_i = \mathrm{clip}\Big(\min\Big\{1, \sqrt{2C/s_i}\Big\}, \beta_{\min}, \beta_{\max}\Big).$$

Peer-relative budget: replace C by a peer-scaled budget $C_{\mathcal{B}} := \rho \cdot \frac{1}{2}\bar{s}$ with $\bar{s} = \frac{1}{B}\sum_j s_j$. Then

$$\beta_i = \mathrm{clip}\Big(\min\Big\{1, \sqrt{\rho\,\bar{s}/s_i}\Big\}, \beta_{\min}, \beta_{\max}\Big). \tag{37}$$

The scale ρ controls overall teacher retention, while the ratio $\bar{s}/s_i$ adapts the interpolation to each sample. Samples with large disagreement relative to their peers receive a smaller teacher weight, and the rule is invariant to a common rescaling of s. The three teacher geometries share this calibration rule while differing in both their interpolation paths and their local peer statistics.

Geometry and calibration. Forward KL follows a geometric product path, reverse KL an arithmetic mixture, and squared Hellinger an arithmetic path in square-root probability space. Each path is paired with its corresponding local curvature. The relative-Pearson substitution stabilizes reverse-KL calibration without changing its TRSD teacher distribution. Clipping β_i to $[\beta_{\min}, \beta_{\max}]$ avoids degenerate targets, while token averaging provides a stable sample-level statistic. An absolute budget fixes the local divergence scale, whereas peer-relative calibration adapts this scale to the current teacher–student disagreement.

Geometry	Effective target q_β	Calibration statistic s
Forward KL	$\propto (p_\theta^t)^{1-\beta}(q_0^t)^\beta$	$V_t = \mathrm{Var}_{p_\theta^t}(\log(q_0^t/p_\theta^t))$
Reverse KL	$(1-\beta)p_\theta^t + \beta q_0^t$	$\widetilde{\chi}_{\zeta,t}^2$ in practice
Squared Hellinger	$\big((1-\beta)\sqrt{p_\theta^t} + \beta\sqrt{q_0^t}\big)^2/Z_\beta$	S_H^t

E Additional Experimental Details

E.1 Training Procedure

SDFT (Shenfeld et al., 2026) and TRSD share the on-policy loop in Algorithm 1. The student samples a completion from the demonstration-free prompt. The reference model then evaluates the demonstration-conditioned teacher on that same completion. As described in Sec. 4, TRSD replaces this teacher with a trust-region-constrained distribution before the student update. The choice $\beta_i = 1$ leaves the teacher unchanged and recovers SDFT. Prompt templates are given in Appendix E.5.

Algorithm 1 Trust-region constrained self-distillation. SDFT is the special case $\beta_i = 1$, for which no trust-region target is constructed.

Require: Demonstration dataset $\mathcal{D} = \{(x, c)\}$
Require: Student π_θ initialized at $\theta = \theta_0$; EMA rate α
Require: Budget scale ρ and clip bounds $[\beta_{\min}, \beta_{\max}]$
$\phi \leftarrow \theta_0$
for each training step **do**
  Sample a minibatch $\mathcal{B} \subset \mathcal{D}$
  for each $(x, c) \in \mathcal{B}$ in parallel **do**
    Sample $y \sim \pi_\theta(\cdot \mid x)$
    for each position $t = 1, \ldots, |y|$ **do**
      $p_\theta^t \leftarrow \pi_\theta(\cdot \mid x, y_{<t})$
      $q_0^t \leftarrow \pi_\phi(\cdot \mid x, c, y_{<t})$
    end for
    Average the local curvature over t to obtain s_i
    Calibrate β_i with Eq. (37)
    Construct $q^\star_{f,\beta_i,t}$ from p_θ^t and q_0^t at each t
  end for
  Update θ by minimizing Eq. (11)
  $\phi \leftarrow \alpha\theta + (1 - \alpha)\phi$
end for

Reference-model EMA. The reference weights ϕ are an exponential moving average of the student, not a separately trained teacher. After every student update,

$$\phi \leftarrow \alpha\theta + (1 - \alpha)\phi.$$

SDFT (Shenfeld et al., 2026) compares three teacher choices. A frozen base model does not track learning progress, while using the current student directly destabilizes on-policy training. An EMA of the student parameters tracks the student while smoothing these updates. We therefore use this EMA for ϕ in Algorithm 1.

E.2 IFEval Metric Breakdown

The main single-task table reports IFEval as the unweighted mean of four standard metrics: prompt- and instruction-level accuracy under strict and loose evaluation (P-S, I-S, P-L, I-L). Tables 3 and 4 list these constituents for the Qwen2.5-7B-Instruct and InternLM2.5-7B-Chat runs in Table 2.

E.3 InternLM2.5 Single-Task Selections

Table 5 lists the InternLM2.5-7B-Chat single-task configurations. The matched SDFT baseline is listed first. The main single-task table uses T1, S1, and M1. We also report T2, S2, and M2 as additional configurations to assess robustness.

E.4 Training Configurations

Table 6 summarizes the common training configuration and the method-specific search spaces used in our experiments.

E.5 Student and Teacher Prompt Construction

SDFT and TRSD use the same prompt construction; they differ only in the distribution used as the effective distillation target. For each training example, the student receives the original dataset prompt without access to the expert demonstration. Abstracting away the model-specific chat-template tokens, its user content is **Student prompt**

```
[ORIGINAL USER REQUEST]
```

Table 3: IFEval constituents for the Qwen2.5-7B-Instruct single-task runs in Table 2. Mean is the unweighted average of P-S, I-S, P-L, and I-L.

Geometry	Method	P-S	I-S	P-L	I-L	Mean
Tool Use						
	Base	57.30	68.35	63.96	73.50	65.78
	SFT	24.77	39.45	27.73	41.97	33.48
	DFT	35.67	49.16	41.96	54.56	45.34
Forward KL	SDFT	34.20	47.72	40.67	53.96	44.14
	TRSD	48.61	59.35	56.75	66.55	57.81
Reverse KL	SDFT	33.46	48.08	39.93	53.72	43.80
	TRSD	47.13	59.47	52.31	63.55	55.62
Squared Hellinger	SDFT	44.73	56.95	49.35	61.51	53.14
	TRSD	49.72	61.75	54.90	65.95	58.08
Science Q&A						
	Base	57.30	68.35	63.96	73.50	65.78
	SFT	35.86	49.52	39.74	52.76	44.47
	DFT	39.37	50.96	48.61	59.95	49.72
Forward KL	SDFT	36.97	49.64	41.96	53.84	45.60
	TRSD	41.77	54.68	52.13	63.67	53.06
Reverse KL	SDFT	41.04	54.08	46.03	58.15	49.82
	TRSD	46.77	58.51	53.60	64.63	55.88
Squared Hellinger	SDFT	40.85	53.60	46.03	58.39	49.72
	TRSD	43.62	55.64	51.76	62.59	53.40
Medical						
	Base	57.30	68.35	63.96	73.50	65.78
	SFT	34.75	48.56	39.74	52.40	43.86
	DFT	37.89	53.00	44.18	58.03	48.28
Forward KL	SDFT	38.45	51.80	46.21	57.79	48.56
	TRSD	51.02	61.75	60.63	70.26	60.91
Reverse KL	SDFT	38.63	51.68	46.40	59.35	49.01
	TRSD	46.21	58.39	51.39	62.95	54.73
Squared Hellinger	SDFT	46.77	58.51	53.23	64.39	55.73
	TRSD	52.87	63.67	62.85	71.70	62.77

The student completion is sampled from this prompt and then held fixed when evaluating the token-level student and teacher distributions.

The teacher is conditioned on the same request together with the corresponding expert response:
Demonstration-conditioned teacher prompt

```
[ORIGINAL USER REQUEST]

This is an example for a response to the question:
[EXPERT DEMONSTRATION]

Now answer with a response of your own, including the thinking process.
```

Thus, the demonstration changes the teacher's distribution along the student-generated completion, but is never included in the student's input. Any leading system or task-context messages supplied by the dataset are preserved when constructing the corresponding chat conversations. The domain-specific source fields are summarized in Table 7.

E.6 Medical Data Construction

The medical domain draws its training pool from the English split of `medical-o1-reasoning-SFT`, the stage-1 reasoning corpus released with HuatuoGPT-o1 (Chen et al., 2024). Evaluation questions are drawn from `medical-o1-verifiable-problem`. We first construct a fixed holdout of 1,000 unique evaluation questions using random seed 42. To reduce train–test contamination, we remove

Table 4: IFEval constituents for the InternLM2.5-7B-Chat single-task runs in Table 2. Mean is the unweighted average of P-S, I-S, P-L, and I-L.

Geometry	Method	P-S	I-S	P-L	I-L	Mean
Tool Use						
	Base	50.46	60.79	54.71	64.27	57.56
	SFT	24.40	37.41	26.25	40.05	32.03
	DFT	22.92	36.45	27.17	40.29	31.71
Forward KL	SDFT	38.82	50.60	42.51	54.44	46.59
	TRSD	42.33	54.80	46.21	57.67	50.25
Reverse KL	SDFT	38.08	50.72	42.70	54.08	46.39
	TRSD	48.06	58.27	51.57	61.39	54.82
Squared Hellinger	SDFT	41.96	53.60	46.21	57.43	49.80
	TRSD	46.58	56.71	50.28	60.19	53.44
Science Q&A						
	Base	50.46	60.79	54.71	64.27	57.56
	SFT	36.78	48.68	43.99	55.04	46.12
	DFT	32.35	44.72	38.45	50.48	41.50
Forward KL	SDFT	38.08	50.24	41.77	53.72	45.95
	TRSD	40.11	51.56	43.62	55.64	47.73
Reverse KL	SDFT	39.19	50.24	44.73	54.92	47.27
	TRSD	43.62	54.32	48.06	58.87	51.22
Squared Hellinger	SDFT	38.08	49.52	42.88	54.08	46.14
	TRSD	40.67	51.80	45.10	55.76	48.33
Medical						
	Base	50.46	60.79	54.71	64.27	57.56
	SFT	34.57	48.44	41.59	55.28	44.97
	DFT	39.19	51.92	45.66	57.79	48.64
Forward KL	SDFT	39.00	51.32	42.14	53.84	46.58
	TRSD	43.62	55.88	47.87	59.23	51.65
Reverse KL	SDFT	38.45	50.96	43.99	55.76	47.29
	TRSD	43.25	54.80	47.50	58.99	51.14
Squared Hellinger	SDFT	37.34	49.76	42.70	54.56	46.09
	TRSD	42.88	55.40	46.40	58.99	50.92

any training question that shares a contiguous sequence of at least 64 characters with a holdout question. This removes 745 of the 19,704 training examples, leaving 18,959 eligible examples.

From this decontaminated pool, we sample 12,288 examples without replacement. Sampling is stratified by quartiles of total sequence length, measured over the formatted prompt and response with the Qwen2.5-7B-Instruct tokenizer. The allocation to each quartile is proportional to its prevalence in the eligible pool, preserving the original length distribution while fixing the training-set scale. The same evaluation holdout is used across all medical experiments.

E.7 Comparisons with SFT and DFT

The main-text Figure 3 focuses on comparisons between each TRSD instantiation and its matched SDFT objective. Figure 7 provides the complementary comparison with SFT and DFT by reporting the margin of the across-geometry TRSD lower envelope over DFT and SFT. Figure 8 further disaggregates these results by divergence geometry. We use the selected continual trajectories $T1{\to}S2{\to}M1$ for forward KL, $T2{\to}S1{\to}M2$ for reverse KL, and $T1{\to}S2{\to}M2$ for squared Hellinger.

The improvement is not driven by a single divergence geometry. Each TRSD instantiation remains ahead of both SFT and DFT on every reported stage–capability pair. The margins are especially pronounced for skills learned earlier in the sequence, showing that the TRSD teacher distribution supports the accumulation of new capabilities while limiting their erosion during subsequent adaptation.

Table 5: InternLM2.5-7B-Chat single-task configurations. The matched SDFT baseline is listed first. The main single-task table uses T1, S1, and M1. T2, S2, and M2 are additional configurations included to assess robustness. Baseline rows report absolute percentages; TRSD rows report percentage-point differences from the corresponding baseline.

Geometry	Task	ID	ρ	$[\beta_{\min}, \beta_{\max}]$	Δ Acc.	Δ IFEval
Forward KL	Tool	B	–	–	48.45	46.59
		T1	0.25	$[0.10, 0.80]$	+1.03	+3.66
		T2	0.05	$[0.05, 1.00]$	+0.00	+3.67
	Science	B	–	–	50.69	45.95
		S1	0.20	$[0.10, 0.80]$	+6.11	+1.78
		S2	0.20	$[0.05, 0.85]$	+3.35	+1.57
	Medical	B	–	–	30.30	46.58
		M1	0.05	$[0.05, 1.00]$	+0.60	+5.07
		M2	0.05	$[0.10, 1.00]$	-2.40	+5.15
Reverse KL	Tool	B	–	–	39.18	46.39
		T1	0.05	$[0.05, 0.80]$	+8.24	+8.43
		T2	0.05	$[0.00, 0.75]$	+6.18	+6.98
	Science	B	–	–	53.85	47.27
		S1	0.30	$[0.05, 0.75]$	+3.15	+3.95
		S2	0.30	$[0.05, 0.80]$	+3.94	+2.98
	Medical	B	–	–	30.20	47.29
		M1	0.20	$[0.10, 0.80]$	-1.30	+3.85
		M2	0.20	$[0.10, 0.85]$	-1.30	+3.85
Squared Hellinger	Tool	B	–	–	43.30	49.80
		T1	0.05	$[0.05, 1.00]$	+4.12	+3.64
		T2	0.05	$[0.20, 0.75]$	+2.06	+4.79
	Science	B	–	–	48.72	46.14
		S1	0.10	$[0.10, 0.80]$	+5.13	+2.19
		S2	0.25	$[0.05, 0.75]$	+4.93	+1.15
	Medical	B	–	–	30.70	46.09
		M1	0.20	$[0.10, 0.80]$	+0.20	+4.83
		M2	0.10	$[0.10, 0.80]$	+0.40	+3.65

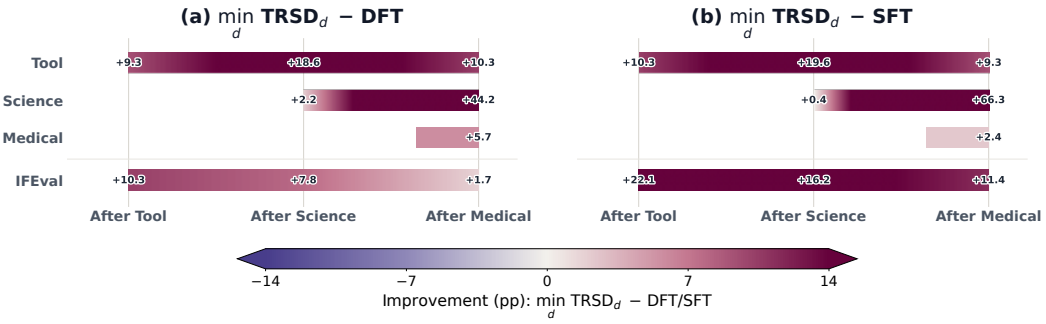

Figure 7: Conservative comparisons with SFT and DFT under sequential Tool Use → Science Q&A → Medical adaptation. Panels (a) and (b) report the percentage-point margins of the across-geometry TRSD lower envelope over DFT and SFT, respectively. At each stage, we show the skills introduced up to that point together with IFEval retention. Positive values indicate that every TRSD divergence instantiation outperforms the corresponding fine-tuning baseline.

E.7.1 Continual Baseline Trajectories

Table 8 reports the absolute stage-wise results of SFT, DFT, and the self-distillation baselines used in Figures 3 and 8. For SDFT, we retain a separate trajectory for each divergence geometry to match the corresponding TRSD instantiation.

Table 6: Training configurations for all compared methods. Effective batch size includes gradient accumulation. Prompt and completion limits are measured in tokens. The reference-model EMA rate α is the update in Algorithm 1.

Configuration	SFT	DFT	SDFT	TRSD
Base model	Qwen2.5-7B-Instruct / InternLM2.5-7B-Chat			
Learning rate	5×10^{-5}			
Optimizer	AdamW			
LR schedule	Cosine with 10% warmup			
Training epochs	$\{2, 4, 6\}$	$\{2, 4, 6\}$	$\{2, 3\}$	$\{2, 3\}$
Effective batch size	$\{16, 32, 64\}$			
Maximum gradient norm	1			
Precision	bfloat16			
Weight decay	0			
Maximum prompt length	1024			
Maximum completion length	1024			
Divergence geometry	–	–	Forward KL Reverse KL Squared Hellinger	Forward KL Reverse KL Squared Hellinger
Reference-model EMA α	–	–	0.01	0.01
Budget scale ρ	–	–	–	$\{0.05, 0.10, 0.15, 0.20, 0.25\}$
$\beta_{\min}$	–	–	–	$\{0, 0.05, 0.10, 0.15, 0.20\}$
$\beta_{\max}$	–	–	–	$\{0.75, 0.80, 0.85, 1.00\}$

Table 7: Dataset fields used to construct the student prompt and the demonstration-conditioned teacher prompt.

Domain	Original prompt	Expert demonstration
Tool Use	`prompt`	`golden_response`
Science Q&A	`messages`	`output_text`
Medical	`messages`	`output_text`

E.7.2 Failure Modes after Sequential SFT and DFT

The large Science Q&A decline of SFT and DFT after the Medical stage reflects a marked change in response behavior, rather than a failure to acquire Science Q&A in the preceding stage. The evaluator requires the final multiple-choice decision to be emitted as a single letter inside the prescribed answer wrapper. As shown in Table 9, both SFT and DFT satisfy this interface on nearly all examples immediately after Science training. Subsequent Medical adaptation sharply reduces this compliance: SFT emits a valid wrapper on only 11 of 507 examples, while DFT often places an entire explanation, rather than a single choice, inside it.

The representative outputs in Table 10 reveal two complementary failure modes. In one SFT case, the model continues to identify option A after Medical training but terminates with only a closing answer tag, making an otherwise correct decision unrecoverable by the task interface. On the same question, DFT places a paragraph-long justification inside the answer field instead of the required letter. Other DFT examples exhibit a genuine decision change: one molecular weight question is answered correctly as C after Science training, but the Medical-stage model miscounts the atoms, selects D, and attributes the disagreement to an error in the provided options. The observed collapse therefore combines loss of output-protocol adherence with degradation in task-specific decision making. Both are consequential forms of functional forgetting in a continually adapted model.

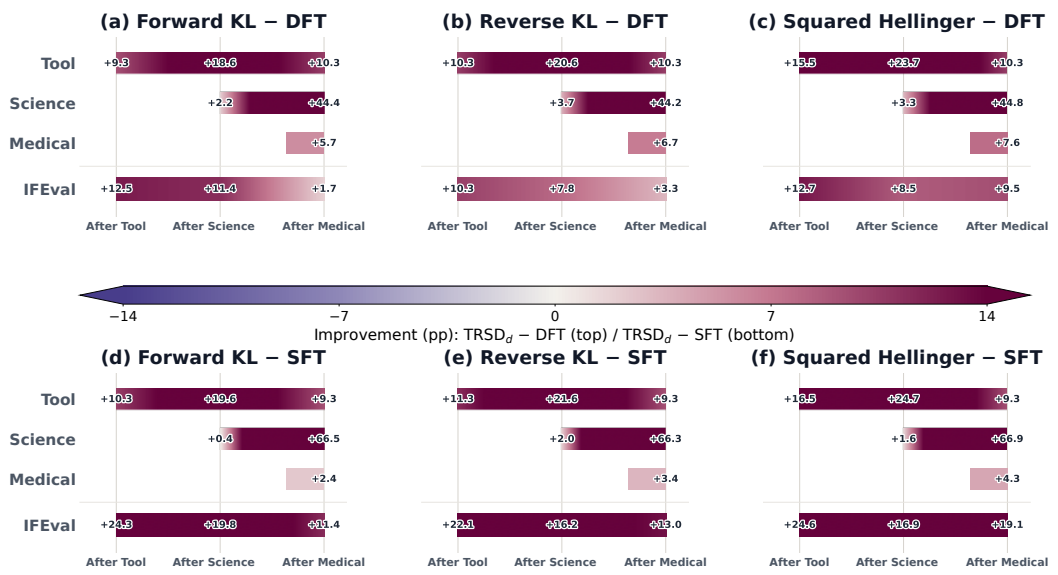

Figure 8: Divergence-specific comparisons with SFT and DFT under sequential Tool Use → Science Q&A → Medical adaptation. The top row reports the percentage-point margin of each TRSD instantiation over DFT, and the bottom row reports the corresponding margin over SFT. At each stage, we show the skills introduced up to that point together with IFEval retention. Positive values indicate higher performance under TRSD.

Table 8: Stage-wise continual-learning results for the selected baseline trajectories (%). A dash indicates that the corresponding task has not yet been introduced.

Stage	Method	Tool	Science	Medical	IFEval
After Tool	SFT	55.67	–	–	33.48
	DFT	56.70	–	–	45.34
	SDFT (Forward KL)	63.92	–	–	44.14
	SDFT (Reverse KL)	64.95	–	–	43.80
	SDFT (Squared Hellinger)	62.89	–	–	53.14
After Science	SFT	43.30	66.86	–	28.08
	DFT	44.30	65.09	–	36.50
	SDFT (Forward KL)	60.82	67.26	–	38.34
	SDFT (Reverse KL)	61.86	66.47	–	36.63
	SDFT (Squared Hellinger)	60.82	68.44	–	42.23
After Medical	SFT	48.45	0.99	34.80	28.24
	DFT	47.40	23.08	31.50	37.90
	SDFT (Forward KL)	47.42	66.27	37.70	35.91
	SDFT (Reverse KL)	57.73	63.51	34.60	35.79
	SDFT (Squared Hellinger)	51.55	63.91	39.80	38.12

E.7.3 Cross-Task Grid Configurations

For each divergence geometry, we retain two configurations at every stage and evaluate their Cartesian product under the Tool Use → Science Q&A → Medical sequence. Table 11 lists the peer-relative budget scale and interpolation clipping interval used by each configuration. Here T, S, and M denote the Tool Use, Science Q&A, and Medical stages, respectively.

Tables 12–14 report the stage-wise results for every configuration in the three $2\times2\times2$ grids. Each table expands from two Tool configurations to four Tool–Science combinations and then to all eight Tool–Science–Medical trajectories. Each stage block begins with the matched SDFT baseline in absolute percentage, followed by the TRSD differences from that baseline in percentage points. Bold

Table 9: Science Q&A behavior before and after the Medical stage. "Wrapped" counts responses containing the prescribed answer wrapper; "letter-only" further requires its content to be a single option letter.

Method	Checkpoint	Accuracy	Wrapped	Letter-only
SFT	After Science	66.86	502/507	502/507
SFT	After Medical	0.99	11/507	11/507
DFT	After Science	65.09	504/507	503/507
DFT	After Medical	23.08	360/507	220/507

Table 10: Representative Science Q&A outputs before and after subsequent Medical adaptation. Each row compares the same evaluation example across checkpoints; ellipses omit the preceding reasoning while preserving the model's final decision and answer-field content.

Method and case	After Science	After Medical
SFT, reactant synthesis (index 0; gold A)	`... Therefore, the correct answer is option A. <answer>A</answer>`	`... Therefore, Option A is the most likely candidate. </answer>`
DFT, reactant synthesis (index 0; gold A)	`... Option A is the only one that matches the structure of the product. <answer>A</answer>`	`<answer>The correct reactant ... is Option A. This option has a structure that matches the product structure ...</answer>`
DFT, molecular weight (index 1; gold C)	`... The closest option to this calculated value is 337.310 g/mol. <answer>C</answer>`	`... the correct answer is: D: 101.150 g/mol This discrepancy may be due to an error in the options ...`

rows are the trajectories reported in the main text. The remaining rows are additional configurations included to assess robustness across the grid.

E.8 InternLM2.5 Sequential Adaptation

Figure 9 reports the selected InternLM2.5-7B-Chat trajectory for each divergence relative to its matched SDFT objective.

For each divergence, we retain two Tool and two Science configurations from Table 5, evaluate all four Tool–Science combinations, and continue two selected Tool–Science parents with M1 and M2. Thus, each table contains two Tool configurations, four Tool–Science configurations, and four Tool–Science–Medical trajectories. The first row of each stage gives the matched SDFT baseline in absolute percentage; all TRSD rows report percentage-point differences from that baseline. Bold rows form the trajectory shown in Figure 9.

E.9 Complete Single-Task Hyperparameter Grids

Tables 18–26 report all 311 Qwen2.5-7B-Instruct configurations used in the single-task robustness analysis. For every divergence geometry and target domain, we sweep two peer-relative budget scales ρ together with the corresponding clipping grid for $[\beta_{\min}, \beta_{\max}]$. Each table first reports the matched SDFT baseline in absolute percentage, followed by the task-accuracy and aggregate IFEval differences of every TRSD configuration from that baseline in percentage points; no result is removed on the basis of either metric.

F Expanded Related Work

Continual adaptation and forgetting. Continual learning studies how a model can acquire new knowledge and skills without erasing capabilities learned previously (De Lange et al., 2021). Classical

Table 11: Configurations retained for the cross-task grids. The budget scale is denoted by ρ in the main text, and the clipping interval specifies $[\beta_{\min}, \beta_{\max}]$.

Geometry	Stage	ID	ρ	$[\beta_{\min}, \beta_{\max}]$
Forward KL	Tool	T1	0.05	$[0.05, 1.00]$
	Tool	T2	0.05	$[0.05, 0.85]$
	Science	S1	0.10	$[0.15, 0.80]$
	Science	S2	0.05	$[0.20, 0.80]$
	Medical	M1	0.10	$[0.10, 0.85]$
	Medical	M2	0.05	$[0.15, 0.80]$
Reverse KL	Tool	T1	0.05	$[0.00, 0.75]$
	Tool	T2	0.05	$[0.05, 0.80]$
	Science	S1	0.20	$[0.20, 0.85]$
	Science	S2	0.20	$[0.20, 1.00]$
	Medical	M1	0.05	$[0.05, 1.00]$
	Medical	M2	0.20	$[0.10, 1.00]$
Squared Hellinger	Tool	T1	0.15	$[0.00, 0.85]$
	Tool	T2	0.05	$[0.20, 0.75]$
	Science	S1	0.25	$[0.10, 0.80]$
	Science	S2	0.25	$[0.05, 1.00]$
	Medical	M1	0.10	$[0.15, 0.85]$
	Medical	M2	0.20	$[0.05, 0.80]$

Table 12: Stage-wise Forward-KL TRSD cross-task grid. The first row of each stage reports the matched SDFT baseline in percent; subsequent rows report TRSD differences from this baseline in percentage points.

Stage	Configuration	Δ Tool	Δ Science	Δ Medical	Δ IFEval
After Tool	SDFT baseline	63.92	–	–	44.14
	T1	**+2.06**	–	–	**+13.67**
	T2	+6.18	–	–	+5.79
After Science	SDFT baseline	60.82	67.26	–	38.34
	T1×S1	-2.06	+1.77	–	+6.49
	T1×S2	**+2.07**	**+0.00**	–	**+9.53**
	T2×S1	-1.03	+1.77	–	+2.14
	T2×S2	+3.10	-0.59	–	+0.95
After Medical	SDFT baseline	47.42	66.27	37.70	35.91
	T1×S1×M1	+9.28	+0.59	-0.70	+3.58
	T1×S1×M2	+9.28	+2.76	+1.30	+4.42
	T1×S2×M1	**+10.31**	**+1.19**	**-0.50**	**+3.73**
	T1×S2×M2	+8.25	+0.40	-0.30	+4.02
	T2×S1×M1	+12.37	+0.79	+1.50	+2.22
	T2×S1×M2	+9.28	+0.40	+0.10	+0.59
	T2×S2×M1	+15.47	+0.20	+0.50	+0.17
	T2×S2×M2	+15.47	-2.17	-0.40	+1.51

approaches mitigate forgetting through mechanisms such as parameter regularization and knowledge preservation (Kirkpatrick et al., 2017; Li & Hoiem, 2018). The challenge remains pronounced in language models, where sequential adaptation can degrade both previously learned tasks and broad instruction-following capabilities (Wang et al., 2023). Recent studies further show that reinforcement fine-tuning and on-policy data can preserve prior behavior more effectively than conventional supervised adaptation (Lai et al., 2026; Shenfeld et al., 2025; Chen et al., 2025). Complementary recent work introduces on-policy replay as a data-level mechanism for preserving historical behavior during continual supervised fine-tuning (Chen et al., 2026). Our work instead examines a distinct source of forgetting within on-policy self-distillation: the sample-wise divergence induced by the demonstration-conditioned teacher.

Table 13: Stage-wise Reverse-KL TRSD cross-task grid. The first row of each stage reports the matched SDFT baseline in percent; subsequent rows report TRSD differences from this baseline in percentage points.

Stage	Configuration	Δ Tool	Δ Science	Δ Medical	Δ IFEval
After Tool	SDFT baseline	64.95	–	–	43.80
	T1	+5.15	–	–	+8.87
	T2	**+2.06**	–	–	**+11.82**
After Science	SDFT baseline	61.86	66.47	–	36.63
	T1×S1	+6.18	+2.76	–	+8.18
	T1×S2	+6.18	+2.37	–	+9.77
	T2×S1	**+3.09**	**+2.37**	–	**+7.63**
	T2×S2	+2.06	+0.20	–	+8.96
After Medical	SDFT baseline	57.73	63.51	34.60	35.79
	T1×S1×M1	+0.00	+1.18	+3.60	+3.63
	T1×S1×M2	+0.00	+1.97	+4.10	+4.61
	T1×S2×M1	+1.03	-0.79	+3.40	+4.03
	T1×S2×M2	+1.03	+2.76	+2.10	+5.93
	T2×S1×M1	+0.00	+5.13	+4.50	+2.98
	T2×S1×M2	**+0.00**	**+3.75**	**+3.60**	**+5.42**
	T2×S2×M1	-1.03	+5.33	+5.90	+2.47
	T2×S2×M2	+0.00	+2.37	+6.40	+4.23

Table 14: Stage-wise Squared-Hellinger TRSD cross-task grid. The first row of each stage reports the matched SDFT baseline in percent; subsequent rows report TRSD differences from this baseline in percentage points.

Stage	Configuration	Δ Tool	Δ Science	Δ Medical	Δ IFEval
After Tool	SDFT baseline	62.89	–	–	53.14
	T1	**+9.27**	–	–	**+4.94**
	T2	+6.18	–	–	+5.32
After Science	SDFT baseline	60.82	68.44	–	42.23
	T1×S1	+7.22	-0.20	–	+4.02
	T1×S2	**+7.22**	**+0.00**	–	**+2.78**
	T2×S1	+3.10	-0.39	–	+4.34
	T2×S2	+3.10	+2.17	–	+2.83
After Medical	SDFT baseline	51.55	63.91	39.80	38.12
	T1×S1×M1	+10.31	+5.72	-1.30	+2.75
	T1×S1×M2	+8.24	+2.95	-1.80	+1.71
	T1×S2×M1	+10.31	+2.76	-1.20	+4.33
	T1×S2×M2	**+6.18**	**+3.94**	**-0.70**	**+9.24**
	T2×S1×M1	+4.12	+1.97	-2.10	+9.52
	T2×S1×M2	+5.15	+1.77	-0.60	+3.31
	T2×S2×M1	+4.12	+3.15	-1.00	+1.07
	T2×S2×M2	+7.21	+2.36	+0.30	+2.58

On-policy and contextual self-distillation. On-policy learning reduces the mismatch between training and inference by optimizing on trajectories induced by the learner itself (Ross et al., 2011; Agarwal et al., 2024). Context distillation instead transfers behavior elicited by additional information into a model operating without that context (Snell et al., 2022), while earlier self-distillation approaches construct model-aligned offline targets through self-generated rewriting (Yang et al., 2024b). Recent methods form an on-policy self-teacher from reasoning traces or other contextual information, such as feedback on the model's own attempt (Zhao et al., 2026; Hübotter et al., 2026; Ye et al., 2026). Their focus is constructing this teacher so the student can absorb privileged context along its own trajectories. SDFT instead uses a demonstration-conditioned teacher for continual adaptation, replacing one-hot supervision so a new skill can be learned with less disruption of prior behavior (Shenfeld et al., 2026). Our focus is the heavy tail that remains: the teacher stays close

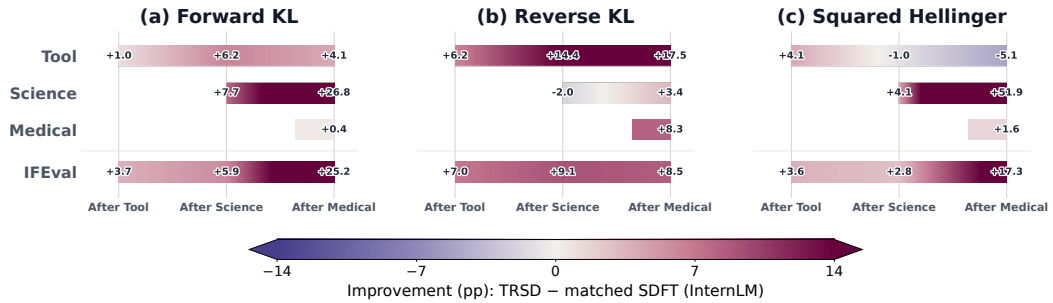

Figure 9: Percentage-point improvements of InternLM2.5-7B-Chat TRSD over the matched SDFT objective during sequential Tool Use → Science Q&A → Medical adaptation. Each panel corresponds to one divergence geometry, and each stage shows the capabilities introduced up to that point together with IFEval retention.

Table 15: InternLM2.5 Forward-KL TRSD cross-task results.

Stage	Configuration	Δ Tool	Δ Science	Δ Medical	Δ IFEval
After Tool	SDFT baseline	48.45	–	–	46.59
	T1	**+1.03**	–	–	**+3.66**
	T2	+0.00	–	–	+3.67
After Science	SDFT baseline	47.42	50.49	–	37.32
	T1×S1	**+6.19**	**+7.69**	–	**+5.87**
	T1×S2	+4.12	+7.10	–	+3.58
	T2×S1	+0.00	+7.10	–	+3.65
	T2×S2	-1.03	+9.86	–	+3.50
After Medical	SDFT baseline	44.33	28.40	26.40	9.91
	T1×S1×M1	+4.12	+26.63	+0.40	+25.22
	T1×S1×M2	**+4.12**	**+26.82**	**+0.40**	**+25.22**
	T2×S2×M1	+1.03	+22.49	+2.20	+25.04
	T2×S2×M2	+1.03	+23.87	+1.50	+25.84

to the student on most examples, while large divergences on the remaining examples contribute disproportionately to forgetting. We constrain the teacher sample by sample, so the new skill is acquired while improving the retention of previously learned capabilities.

Trust-region and divergence-aware teacher construction. Trust-region methods stabilize policy optimization by limiting the movement induced by each update (Schulman et al., 2015). Recent work extends this principle to on-policy distillation. Trust Region Policy Distillation (TOP-D) directly interpolates an external teacher with the current student in probability space, using a fixed global coefficient to bound the log-ratio reward and support efficient internal optimization (Xie et al., 2026). TRSD shares the broader idea of moderating teacher-induced updates, but derives the effective teacher from a divergence-constrained fidelity problem. The resulting path depends on the chosen f-divergence, and its interpolation coefficient adapts to the divergence geometry of each example. This formulation connects TRSD's trust-region-constrained teacher construction to the heavy-tailed self-teacher shifts that arise in continual adaptation, rather than to teacher–student capacity gaps or optimization variance alone.

Table 16: InternLM2.5 Reverse-KL TRSD cross-task results.

Stage	Configuration	Δ Tool	Δ Science	Δ Medical	Δ IFEval
After Tool	SDFT baseline	39.18	–	–	46.39
	T1	+8.24	–	–	+8.43
	T2	**+6.18**	–	–	**+6.98**
After Science	SDFT baseline	36.08	60.36	–	36.60
	T1×S1	+13.40	-2.56	–	+5.75
	T1×S2	+14.43	-4.14	–	+5.52
	T2×S1	+16.49	-0.79	–	+7.85
	T2×S2	**+14.43**	**-1.97**	–	**+9.10**
After Medical	SDFT baseline	36.08	50.49	23.60	27.34
	T1×S2×M1	+16.49	+0.99	+8.40	+9.25
	T1×S2×M2	+16.49	+0.99	+8.00	+9.28
	T2×S2×M1	**+17.53**	**+3.35**	**+8.30**	**+8.47**
	T2×S2×M2	+14.43	+2.76	+7.10	+13.09

Table 17: InternLM2.5 Squared-Hellinger TRSD cross-task results.

Stage	Configuration	Δ Tool	Δ Science	Δ Medical	Δ IFEval
After Tool	SDFT baseline	43.30	–	–	49.80
	T1	**+4.12**	–	–	**+3.64**
	T2	+2.06	–	–	+4.79
After Science	SDFT baseline	46.39	54.24	–	42.86
	T1×S1	**-1.03**	**+4.14**	–	**+2.80**
	T1×S2	-2.06	+1.38	–	+2.86
	T2×S1	+1.03	+1.78	–	+2.30
	T2×S2	+0.00	+2.76	–	+3.53
After Medical	SDFT baseline	46.39	0.00	29.10	23.81
	T1×S1×M1	-6.19	+50.89	-1.00	+14.96
	T1×S1×M2	**-5.15**	**+51.87**	**+1.60**	**+17.30**
	T1×S2×M1	-5.15	+52.66	+1.40	+14.24
	T1×S2×M2	-4.12	+52.47	+0.90	+13.51

Table 18: Qwen2.5-7B-Instruct single-task Forward KL TRSD sweep on Tool. The first row reports the matched SDFT baseline in percent; grid entries report TRSD differences in percentage points.

SDFT baseline (absolute): Task Acc. 63.92; IFEval 44.14					
$\rho = 0.05$			$\rho = 0.10$		
$[\beta_{\min}, \beta_{\max}]$	Δ Acc.	Δ IFEval	$[\beta_{\min}, \beta_{\max}]$	Δ Acc.	Δ IFEval
$[0.00, 0.75]$	+1.03	+4.52	$[0.00, 0.75]$	+6.18	+0.58
$[0.00, 0.80]$	-3.10	+0.52	$[0.00, 0.80]$	-4.13	+8.10
$[0.00, 0.85]$	+3.09	+7.24	$[0.00, 0.85]$	-2.06	+0.14
$[0.00, 1.00]$	+0.00	+8.29	$[0.00, 1.00]$	+5.15	+0.71
$[0.05, 0.75]$	-2.06	+2.71	$[0.05, 0.75]$	+0.00	+3.39
$[0.05, 0.80]$	+0.00	+10.70	$[0.05, 0.80]$	+0.00	-0.27
$[0.05, 0.85]$	+6.18	+5.79	$[0.05, 0.85]$	-8.25	+2.12
$[0.05, 1.00]$	+2.06	+13.67	$[0.05, 1.00]$	-4.13	+3.97
$[0.10, 0.75]$	+1.03	+3.06	$[0.10, 0.75]$	+3.09	-0.83
$[0.10, 0.80]$	-2.06	+1.43	$[0.10, 0.80]$	+1.03	+9.36
$[0.10, 0.85]$	-1.03	+6.15	$[0.10, 0.85]$	-5.16	-0.15
$[0.10, 1.00]$	-4.13	+4.19	$[0.10, 1.00]$	+1.03	+1.32
$[0.15, 0.75]$	+5.15	+4.30	$[0.15, 0.75]$	+0.00	+3.68
$[0.15, 0.80]$	+2.06	+0.46	$[0.15, 0.80]$	+3.09	+3.17
$[0.15, 0.85]$	+0.00	+9.85	$[0.15, 0.85]$	-1.03	-1.22
$[0.15, 1.00]$	-1.03	+6.00	$[0.15, 1.00]$	-3.10	+2.21
$[0.20, 0.75]$	+1.03	+4.54	$[0.20, 0.75]$	+2.06	-0.04
$[0.20, 0.80]$	+4.12	+1.59	$[0.20, 0.80]$	-2.06	+4.02
$[0.20, 0.85]$	+2.06	+3.82	$[0.20, 0.85]$	+2.06	+4.83
$[0.20, 1.00]$	-2.06	+2.68	$[0.20, 1.00]$	+0.00	+0.46

Table 19: Qwen2.5-7B-Instruct single-task Forward KL TRSD sweep on Science. The first row reports the matched SDFT baseline in percent; grid entries report TRSD differences in percentage points.

SDFT baseline (absolute): Task Acc. 65.88; IFEval 45.60					
	$\rho = 0.05$			$\rho = 0.10$	
$[\beta_{\min}, \beta_{\max}]$	Δ Acc.	Δ IFEval	$[\beta_{\min}, \beta_{\max}]$	Δ Acc.	Δ IFEval
$[0.00, 0.75]$	+2.17	+7.09	$[0.00, 0.75]$	+4.14	+3.18
$[0.00, 0.80]$	+2.36	-0.78	$[0.00, 0.80]$	+1.38	+2.06
$[0.00, 0.85]$	+0.39	+0.60	$[0.00, 0.85]$	+1.77	-2.33
$[0.00, 1.00]$	-0.20	+1.05	$[0.00, 1.00]$	+1.77	-0.33
$[0.05, 0.75]$	+2.17	+1.23	$[0.05, 0.75]$	+4.73	+0.31
$[0.05, 0.80]$	-0.40	+0.86	$[0.05, 0.80]$	+3.55	-0.23
$[0.05, 0.85]$	+2.96	+2.95	$[0.05, 0.85]$	+2.76	-0.29
$[0.05, 1.00]$	+2.76	-1.01	$[0.05, 1.00]$	+1.18	+0.69
$[0.10, 0.75]$	+2.56	-0.41	$[0.10, 0.75]$	+0.00	+1.97
$[0.10, 0.80]$	+1.97	+0.18	$[0.10, 0.80]$	+1.58	+3.79
$[0.10, 0.85]$	-2.37	+2.85	$[0.10, 0.85]$	+1.38	+0.16
$[0.10, 1.00]$	-0.99	+1.43	$[0.10, 1.00]$	-0.20	+0.21
$[0.15, 0.75]$	-0.79	+6.49	$[0.15, 0.75]$	-0.79	+2.34
$[0.15, 0.80]$	+0.59	+1.71	$[0.15, 0.80]$	+2.36	+7.46
$[0.15, 0.85]$	-1.19	+5.29	$[0.15, 0.85]$	+2.36	+3.26
$[0.15, 1.00]$	+0.39	+2.38	$[0.15, 1.00]$	+1.18	+1.58
$[0.20, 0.75]$	+2.17	+0.98	$[0.20, 0.75]$	+2.17	+4.65
$[0.20, 0.80]$	+2.36	+6.52	$[0.20, 0.80]$	+0.59	+4.80
$[0.20, 0.85]$	+1.77	+1.45	$[0.20, 0.85]$	+2.76	+2.04
$[0.20, 1.00]$	+0.98	+1.59	$[0.20, 1.00]$	+4.73	-0.39

Table 20: Qwen2.5-7B-Instruct single-task Forward KL TRSD sweep on Medical. The first row reports the matched SDFT baseline in percent; grid entries report TRSD differences in percentage points.

SDFT baseline (absolute): Task Acc. 40.20; IFEval 48.56					
	$\rho = 0.05$			$\rho = 0.10$	
$[\beta_{\min}, \beta_{\max}]$	Δ Acc.	Δ IFEval	$[\beta_{\min}, \beta_{\max}]$	Δ Acc.	Δ IFEval
$[0.00, 0.80]$	-2.40	+10.57	$[0.00, 0.80]$	-3.60	+4.71
$[0.00, 0.85]$	-3.20	+4.80	$[0.00, 0.85]$	-0.60	+10.69
$[0.00, 1.00]$	+0.50	+6.55	$[0.00, 1.00]$	-0.50	+6.41
$[0.05, 0.80]$	-2.20	+8.21	$[0.05, 0.80]$	-1.60	-1.85
$[0.05, 0.85]$	-0.10	+8.06	$[0.05, 0.85]$	-1.50	+6.57
$[0.05, 1.00]$	-4.60	+2.87	$[0.05, 1.00]$	-2.80	+7.23
$[0.10, 0.80]$	-0.20	+3.38	$[0.10, 0.80]$	-0.30	+4.39
$[0.10, 0.85]$	+0.40	+9.68	$[0.10, 0.85]$	+0.70	+12.35
$[0.10, 1.00]$	-3.60	+5.90	$[0.10, 1.00]$	-4.50	+8.46
$[0.15, 0.80]$	-0.70	+12.34	$[0.15, 0.80]$	+1.70	+6.68
$[0.15, 0.85]$	-2.40	+12.19	$[0.15, 0.85]$	-1.70	+0.85
$[0.15, 1.00]$	-0.90	+2.83	$[0.15, 1.00]$	-0.90	+9.84

Table 21: Qwen2.5-7B-Instruct single-task Reverse KL TRSD sweep on Tool. The first row reports the matched SDFT baseline in percent; grid entries report TRSD differences in percentage points.

SDFT baseline (absolute): Task Acc. 64.95; IFEval 43.80					
	$\rho = 0.05$			$\rho = 0.20$	
$[\beta_{\min}, \beta_{\max}]$	Δ Acc.	Δ IFEval	$[\beta_{\min}, \beta_{\max}]$	Δ Acc.	Δ IFEval
$[0.00, 0.75]$	+5.15	+8.87	$[0.00, 0.75]$	-1.05	+3.40
$[0.00, 0.80]$	-3.09	+9.15	$[0.00, 0.80]$	-2.05	+7.60
$[0.00, 0.85]$	-2.06	+5.96	$[0.00, 0.85]$	-0.05	+6.30
$[0.00, 1.00]$	+0.00	+9.02	$[0.00, 1.00]$	-0.05	+3.10
$[0.05, 0.75]$	+3.09	+7.89	$[0.05, 0.75]$	-6.15	+6.30
$[0.05, 0.80]$	+2.06	+11.82	$[0.05, 0.80]$	+5.15	+8.20
$[0.05, 0.85]$	-2.06	+5.93	$[0.05, 0.85]$	-1.05	+4.60
$[0.05, 1.00]$	+5.15	+7.02	$[0.05, 1.00]$	-3.05	+3.80
$[0.10, 0.75]$	+5.15	+7.19	$[0.10, 0.75]$	+2.05	+4.90
$[0.10, 0.80]$	+3.09	+7.81	$[0.10, 0.80]$	+5.15	+6.80
$[0.10, 0.85]$	+0.00	+5.39	$[0.10, 0.85]$	-7.25	+5.40
$[0.10, 1.00]$	-3.09	+7.40	$[0.10, 1.00]$	-3.05	+5.70
$[0.15, 0.75]$	+4.12	+8.30	$[0.15, 0.75]$	-2.05	+10.80
$[0.15, 0.80]$	+3.09	+10.27	$[0.15, 0.80]$	-6.15	+6.80
$[0.15, 0.85]$	+3.09	+5.76	$[0.15, 0.85]$	-9.25	+7.10
$[0.15, 1.00]$	-4.13	+5.62	$[0.15, 1.00]$	-2.05	+7.70
$[0.20, 0.75]$	+1.03	+5.57	$[0.20, 0.75]$	-8.25	+5.90
$[0.20, 0.80]$	+1.03	+6.47	$[0.20, 0.80]$	+1.05	+5.80
$[0.20, 0.85]$	+3.09	+4.92	$[0.20, 0.85]$	-7.25	+3.90
$[0.20, 1.00]$	-1.03	+10.05	$[0.20, 1.00]$	-2.05	+6.10

Table 22: Qwen2.5-7B-Instruct single-task Reverse KL TRSD sweep on Science. The first row reports the matched SDFT baseline in percent; grid entries report TRSD differences in percentage points.

SDFT baseline (absolute): Task Acc. 67.06; IFEval 49.82					
	$\rho = 0.05$			$\rho = 0.20$	
$[\beta_{\min}, \beta_{\max}]$	Δ Acc.	Δ IFEval	$[\beta_{\min}, \beta_{\max}]$	Δ Acc.	Δ IFEval
$[0.00, 0.75]$	+0.79	+2.50	$[0.00, 0.75]$	-1.16	+6.78
$[0.00, 0.80]$	-1.38	+1.76	$[0.00, 0.80]$	-0.16	+5.78
$[0.00, 0.85]$	-1.38	+6.16	$[0.00, 0.85]$	-0.36	+5.78
$[0.00, 1.00]$	-2.37	+1.69	$[0.00, 1.00]$	-0.56	+5.78
$[0.05, 0.75]$	-0.59	+2.10	$[0.05, 0.75]$	-0.16	+4.18
$[0.05, 0.80]$	-3.55	+1.06	$[0.05, 0.80]$	-0.16	+4.18
$[0.05, 0.85]$	+0.20	+2.84	$[0.05, 0.85]$	-1.36	+3.98
$[0.05, 1.00]$	+0.20	+0.94	$[0.05, 1.00]$	-0.36	+3.28
$[0.10, 0.75]$	-0.20	+2.22	$[0.10, 0.75]$	-0.16	+3.28
$[0.10, 0.80]$	-0.79	+3.57	$[0.10, 0.80]$	-2.16	+4.58
$[0.10, 0.85]$	-1.58	+1.14	$[0.10, 0.85]$	-1.56	+4.58
$[0.10, 1.00]$	-1.77	+2.05	$[0.10, 1.00]$	+0.04	+4.78
$[0.15, 0.75]$	-3.55	+3.39	$[0.15, 0.75]$	-2.76	+2.98
$[0.15, 0.80]$	+0.20	+3.71	$[0.15, 0.80]$	-0.56	+3.48
$[0.15, 0.85]$	-4.73	+2.24	$[0.15, 0.85]$	-0.56	+4.98
$[0.15, 1.00]$	+0.00	+5.36	$[0.15, 1.00]$	-3.16	+3.18
$[0.20, 0.75]$	+1.58	-0.61	$[0.20, 0.75]$	-0.36	+4.58
$[0.20, 0.80]$	-1.58	-0.11	$[0.20, 0.80]$	-0.16	+5.78
$[0.20, 0.85]$	-1.77	-1.37	$[0.20, 0.85]$	+0.54	+6.08
$[0.20, 1.00]$	+1.38	+4.35	$[0.20, 1.00]$	+0.24	+5.58

Table 23: Qwen2.5-7B-Instruct single-task Reverse KL TRSD sweep on Medical. The first row reports the matched SDFT baseline in percent; grid entries report TRSD differences in percentage points.

SDFT baseline (absolute): Task Acc. 38.60; IFEval 49.01					
	$\rho = 0.05$			$\rho = 0.20$	
$[\beta_{\min}, \beta_{\max}]$	Δ Acc.	Δ IFEval	$[\beta_{\min}, \beta_{\max}]$	Δ Acc.	Δ IFEval
$[0.00, 0.80]$	-0.80	+6.12	$[0.00, 0.80]$	+2.70	+5.92
$[0.00, 0.85]$	+0.30	+3.16	$[0.00, 0.85]$	+1.00	+7.71
$[0.00, 1.00]$	+2.50	+4.08	$[0.00, 1.00]$	+1.90	+4.60
$[0.05, 0.80]$	+1.30	+4.41	$[0.05, 0.80]$	+2.00	+6.45
$[0.05, 0.85]$	+1.60	+5.99	$[0.05, 0.85]$	+0.50	+0.51
$[0.05, 1.00]$	+3.70	+5.72	$[0.05, 1.00]$	+2.30	+0.39
$[0.10, 0.80]$	+2.00	+3.33	$[0.10, 0.80]$	+2.00	+7.75
$[0.10, 0.85]$	+2.30	+5.90	$[0.10, 0.85]$	+1.80	+5.69
$[0.10, 1.00]$	+1.00	+6.88	$[0.10, 1.00]$	+0.00	+10.75
$[0.15, 0.80]$	+1.00	+1.40	$[0.15, 0.80]$	+0.80	+7.25
$[0.15, 0.85]$	+2.50	+3.79	$[0.15, 0.85]$	-0.20	+6.79
$[0.15, 1.00]$	+1.60	+7.02	$[0.15, 1.00]$	-1.40	+4.96

Table 24: Qwen2.5-7B-Instruct single-task Squared Hellinger TRSD sweep on Tool. The first row reports the matched SDFT baseline in percent; grid entries report TRSD differences in percentage points.

SDFT baseline (absolute): Task Acc. 62.89; IFEval 53.14					
	$\rho = 0.05$			$\rho = 0.15$	
$[\beta_{\min}, \beta_{\max}]$	Δ Acc.	Δ IFEval	$[\beta_{\min}, \beta_{\max}]$	Δ Acc.	Δ IFEval
$[0.00, 0.75]$	+4.12	+1.40	$[0.00, 0.75]$	+0.00	-2.26
$[0.00, 0.80]$	+5.15	+3.20	$[0.00, 0.80]$	+8.24	+2.89
$[0.00, 0.85]$	+0.00	+1.65	$[0.00, 0.85]$	+9.27	+4.94
$[0.00, 1.00]$	+3.09	+3.73	$[0.00, 1.00]$	+6.18	+2.17
$[0.05, 0.75]$	+2.06	+0.48	$[0.05, 0.75]$	+6.18	+4.73
$[0.05, 0.80]$	+2.06	+4.35	$[0.05, 0.80]$	+9.27	+2.74
$[0.05, 0.85]$	+2.06	+1.91	$[0.05, 0.85]$	+1.03	+0.14
$[0.05, 1.00]$	+2.06	+8.03	$[0.05, 1.00]$	+2.06	+0.91
$[0.10, 0.75]$	+4.12	+0.36	$[0.10, 0.75]$	+4.12	+4.31
$[0.10, 0.80]$	+2.06	+3.46	$[0.10, 0.80]$	+4.12	+0.46
$[0.10, 0.85]$	+3.09	+1.54	$[0.10, 0.85]$	-1.03	+0.85
$[0.10, 1.00]$	+0.00	+4.52	$[0.10, 1.00]$	+7.21	+0.54
$[0.15, 0.75]$	-2.07	+0.46	$[0.15, 0.75]$	+2.06	+4.50
$[0.15, 0.80]$	+5.15	+4.93	$[0.15, 0.80]$	+0.00	+1.82
$[0.15, 0.85]$	+6.18	+1.61	$[0.15, 0.85]$	+1.03	+1.44
$[0.15, 1.00]$	+1.03	+3.38	$[0.15, 1.00]$	+1.03	-1.42
$[0.20, 0.75]$	+6.18	+5.32	$[0.20, 0.75]$	+1.03	+0.93
$[0.20, 0.80]$	+3.09	+3.75	$[0.20, 0.80]$	-3.10	+4.92
$[0.20, 0.85]$	+3.09	+3.09	$[0.20, 0.85]$	+2.06	+4.98
$[0.20, 1.00]$	-2.07	+2.59	$[0.20, 1.00]$	-1.03	+0.41

Table 25: Qwen2.5-7B-Instruct single-task Squared Hellinger TRSD sweep on Science. The first row reports the matched SDFT baseline in percent; grid entries report TRSD differences in percentage points.

SDFT baseline (absolute): Task Acc. 65.88; IFEval 49.72					
	$\rho = 0.15$			$\rho = 0.25$	
$[\beta_{\min}, \beta_{\max}]$	Δ Acc.	Δ IFEval	$[\beta_{\min}, \beta_{\max}]$	Δ Acc.	Δ IFEval
$[0.00, 0.75]$	-3.95	+2.35	$[0.00, 0.75]$	-0.99	+0.74
$[0.00, 0.80]$	-1.97	+3.24	$[0.00, 0.80]$	-1.78	+1.09
$[0.00, 0.85]$	-2.17	+2.43	$[0.00, 0.85]$	-1.97	+2.91
$[0.00, 1.00]$	-1.78	+3.13	$[0.00, 1.00]$	-1.78	+2.31
$[0.05, 0.75]$	-1.58	+1.59	$[0.05, 0.75]$	+1.18	+3.87
$[0.05, 0.80]$	-1.58	+2.73	$[0.05, 0.80]$	+0.00	+3.22
$[0.05, 0.85]$	-1.97	+1.21	$[0.05, 0.85]$	-2.17	+2.67
$[0.05, 1.00]$	-3.36	+1.58	$[0.05, 1.00]$	+0.41	+4.58
$[0.10, 0.75]$	-1.58	+4.22	$[0.10, 0.75]$	-1.97	+4.73
$[0.10, 0.80]$	+0.19	-0.13	$[0.10, 0.80]$	+2.19	+3.68
$[0.10, 0.85]$	+0.00	+0.49	$[0.10, 0.85]$	-1.58	+4.48
$[0.10, 1.00]$	+1.38	-0.37	$[0.10, 1.00]$	-0.79	+4.53
$[0.15, 0.75]$	-2.37	+1.04	$[0.15, 0.75]$	+1.58	+1.30
$[0.15, 0.80]$	+0.00	-2.20	$[0.15, 0.80]$	+0.79	+1.19
$[0.15, 0.85]$	-2.37	-2.23	$[0.15, 0.85]$	-0.20	+2.24
$[0.15, 1.00]$	-0.20	-0.33	$[0.15, 1.00]$	+2.17	+2.51
$[0.20, 0.75]$	+0.59	+2.43	$[0.20, 0.75]$	-1.78	+1.36
$[0.20, 0.80]$	-1.58	+0.64	$[0.20, 0.80]$	-2.17	+0.85
$[0.20, 0.85]$	-1.58	+1.27	$[0.20, 0.85]$	+0.19	+2.15
$[0.20, 1.00]$	-2.96	+0.19	$[0.20, 1.00]$	-0.20	+0.54

Table 26: Qwen2.5-7B-Instruct single-task Squared Hellinger TRSD sweep on Medical. The first row reports the matched SDFT baseline in percent; grid entries report TRSD differences in percentage points.

SDFT baseline (absolute): Task Acc. 37.60; IFEval 55.73					
	$\rho = 0.10$			$\rho = 0.20$	
$[\beta_{\min}, \beta_{\max}]$	Δ Acc.	Δ IFEval	$[\beta_{\min}, \beta_{\max}]$	Δ Acc.	Δ IFEval
$[0.00, 0.80]$	+2.00	-0.08	$[0.00, 0.80]$	+1.50	+0.67
$[0.00, 0.85]$	+2.50	+0.46	$[0.00, 0.85]$	+2.00	-3.39
$[0.00, 1.00]$	+3.00	+5.96	$[0.00, 1.00]$	+2.30	+0.37
$[0.05, 0.80]$	+3.50	+0.64	$[0.05, 0.80]$	+4.00	+4.97
$[0.05, 0.85]$	-0.20	-2.99	$[0.05, 0.85]$	+2.80	+0.52
$[0.05, 1.00]$	+0.60	+3.92			
$[0.10, 0.80]$	+1.50	+7.81	$[0.10, 0.80]$	+2.10	-0.65
$[0.10, 0.85]$	+0.80	+0.76	$[0.10, 0.85]$	-0.70	+3.70
$[0.10, 1.00]$	-0.20	+1.01	$[0.10, 1.00]$	+2.90	+4.18
$[0.15, 0.80]$	+1.10	-3.31	$[0.15, 0.80]$	+2.60	+0.04
$[0.15, 0.85]$	+2.60	+7.04	$[0.15, 0.85]$	+1.90	+5.75
$[0.15, 1.00]$	+3.40	+4.63	$[0.15, 1.00]$	+2.60	+6.56