Parameter-Efficient Is Not Function-Stable: Function Trust for Model Adaptation

Anonymous authors
Paper under double-blind review

Abstract

Parameter efficiency constrains where a model changes, but not how far its predictive function moves. We introduce Function Trust, a parameterization-agnostic framework that adapts a model subject to a user-specified budget on the forward KL divergence from a frozen source predictor over cached source inputs. A primal–dual Powell–Hestenes–Rockafellar augmented Lagrangian approximately solves the constrained problem, while the budget formulation remains independent of the solver. Conditional on achieving the replay constraint within tolerance ξ, we bound population functional drift by $\rho_{\mathrm{KL}} + \xi + \Gamma_n$ and source-risk change by $L\sqrt{(\rho_{\mathrm{KL}} + \xi + \Gamma_n)/2}$, where Γ_n captures replay-to-population generalization; a margin corollary controls deterministic prediction changes. Across TinyBERT and Qwen3-0.6B, Function Trust consistently reduces source forgetting and held-out KL for LoRA, full fine-tuning, and our structured BridgeOFT adapter, with small or parameterization-dependent changes in target quality. Yet equal replay budgets need not imply equal held-out behavior: increasing TinyBERT replay from 64 to 4096 examples reduces validation KL from 0.0544 to 0.0255, while low-margin replay at fixed $n = 256$ reduces it from 0.0421 to 0.0209. Fixed-penalty and augmented-Lagrangian frontiers overlap; our contribution is therefore not solver dominance. The evidence instead shows that functional stability is not a by-product of parameter efficiency: it must be explicitly budgeted, measured, and audited.

1 Introduction

Fine-tuning can improve a target capability while degrading behavior acquired before adaptation. Updating few parameters does not prevent this failure: LoRA and related parameter-efficient fine-tuning (PEFT) methods constrain where parameter updates occur, not the resulting movement in function space (Hu et al., 2022; Han et al., 2024). Continual fine-tuning studies observe the same stability–plasticity tension in language models (Luo et al., 2023). This paper studies the narrower problem of preserving a fixed source anchor during one adaptation step; it does not claim to evaluate a sequential continual-learning trajectory. We control the protected object directly—the source predictive distribution.

We propose Function Trust as budgeted functional retention. Given a frozen source model p_0, an adapted model p_θ, and source replay inputs, adaptation solves

$$\min_{\theta} \mathcal{L}_{\mathrm{target}}(\theta) \quad \text{s.t.} \quad \widehat{D}_{\mathcal{R}}(\theta) = \frac{1}{n} \sum_{x_i \in \mathcal{R}} \mathrm{KL}(p_0(\cdot \mid x_i) \| p_\theta(\cdot \mid x_i)) \leq \rho_{\mathrm{KL}}. \tag{1}$$

Unlike a fixed regularization weight λ, ρ_{KL} specifies the quantity to be limited. We use a primal–dual augmented-Lagrangian solver, but separate the constraint from its numerical realization: the theory is conditional on achieved feasibility, and the experiments do not establish superiority over tuned fixed penalties.

Our second proposal, BridgeOFT, addresses the complementary question of parameterization. Standard OFT preserves Euclidean geometry but does not account for a mismatch between source and target activation covariances; conversely, a deterministic transport map

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.

can align second moments but offers no task-adaptive rotational freedom. BridgeOFT combines both: a Gaussian optimal-transport (OT) bridge supplies a covariance-aware path, and a blockwise Cayley-parametrized orthogonal gauge supplies trainable flexibility without dense cubic-scale parameters. The two proposals are independent: BridgeOFT operates without Function Trust, whereas Function Trust can constrain BridgeOFT, LoRA, or full fine-tuning. Experiments isolate their roles: covariance geometry changes target behavior, but the functional constraint accounts for the consistent reduction in drift.

Precise novelty. Output distillation, KL trust regions, replay, and Lagrangian optimization are established ingredients. Our novelty lies in formulating and auditing them as a user-specified source-function budget shared across heterogeneous adaptation parameterizations. We pair this formulation with a replay-generalization guarantee, fixed-λ controls, and direct evidence that replay-local trust need not yield held-out preservation. Separately, BridgeOFT combines activation-covariance Gaussian transport with a trainable orthogonal gauge; orthogonal structure alone is not presented as a retention guarantee.

Our contributions are:

- Function Trust, an explicit replay-distribution functional budget applicable to structured PEFT, LoRA, and full fine-tuning;
- a conditional theorem relating empirical feasibility to population drift and source-risk degradation, with replay coverage exposed through an explicit generalization term;
- BridgeOFT, a Gaussian-OT structured adapter with a trainable blockwise orthogonal gauge, evaluated independently of the retention constraint;
- controlled evidence that replay coverage and composition determine the held-out meaning of an empirical budget, supported by scaling, margin-aware selection, feasibility, fixed-penalty, and cross-model analyses.

2 Related Work

Parameter-efficient and orthogonal adaptation. Adapters, prefix tuning, and LoRA reduce the number of trainable parameters (Houlsby et al., 2019; Li & Liang, 2021; Hu et al., 2022). Orthogonal fine-tuning (OFT) rotates pretrained weights while preserving geometric quantities; BOFT improves its scalability with butterfly factorization, and later variants use Givens rotations or Householder reflections (Qiu et al., 2023; Liu et al., 2023; Ma et al., 2024; Yuan et al., 2024). These methods motivate structured rotations, but BridgeOFT differs by deriving a frozen path from source/target activation covariances and learning an orthogonal gauge on that path. Our ablations also distinguish orthogonal parameterization from functional retention.

Forgetting and functional regularization. Learning without Forgetting matches old outputs during new-task training, while knowledge distillation transfers predictive distributions from a teacher (Li & Hoiem, 2016; Hinton et al., 2015). Elastic Weight Consolidation protects important parameters rather than outputs (Kirkpatrick et al., 2017). Bayesian functional regularization stores sparse function-space summaries, and projected functional regularization protects self-supervised representations (Titsias et al., 2020; Gomez-Villa et al., 2021). Recent LLM studies also use replay and output-level KL imitation: context-free synthetic data estimates a model-distribution KL without pretraining data, while GeRe evaluates fixed general-sample replay with label fitting and KL imitation (Huang et al., 2025; Zhang et al., 2025). Function Trust does not claim any of these ingredients as new. Its distinction is to specify forward KL to a fixed source predictor on source replay as an explicit inequality, rather than as an unconstrained distillation term or fixed-weight penalty, and to expose the same constraint under LoRA, BridgeOFT, and full fine-tuning.

Trust regions and constrained learning. Explicit KL regions are central in policy optimization, where TRPO and constrained policy optimization control update or cost distributions

(Schulman et al., 2015; Achiam et al., 2017). In continual learning, OCAR explicitly constrains KL variation on replay data, but uses a previous-iterate/local online formulation and curvature-aware K-FAC updates (Urettini & Carta, 2025); Trust Region Continual Learning combines generative replay with a Fisher-metric local trust region in diffusion settings (Wang et al., 2026b). Concurrent Anchored Learning controls supervised LLM fine-tuning with a moving target interpolated between the current model and a frozen reference and derives local per-iteration KL bounds without replay (Wang et al., 2026a). Our distinction is narrower: Function Trust uses a fixed source anchor, an aggregate replay-distribution budget, and a held-out source-population audit in supervised target-task adaptation. Generic Lagrangian and proxy-Lagrangian work establishes optimization machinery and data-dependent constraint considerations (Fioretto et al., 2020; Cotter et al., 2018a;b); we make no optimizer-level novelty or convergence claim.

Activation transport. OT aligns source and target distributions in domain adaptation and can align neurons before model fusion (Damodaran et al., 2018; Ju & Guan, 2022; Singh & Jaggi, 2019). Model soups instead average compatible fine-tuned weights (Wortsman et al., 2022). BridgeOFT uses closed-form Gaussian transport at activation interfaces, followed by a learned orthogonal gauge; it neither averages model weights nor treats geometry as a retention certificate.

3 Method

3.1 Function Trust

Let P_s be the protected source joint distribution with input marginal P_s^X, $p_0(y \mid x)$ the frozen source predictor, and $p_\theta(y \mid x)$ the adapted predictor. Define population drift and its replay estimate as

$$D_s(\theta) = \mathbb{E}_{x\sim P_s^X}\left[\mathrm{KL}(p_0(\cdot \mid x)\|p_\theta(\cdot \mid x))\right], \tag{2}$$

$$\widehat{D}_{\mathcal{R}}(\theta) = \frac{1}{n}\sum_{i=1}^{n} \mathrm{KL}(p_0(\cdot \mid x_i)\|p_\theta(\cdot \mid x_i)). \tag{3}$$

The target objective is the expected supervised loss, $\mathcal{L}_{\text{target}}(\theta) = \mathbb{E}_{(x,y)\sim P_t}[\ell_t(p_\theta(\cdot \mid x), y)]$, approximated by minibatches from the target-training split. The replay cache stores source inputs and frozen source logits, so source labels and source-model forward passes are unnecessary during target optimization. We call this a fixed-anchor budget because the reference predictor is held fixed throughout adaptation; it is not a global guarantee over all inputs.

For $g(\theta) = \widehat{D}_{\mathcal{R}}(\theta) - \rho_{\mathrm{KL}}$, we optimize the Powell–Hestenes–Rockafellar surrogate

$$\mathcal{L}_{\mathrm{PHR}}(\theta, \mu) = \mathcal{L}_{\text{target}}(\theta) + \frac{[\mu + \beta g(\theta)]_+^2 - \mu^2}{2\beta}, \tag{4}$$

with nonnegative dual variable μ and penalty coefficient β. We initialize $(\mu, \beta) = (0, 10)$ and update dual statistics outside backpropagation; Appendix B gives the exact update, stall rule, frequency, and cap. This is an approximate nonconvex solver, so we report achieved drift and residuals rather than assuming feasibility. No-Trust controls use the same target schedule but no replay forward pass; replay access is therefore part of the Trust mechanism.

3.2 A conditional source-risk guarantee

The following result formalizes what an achieved replay budget does—and does not—guarantee. Let $\mathcal{K} = \{x \mapsto \mathrm{KL}(p_0(\cdot \mid x)\|p_\theta(\cdot \mid x)) : \theta \in \Theta\}$. Assume that, with probability at least $1 - \delta$ over an i.i.d. replay sample, a uniform deviation bound

$$\sup_{\theta\in\Theta}\left|D_s(\theta) - \widehat{D}_{\mathcal{R}}(\theta)\right| \le \Gamma_n(\mathcal{K}, \delta) \tag{5}$$

holds. For bounded KL values in $[0, B]$, Γ_n can be instantiated by standard Rademacher or finite-class bounds; for example, a finite candidate set of size M gives $B\sqrt{\log(M/\delta)/(2n)}$ by Hoeffding and a union bound.

Theorem 1 (Replay budget implies bounded stochastic source-risk drift). Suppose an adaptation procedure returns $\widehat{\theta}$ with $\widehat{D}_{\mathcal{R}}(\widehat{\theta}) \le \rho_{\mathrm{KL}} + \xi$, where $\xi \ge 0$ is the measured feasibility tolerance. For a stochastic predictor $\widehat{y} \sim p_\theta(\cdot \mid x)$ and any loss $\ell(\widehat{y}, y) \in [0, L]$, define $R_s(\theta) = \mathbb{E}_{(x,y)\sim P_s, \widehat{y}\sim p_\theta}[\ell(\widehat{y}, y)]$. On the event in Equation 5,

$$D_s(\widehat{\theta}) \le \rho_{\mathrm{KL}} + \xi + \Gamma_n, \qquad \left|R_s(\widehat{\theta}) - R_s(\theta_0)\right| \le L\sqrt{\frac{\rho_{\mathrm{KL}} + \xi + \Gamma_n}{2}}. \tag{6}$$

The proof is in Appendix A. The result is conditional: it does not certify optimization convergence, and Γ_n can be large when replay poorly covers the source distribution. It applies directly to i.i.d. replay sampled from P_s^X. The low-margin replay intervention below is deliberately biased, motivated by Corollary 1, and is not covered by the stated replay-to-population generalization guarantee.

Corollary 1 (Deterministic top-1 prediction). Let $h_0(x) = \arg\max_y p_0(y \mid x)$ and let $m_0(x)$ be the gap between the largest and second-largest source probabilities. For every $\gamma > 0$, the probability of a top-1 prediction change satisfies

$$\Pr[h_{\widehat{\theta}}(X) \ne h_0(X)] \le \Pr[m_0(X) \le \gamma] + \frac{\sqrt{2(\rho_{\mathrm{KL}} + \xi + \Gamma_n)}}{\gamma}. \tag{7}$$

Consequently, the absolute change in any deterministic source 0–1 risk is no larger than this disagreement probability. This statement concerns agreement with the source predictor, not agreement with unknown source labels.

3.3 BridgeOFT

At a selected layer interface, let C_s and C_t be regularized uncentered source and target activation second moments. In the primary classification setup, C_s uses 1,000 examples from the source-training split and C_t uses 512 examples from the target-training split, sampled independently of evaluation data. The Gaussian OT endpoint and interpolation are

$$T = C_t^{-1/2}\left(C_t^{1/2} C_s C_t^{1/2}\right)^{1/2} C_t^{-1/2}, \qquad H_\tau = (1-\tau)I + \tau T. \tag{8}$$

Thus T maps target-activation geometry toward the source covariance. With $Z_t = C_t^{1/2}$ and trainable blockwise skew-symmetric A_θ, the deployed transform is

$$M_{\tau,\theta} = H_\tau Z_t Q_{\tau,\theta} Z_t^{-1}, \qquad Q_{\tau,\theta} = \mathrm{Cayley}(\tau A_\theta). \tag{9}$$

Transport factors are frozen after calibration; only the orthogonal gauge and target head are trained. The continuation variable τ is scheduled, not learned; endpoint experiments set $\tau = 1$. Appendix B gives regularization and layer-placement details. Thus BridgeOFT is a structured PEFT parameterization, whereas Function Trust wraps an arbitrary parameterization.

4 Experimental Protocol

Primary tasks and models. We adapt TinyBERT from SST-2 to MRPC with matched seeds $\{13, 21, 42, 87, 123\}$ (Jiao et al., 2019; Wang et al., 2018). We compare head-only, LoRA, BridgeOFT, and full fine-tuning, with matched Trust/no-Trust pairs. A Qwen3-0.6B study repeats the frozen-backbone PEFT comparisons over seeds $\{13, 21, 42\}$ (Yang et al., 2025). The default replay cache has 1,024 source inputs and the default requested budget is $\rho_{\mathrm{KL}} = 0.01$; replay inputs are sampled offline from the source-training split and are not regenerated during adaptation.

Appendix B records the optimizer, learning rates, batches, epochs, LoRA configuration, calibration, and solver schedule for the primary runs.

Metrics and selection. Target quality is MRPC F1. Source forgetting is the drop in SST-2 validation accuracy from the source anchor. We also report full-validation source KL and prediction flips. Checkpoints are selected by target-validation F1 before source-retention evaluation. Values are means and sample standard deviations over seeds. Because $n = 5$ and $n = 3$ are small, uncertainty is descriptive; we do not claim statistical significance.

Table 1: Main matched-seed results (mean ± sample standard deviation). Forgetting is source SST-2 accuracy drop; lower is better.

Model	Method	n	MRPC F1↑	Forgetting↓	KL↓	Flip↓
TinyBERT	Head-only	5	0.8179 ± 0.0007	0.0000 ± 0.0000	0.0000 ± 0.0000	0.0000 ± 0.0000
TinyBERT	LoRA	5	0.8385 ± 0.0091	0.0438 ± 0.0213	0.0931 ± 0.0238	0.0764 ± 0.0197
TinyBERT	LoRA + Trust	5	0.8376 ± 0.0091	0.0144 ± 0.0029	0.0326 ± 0.0030	0.0365 ± 0.0036
TinyBERT	BridgeOFT	5	0.8645 ± 0.0063	0.1729 ± 0.0239	0.4374 ± 0.0482	0.2472 ± 0.0243
TinyBERT	BridgeOFT + Trust	5	0.8621 ± 0.0060	0.0344 ± 0.0068	0.1552 ± 0.0206	0.0839 ± 0.0047
TinyBERT	Full FT	5	0.8676 ± 0.0055	0.0878 ± 0.0607	0.2253 ± 0.1483	0.1466 ± 0.0735
TinyBERT	Full FT + Trust	5	0.8668 ± 0.0040	0.0300 ± 0.0151	0.0750 ± 0.0291	0.0612 ± 0.0225
Qwen3-0.6B	Head-only	3	0.8140 ± 0.0191	0.0000 ± 0.0000	0.0000 ± 0.0000	0.0000 ± 0.0000
Qwen3-0.6B	LoRA	3	0.8674 ± 0.0056	0.0822 ± 0.0339	0.3109 ± 0.0992	0.1181 ± 0.0323
Qwen3-0.6B	LoRA + Trust	3	0.8719 ± 0.0035	0.0176 ± 0.0083	0.0603 ± 0.0124	0.0336 ± 0.0065
Qwen3-0.6B	BridgeOFT	3	0.8297 ± 0.0032	0.0088 ± 0.0086	0.0170 ± 0.0098	0.0153 ± 0.0074
Qwen3-0.6B	BridgeOFT + Trust	3	0.8202 ± 0.0099	0.0038 ± 0.0043	0.0098 ± 0.0029	0.0092 ± 0.0034

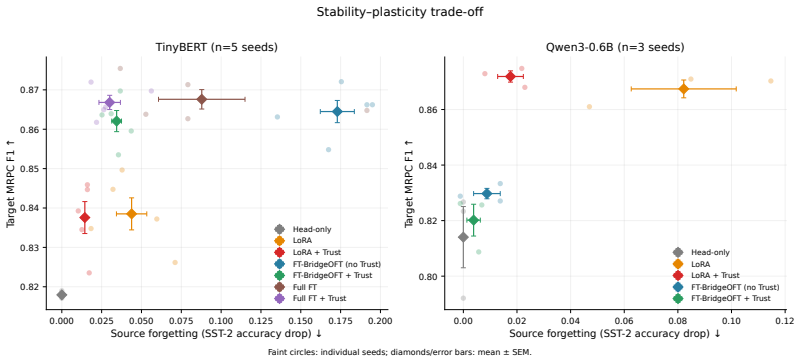


Figure 1: Target MRPC F1 versus source forgetting. Across all matched families, Function Trust shifts the operating point toward lower forgetting; the change in target quality depends on the model and parameterization.

Claim-directed extensions. A geometry-by-Trust factorial tests whether retention comes from covariance geometry or the functional constraint. Fixed-λ and ρ_{KL} sweeps test the stability–plasticity frontier. SST-2→RTE and MRPC→SST-2 test task-direction dependence. The extensions additionally vary replay coverage and composition, audit source margins and checkpoint feasibility, and probe Qwen replay scaling. Replay-versus-validation KL is treated as a finite held-out source diagnostic, not as a direct estimator of Γ_n. Additional synthetic and three-seed generative transfer validations appear in the appendix.

5 Results

5.1 Function Trust reduces drift across parameterizations

Table 1 shows that Function Trust lowers forgetting and full-validation KL in every matched family. On TinyBERT, forgetting falls from 0.0438 to 0.0144 for LoRA, from 0.1729 to 0.0344 for BridgeOFT, and from 0.0878 to 0.0300 for full fine-tuning; the corresponding changes in mean MRPC F1 are −0.0009, −0.0024, and −0.0008. For Qwen LoRA, Function Trust raises mean F1 by 0.0045 while reducing forgetting by 0.0646. We interpret these values as effect sizes rather than significance claims.

5.2 The functional constraint, not covariance geometry, explains retention

The seed-42 factorial holds the BridgeOFT parameter count fixed while varying covariance, random, and identity bases. Without Function Trust, forgetting is 0.1950, 0.1204, and 0.1032; with Function Trust, it falls to 0.0310, 0.0401, and 0.0344. Three-seed means under Function Trust are likewise close (covariance 0.0337, random 0.0432, identity 0.0352).

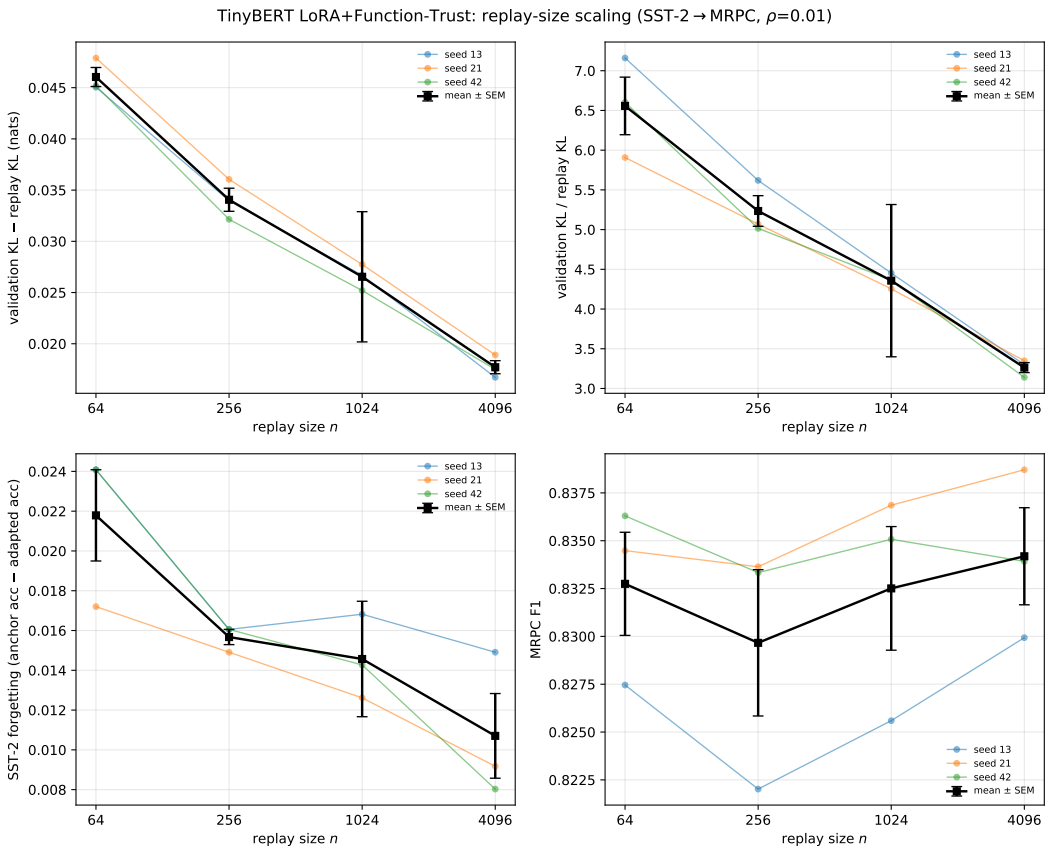


Figure 2: TinyBERT LoRA+Trust replay-size scaling on SST-2→MRPC at requested $\rho_{\mathrm{KL}} = 0.01$, using three matched seeds $\{13, 21, 42\}$ at every $n \in \{64, 256, 1024, 4096\}$. The black markers show mean ± SEM; colored curves show individual seeds. Increasing replay coverage reduces held-out source-validation KL, forgetting, and prediction flips in this tested configuration, while target F1 remains broadly stable. The validation–replay ratio is an empirical diagnostic, not an estimate of Γ_n.

Thus, within the tested configurations, geometry affects target behavior while the explicit functional constraint accounts for most of the retention gain. Appendix Table 4 reports the complete factorial.

5.3 Replay coverage determines the held-out meaning of the budget

The replay-size extension holds TinyBERT LoRA+Trust and the requested budget fixed at $\rho_{\mathrm{KL}} = 0.01$ while varying $n \in \{64, 256, 1024, 4096\}$ over matched seeds $\{13, 21, 42\}$. From $n = 64$ to $n = 4096$, mean held-out KL decreases from 0.0544 to 0.0255, forgetting from 0.0218 to 0.0107, and prediction flips from 0.0516 to 0.0313, while MRPC F1 remains between 0.8297 and 0.8342. Thus, the same requested replay-space budget can correspond to materially different held-out drift as coverage changes. This diagnostic illustrates the replay-generalization problem; it is neither an estimator of Γ_n nor a theoretical rate. Figure 2 reports per-seed values and mean ± SEM.

5.4 Replay composition protects functionally fragile regions

The deterministic margin corollary identifies low-margin source predictions as especially vulnerable to top-1 changes. The margin audit supports this mechanism empirically: pre-

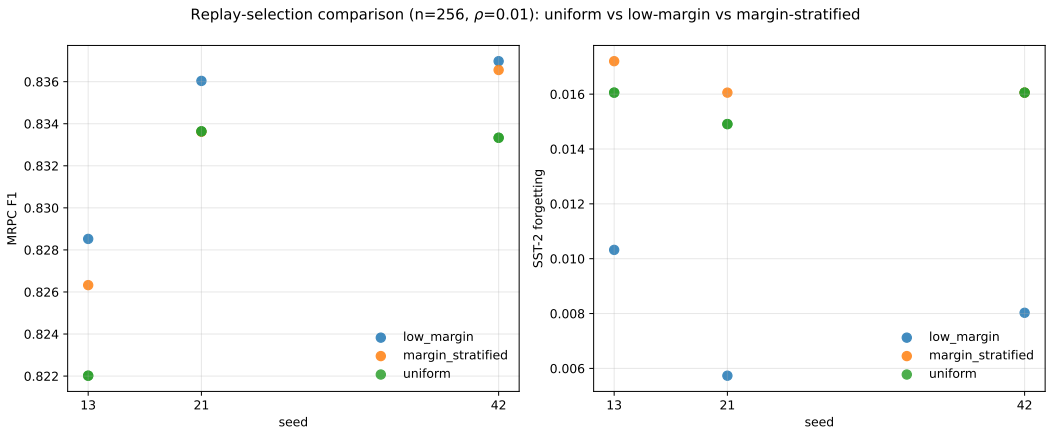


Figure 3: Replay-selection comparison at $n = 256$ and requested $\rho_{\mathrm{KL}} = 0.01$. Each point is a seed; low-margin replay improves retention metrics relative to uniform replay at the same cardinality, while margin-stratified replay is not uniformly better. The result motivates composition-aware replay but does not establish universal dominance or unbiased coverage.

diction flips are concentrated in low-margin bins across audited checkpoints. The continuous margin–KL relation is noisier; its Spearman correlations range from -0.49 to -0.78, so we do not treat them as validation of the theorem or as evidence of a calibrated KL relation. The intended sequence is Corollary 1 $\rightarrow$ low-margin fragility $\rightarrow$ empirical flip analysis $\rightarrow$ low-margin replay intervention.

We compare uniform, low-margin, and margin-stratified replay at $n = 256$ and $\rho_{\mathrm{KL}} = 0.01$ over three seeds. Relative to uniform replay, low-margin replay reduces mean held-out KL from 0.0421 to 0.0209, forgetting from 0.0157 to 0.0080, and flip rate from 0.0417 to 0.0264, while replay KL remains similar (0.0081 versus 0.0088). The direction is consistent across seeds, and target F1 changes from 0.8297 to 0.8338. Margin-stratified replay instead yields 0.0483 held-out KL and 0.0164 forgetting. Replay composition therefore matters in addition to cardinality: low-margin replay preferentially protects functionally fragile regions. The intervention is deliberately biased; it neither provides unbiased coverage of P_s nor directly minimizes Γ_n, and its advantage is not claimed to be universal.

5.5 Budget controllability, feasibility, and solver comparison

The three-seed LoRA+Trust budget study provides budget controllability rather than a calibration-equality claim. As requested ρ_{KL} decreases from 0.03 to 0.003, mean achieved replay KL decreases from 0.0204 to 0.0030, full-validation KL from 0.0467 to 0.0195, and forgetting from 0.0210 to 0.0107. The feasibility residual is $\xi = [\widehat{D}_{\mathcal{R}} - \rho_{\mathrm{KL}}]_+$, where $[z]_+ = \max(0, z)$: its mean is 0 at $\rho_{\mathrm{KL}} = 0.03$, 1.2×10^{-5} at 0.01, and 1.7×10^{-4} at 0.003, with selected-checkpoint feasibility rates of 3/3, 2/3, and 1/3, respectively. We therefore do not call $|\widehat{D}_{\mathcal{R}} - \rho_{\mathrm{KL}}|$ a conventional calibration error: the constraint is an upper bound, not an equality. Aggressive budgets can have small positive violations, so requested ρ_{KL} is not itself a certificate.

For the canonical five-seed TinyBERT settings at $\rho_{\mathrm{KL}} = 0.01$, selected-checkpoint feasibility is 4/5 for LoRA, 2/5 for BridgeOFT, and 4/5 for full fine-tuning. Mean (replay KL, residual ξ) is $(0.0085, 7.2 \times 10^{-6})$ for LoRA, $(0.0151, 0.0054)$ for BridgeOFT, and $(0.0098, 0.0017)$ for full fine-tuning. The appendix separates canonical settings from aggressive budget sweeps, sensitivity cells, and task extensions; heterogeneous family aggregates are not used as canonical solver-success rates. Function Trust therefore makes feasibility auditable, while the nonconvex PHR solver does not guarantee attainment of every requested budget.

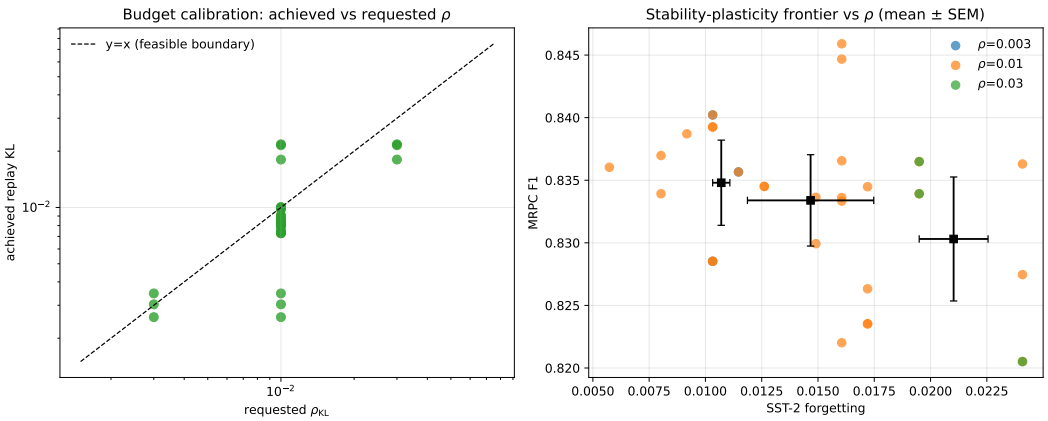


Figure 4: Three-seed requested-versus-achieved budget study for TinyBERT LoRA+Trust. Left: achieved replay KL against requested ρ_{KL} with the feasible boundary. Right: target F1 versus source forgetting, with mean ± SEM markers. The figure supports budget ordering and exposes residual violations; it does not treat achieved drift as equal to the request.

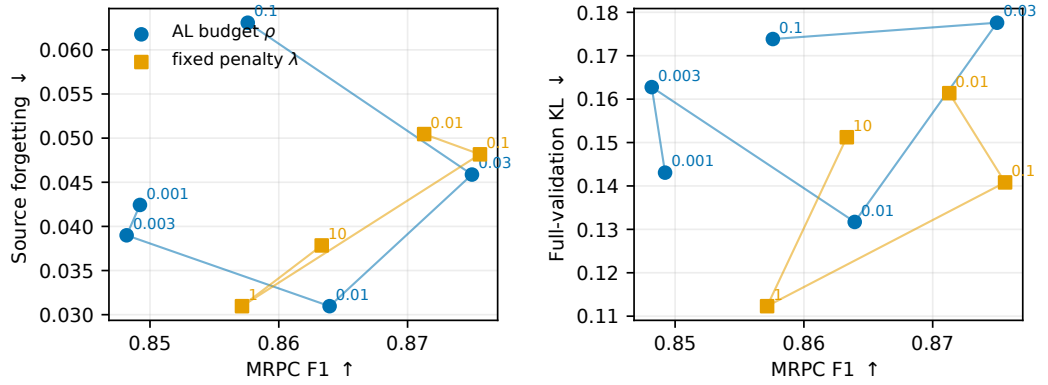


Figure 5: Exploratory seed-42 stability–plasticity sweeps. Labels are requested AL budgets ρ_{KL} (blue circles) or fixed penalty weights λ (orange squares). Lower-right is better. The non-monotone full-validation KL panel exposes replay-to-held-out mismatch; neither optimizer family dominates.

Figure 5 compares single-seed requested budgets ρ_{KL} with fixed penalties λ. The frontiers overlap. Fixed $\lambda = 0.1$ attains the best target F1 in this sweep (0.8756), narrowly above AL with $\rho_{\mathrm{KL}} = 0.03$ (0.8750). Fixed $\lambda = 1$ and AL with $\rho_{\mathrm{KL}} = 0.01$ both reach forgetting 0.0310, with different target scores. A three-seed confirmatory control gives the same qualitative conclusion: fixed $\lambda = 1$ has higher mean F1 than the default AL point at similar forgetting. Therefore, the contribution is not "AL beats a penalty." The advantage of Equation 1 is semantic: ρ_{KL} requests a drift budget, while λ specifies an indirect trade-off coefficient. Whether a particular solver realizes the request must be audited.

5.6 Replay feasibility does not guarantee held-out source preservation

A feasibility-aware checkpoint-selection audit reinforces the replay-generalization limitation. Among 69 audited Trust runs, target-only checkpoint selection was infeasible in 17 cases; restricting selection to feasible checkpoints corrected 11 of these, but reduced target F1 by method-dependent amounts (mean changes ranged from −0.0210 for full fine-tuning to

−0.0054 for LoRA in the headline audit), while average forgetting over corrected runs was unchanged or improved only marginally. Six runs had no feasible checkpoint at all. This is a negative diagnostic, not a proposed selection algorithm: enforcing replay feasibility does not by itself resolve replay-to-held-out mismatch, so $\widehat{D}_{\mathcal{R}} \leq \rho_{\mathrm{KL}} \not\Rightarrow$ small held-out source drift. It empirically illustrates the population-generalization issue represented by Γ_n in Theorem 1.

Replay–held-out mismatch. For the selected TinyBERT BridgeOFT+Trust checkpoint, replay-subset KL is 0.0170 while held-out source-validation mean KL is 0.1317; across the canonical audited checkpoints, replay KL understates held-out KL by roughly 6–21×. In the matched TinyBERT LoRA replay-size study, the held-out-to-replay KL ratio decreases from approximately 6.5× at $n = 64$ to 3.3× at $n = 4096$; these are ratios, not additive gaps. Held-out KL is heavy-tailed (median 0.0034, 99th percentile 0.806, maximum 3.39), with an 8.6% prediction-flip rate in the audited seed-42 checkpoint. This is an empirical diagnostic related to Γ_n in Theorem 1, not a direct estimate of it: a small empirical budget is only useful for held-out preservation when replay covers the protected distribution or the generalization term is controlled.

Operationally, one should sweep ρ_{KL} on a reserved source calibration split and choose the largest budget meeting the held-out tolerance, recalibrating after any change in replay, task, model, or adapter. Without source validation, replay budgets cannot be translated reliably into population-retention targets.

5.7 The stability–plasticity optimum is regime dependent

The Qwen BridgeOFT pair provides the clearest boundary case: unconstrained BridgeOFT is already stable, and adding Function Trust reduces forgetting from 0.0088 to 0.0038 but lowers F1 from 0.8297 to 0.8202. On SST-2→RTE, Function Trust benefits full fine-tuning but changes already-stable LoRA little; in the reverse MRPC→SST-2 pilot, head-only outperforms all adaptation methods. The appropriate operating point therefore depends on task direction, model scale, and parameterization. The budget ρ_{KL} makes this choice explicit rather than eliminating it.

6 Discussion and Limitations

Function Trust is most valuable when unconstrained adaptation is unstable, but can overconstrain an already stable adapter. BridgeOFT is complementary: its geometry is not a retention certificate and does not uniformly outperform LoRA or full fine-tuning. Replay coverage and composition both govern held-out drift; low-margin replay helps in the tested TinyBERT setting but remains biased.

The theorem is conditional on achieved feasibility, controlled KL complexity, and i.i.d. replay; it bounds stochastic risk, while deterministic classification also depends on margins. It therefore excludes biased low-margin replay. The nonconvex PHR solver is likewise no feasibility certificate: ρ_{KL} is a requested budget, and replay feasibility need not imply held-out preservation. Controlling Γ_n remains open.

The empirical scope is modest: five TinyBERT seeds, three canonical Qwen seeds, two forward tasks, one reverse pilot, and classification-centered primary experiments. These single-step adaptations do not test compounding drift along a multi-task trajectory. TinyBERT extensions use three matched seeds; new Qwen replay-size endpoints are single-seed diagnostics, not a scaling law. An SST-2 anchor cannot certify broad Qwen knowledge, and the generative transfer study uses validation subsets without a matched LoRA baseline. BridgeOFT sensitivity sweeps are exploratory. We also omit EWC and matched adapter-weight L_2 baselines.

Function Trust costs more than target-only fine-tuning: cached source logits avoid source-model inference, but replay forward/backward passes and activations add time and memory. Appendix Figure 12 reports system-dependent diagnostics.

References

Joshua Achiam, David Held, Aviv Tamar, and Pieter Abbeel. Constrained policy optimization. 2017. URL https://arxiv.org/abs/1705.10528v1.

Andrew Cotter, Maya Gupta, Heinrich Jiang, Nathan Srebro, Karthik Sridharan, Serena Wang, Blake Woodworth, and Seungil You. Training well-generalizing classifiers for fairness metrics and other data-dependent constraints. 2018a. URL https://arxiv.org/abs/1807.00028v2.

Andrew Cotter, Heinrich Jiang, and Karthik Sridharan. Two-player games for efficient non-convex constrained optimization. 2018b. URL https://arxiv.org/abs/1804.06500v2.

Bharath Bhushan Damodaran, Benjamin Kellenberger, Rémi Flamary, Devis Tuia, and Nicolas Courty. Deepjdot: Deep joint distribution optimal transport for unsupervised domain adaptation. 2018. URL https://arxiv.org/abs/1803.10081v3.

Ferdinando Fioretto, Pascal Van Hentenryck, Terrence W. K. Mak, Cuong Tran, Federico Baldo, and Michele Lombardi. Lagrangian duality for constrained deep learning. 2020. URL https://arxiv.org/abs/2001.09394v2.

Alex Gomez-Villa, Bartlomiej Twardowski, Lu Yu, Andrew D. Bagdanov, and Joost van de Weijer. Continually learning self-supervised representations with projected functional regularization. 2021. URL https://arxiv.org/abs/2112.15022v2.

Zeyu Han, Chao Gao, Jinyang Liu, Jeff Zhang, and Sai Qian Zhang. Parameter-efficient fine-tuning for large models: A comprehensive survey. 2024. URL https://arxiv.org/abs/2403.14608v7.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. 2015. URL https://arxiv.org/abs/1503.02531v1.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin de Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for nlp. 2019. URL https://arxiv.org/abs/1902.00751v2.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. In International Conference on Learning Representations, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Yiming Huang et al. Context-free synthetic data mitigates forgetting. 2025. URL https://arxiv.org/abs/2505.13811v1.

Xiaoqi Jiao, Yichun Yin, Lifeng Shang, Xin Jiang, Xiao Chen, Linlin Li, Fang Wang, and Qun Liu. Tinybert: Distilling bert for natural language understanding. 2019. URL https://arxiv.org/abs/1909.10351v5.

Ce Ju and Cuntai Guan. Deep optimal transport for domain adaptation on spd manifolds. 2022. URL https://arxiv.org/abs/2201.05745v6.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, Demis Hassabis, Claudia Clopath, Dharshan Kumaran, and Raia Hadsell. Overcoming catastrophic forgetting in neural networks. Proceedings of the National Academy of Sciences, 114(13):3521–3526, 2017. URL https://arxiv.org/abs/1612.00796v2.

Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pp. 4582–4597. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.acl-long.353. URL https://aclanthology.org/2021.acl-long.353/.

Zhizhong Li and Derek Hoiem. Learning without forgetting. 2016. URL https://arxiv.org/abs/1606.09282v3.

Weiyang Liu, Zeju Qiu, Yao Feng, Yuliang Xiu, Yuxuan Xue, Longhui Yu, Haiwen Feng, Zhen Liu, Juyeon Heo, Songyou Peng, Yandong Wen, Michael J. Black, Adrian Weller, and Bernhard Schölkopf. Parameter-efficient orthogonal finetuning via butterfly factorization. 2023. URL https://arxiv.org/abs/2311.06243v2.

Yun Luo, Zhen Yang, Fandong Meng, Yafu Li, Jie Zhou, and Yue Zhang. An empirical study of catastrophic forgetting in large language models during continual fine-tuning. 2023. URL https://arxiv.org/abs/2308.08747v5.

Xinyu Ma, Xu Chu, Zhibang Yang, Yang Lin, Xin Gao, and Junfeng Zhao. Parameter efficient quasi-orthogonal fine-tuning via givens rotation. 2024. URL https://arxiv.org/abs/2404.04316v2.

Zeju Qiu, Weiyang Liu, Haiwen Feng, Yuxuan Xue, Yao Feng, Zhen Liu, Dan Zhang, Adrian Weller, and Bernhard Schölkopf. Controlling text-to-image diffusion by orthogonal finetuning. 2023. URL https://arxiv.org/abs/2306.07280v3.

John Schulman, Sergey Levine, Philipp Moritz, Michael I. Jordan, and Pieter Abbeel. Trust region policy optimization. In Proceedings of the 32nd International Conference on Machine Learning, volume 37 of Proceedings of Machine Learning Research, pp. 1889–1897. PMLR, 2015. URL https://proceedings.mlr.press/v37/schulman15.html.

Sidak Pal Singh and Martin Jaggi. Model fusion via optimal transport. 2019. URL https://arxiv.org/abs/1910.05653v6.

Michalis K. Titsias, Jonathan Schwarz, Alexander G. de G. Matthews, Razvan Pascanu, and Yee Whye Teh. Functional regularisation for continual learning with gaussian processes. 2020. URL https://arxiv.org/abs/1901.11356v4.

Edoardo Urettini and Antonio Carta. Online curvature-aware replay: Leveraging 2nd order information for online continual learning. 2025. URL https://arxiv.org/abs/2502.01866v1.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. GLUE: A multi-task benchmark and analysis platform for natural language understanding. 2018. URL https://arxiv.org/abs/1804.07461v3.

Xinyu Wang, Changzhi Sun, Yuanbin Wu, and Xiaoling Wang. Stabilizing llm supervised fine-tuning via explicit distributional control. 2026a. URL https://arxiv.org/abs/2605.04468v1.

Zekun Wang, Anant Gupta, and Christopher J. MacLellan. Trust region continual learning as an implicit meta-learner. 2026b. URL https://arxiv.org/abs/2602.02417v2.

Mitchell Wortsman, Gabriel Ilharco, Samir Yitzhak Gadre, Rebecca Roelofs, Raphael Gontijo-Lopes, Ari S. Morcos, Hongseok Namkoong, Ali Farhadi, Yair Carmon, Simon Kornblith, and Ludwig Schmidt. Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. In Proceedings of the 39th International Conference on Machine Learning, volume 162 of Proceedings of Machine Learning Research, pp. 23965–23998. PMLR, 2022. URL https://proceedings.mlr.press/v162/wortsman22a.html.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. 2025. URL https://arxiv.org/abs/2505.09388v1.

Shen Yuan, Haotian Liu, and Hongteng Xu. Bridging the gap between low-rank and orthogonal adaptation via householder reflection adaptation. 2024. URL https://arxiv.org/abs/2405.17484v3.

Yunan Zhang, Shuoran Jiang, Mengchen Zhao, Yuefeng Li, Yang Fan, Xiangping Wu, and Qingcai Chen. Gere: Towards efficient anti-forgetting in continual learning of llm via general samples replay. 2025. URL https://arxiv.org/abs/2508.04676v1.

Ethics statement

This work studies adaptation on standard language-understanding benchmarks and introduces no new personal dataset or user-facing deployment. The main risk is overstating a local replay constraint as global preservation. We mitigate this by reporting held-out source audits, negative cases, solver residual caveats, and separate claims for target quality and retention.

Reproducibility statement

The repository records seeds, checkpoint semantics, trainable-parameter counts, environment metadata, and audit ledgers. The repository-relative extension manifest and aggregate result files record replay-size, replay-selection definitions, requested budgets, achieved replay KL, residual ξ, feasibility, margin, Qwen, and BridgeOFT sensitivity runs; the generative-transfer archive additionally provides the pinned data configuration, per-seed records, and machine-readable aggregates. Figure-generation scripts regenerate the extension figures. Tables use mean $\pm$ sample standard deviation unless a caption states SEM; figures show individual seeds and mean $\pm$ SEM when stated. The paper tables derive from the manifest and current reporting scripts. A local verification run passed 394 tests with 1 skip.

AI use statement

Generative AI tools assisted with literature discovery, manuscript drafting and editing, and figure formatting. The authors manually verified the cited literature, quantitative results, mathematical statements, and AI-assisted revisions. The authors take responsibility for the final text, claims, and artifacts.

A Proofs

Proof of Theorem 1. On the uniform-deviation event,

$$D_s(\widehat{\theta}) \le \widehat{D}_{\mathcal{R}}(\widehat{\theta}) + \Gamma_n \le \rho_{\mathrm{KL}} + \xi + \Gamma_n.$$

For fixed (x, y), the difference between expected losses under $p_{\widehat{\theta}}(\cdot \mid x)$ and $p_0(\cdot \mid x)$ is at most L times their total-variation distance. Taking expectations, applying Pinsker's inequality pointwise, and then Jensen's inequality gives

$$\begin{aligned}\left|R_s(\widehat{\theta}) - R_s(\theta_0)\right| &\le L\,\mathbb{E}_X\,\mathrm{TV}(p_0, p_{\widehat{\theta}}) \\ &\le L\,\mathbb{E}_X\sqrt{\tfrac{1}{2}\mathrm{KL}(p_0 \| p_{\widehat{\theta}})} \\ &\le L\sqrt{D_s(\widehat{\theta})/2} \\ &\le L\sqrt{(\rho_{\mathrm{KL}} + \xi + \Gamma_n)/2}.\end{aligned}$$

□

Proof of Corollary 1. If the adapted top-1 prediction differs from the source prediction at x, then changing the ordering of the two leading classes requires $\mathrm{TV}(p_0, p_{\widehat{\theta}}) \ge m_0(x)/2$. Split inputs according to $m_0(x) \le \gamma$. On the complementary set, $\mathbf{1}\{h_{\widehat{\theta}}(x) \ne h_0(x)\} \le 2\,\mathrm{TV}(p_0, p_{\widehat{\theta}})/\gamma$. Taking expectations, then applying Pinsker, Jensen, and Theorem 1, yields Equation 7; the final risk statement follows because two deterministic classifiers can differ in 0–1 loss only on their disagreement set. □

B Function Trust extension details

Canonical implementation. The source anchor is trained for three epochs with AdamW (learning rate 2×10^{-5}, batch size 32 for TinyBERT and 8 for Qwen, weight decay 0.01). Primary target runs use AdamW for five epochs, maximum length 128, gradient clipping at

1.0, and batch size 16 (TinyBERT) or 8 (Qwen). The target head and LoRA use learning rate 10^{-4}; BridgeOFT gauge/radial parameters use $10^{-3}/5 \times 10^{-4}$; full-backbone updates use 2×10^{-5}. LoRA uses rank 8, scale 16, dropout 0.1, and query/key/value projections. For calibration, we sample 1,000 SST-2 training examples with the run seed and 512 target-training examples with seed plus one. At every TinyBERT attention/feed-forward input and the last four Qwen attention/MLP inputs, we accumulate token-masked uncentered moments in float64 and form contiguous width-32 blocks. For raw block C, regularization applies $C \leftarrow C + 10^{-6} I$, then $C \leftarrow 0.95C + 0.05\,\mathrm{tr}(C)I/d$, followed by eigenvalue flooring at $\max(10^{-7}, 10^{-4}\bar{\lambda})$. Factors remain static; continuation is linear from $\tau = 0$ to 1 and is not learned. The default PHR state is $(\mu_0, \beta_0) = (0, 10)$; each step updates $\bar{g} \leftarrow 0.9\bar{g} + 0.1g$, and every 50 steps applies $\mu \leftarrow [\mu + \beta\bar{g}]_+$. If the positive violation exceeds 90% of its previous value, β doubles up to 1,000. These values are fixed across canonical matched comparisons.

Table 2 reports the three-seed aggregate extension results. Entries are means over seeds with sample standard deviation; figures retain mean ± SEM where explicitly stated.

Table 2: Function Trust extensions on TinyBERT LoRA+Trust, SST-2→MRPC. Values are means ± sample standard deviation over three matched seeds.

Setting	MRPC F1↑	Forgetting↓	Val. KL↓	Flip↓	Replay KL↓
$n = 64$	0.8327 ± 0.0047	0.0218 ± 0.0040	0.0544 ± 0.0028	0.0516 ± 0.0023	0.0084 ± 0.0013
$n = 256$	0.8297 ± 0.0066	0.0157 ± 0.0007	0.0421 ± 0.0025	0.0417 ± 0.0043	0.0081 ± 0.0008
$n = 1024$	0.8324 ± 0.0081	0.0134 ± 0.0035	0.0321 ± 0.0025	0.0371 ± 0.0035	0.0089 ± 0.0010
$n = 4096$	0.8342 ± 0.0044	0.0107 ± 0.0037	0.0255 ± 0.0015	0.0313 ± 0.0046	0.0078 ± 0.0005
uniform, $n = 256$	0.8297 ± 0.0066	0.0157 ± 0.0007	0.0421 ± 0.0025	0.0417 ± 0.0043	0.0081 ± 0.0008
low-margin, $n = 256$	0.8338 ± 0.0046	0.0080 ± 0.0023	0.0209 ± 0.0036	0.0264 ± 0.0040	0.0088 ± 0.0004
margin-stratified, $n = 256$	0.8322 ± 0.0053	0.0164 ± 0.0007	0.0483 ± 0.0046	0.0447 ± 0.0030	0.0084 ± 0.0013
$\rho_{\mathrm{KL}} = 0.003$	0.8348 ± 0.0059	0.0107 ± 0.0007	0.0195 ± 0.0006	0.0291 ± 0.0018	0.0030 ± 0.0004
$\rho_{\mathrm{KL}} = 0.010$	0.8324 ± 0.0081	0.0134 ± 0.0035	0.0321 ± 0.0025	0.0371 ± 0.0035	0.0089 ± 0.0010
$\rho_{\mathrm{KL}} = 0.030$	0.8303 ± 0.0086	0.0210 ± 0.0026	0.0467 ± 0.0025	0.0455 ± 0.0035	0.0204 ± 0.0021

Margin and feasibility audits. The margin analysis uses $m_0(x) = p_{0,\mathrm{top1}}(x) - p_{0,\mathrm{top2}}(x)$ on 872 SST-2 validation examples per checkpoint; the supplementary artifacts provide the complete binned flip/KL curves and Spearman table. The feasibility audit defines $\xi = [\widehat{D}_{\mathcal{R}} - \rho_{\mathrm{KL}}]_+$ on the selected checkpoint. We report canonical/default settings separately from aggressive ρ_{KL} sweeps, BridgeOFT block/calibration sensitivity, and other task extensions: canonical TinyBERT $\rho_{\mathrm{KL}} = 0.01$ rates are LoRA 4/5, BridgeOFT 2/5, and full fine-tuning 4/5; the aggressive replay=1024 LoRA sweep gives 3/3, 2/3, and 1/3 at $\rho_{\mathrm{KL}} = 0.03, 0.01, 0.003$, respectively. These stratified results should not be interpreted as a single family-wide solver-success rate. The feasibility-aware selection audit is a negative diagnostic documenting that local feasibility alone does not certify held-out source retention.

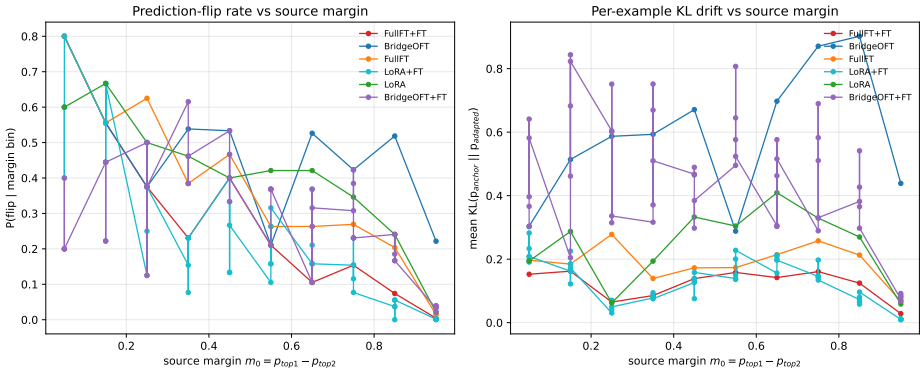


Figure 6: Empirical validation of the margin mechanism. Left: prediction-flip rate by source-margin bin. Right: per-example KL drift by source-margin bin. The discrete flip relation is more directly interpretable for Corollary 1; the continuous KL relation is noisier.

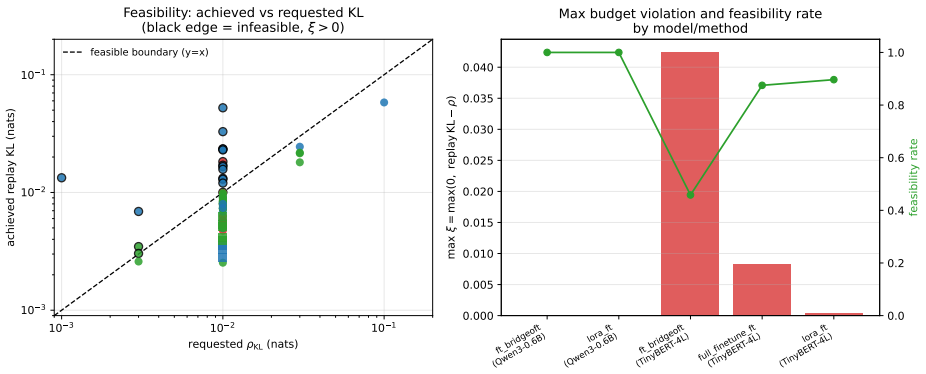


Figure 7: Feasibility audit. Left: selected-checkpoint achieved replay KL versus requested budget; black edges indicate positive ξ. Right: maximum residual and feasibility rate stratified by model/method. Heterogeneous families and aggressive budgets are kept separate rather than summarized by one overall rate.

Qwen replay-size diagnostic. For Qwen3-0.6B LoRA+Trust, the new $n = 256$ and 4096 points are seed-42 runs, while $n = 1024$ is the three-seed canonical anchor. The corresponding MRPC F1 values are 0.8675/0.8719/0.8609, forgetting is 0.0149/0.0176/0.0092, validation KL is 0.0601/0.0603/0.0300, and flip rate is 0.0356/0.0336/0.0206 for $n = 256/1024/4096$, respectively. The larger replay point reduces measured drift but also lowers target F1; because only the new sizes are single-seed, this is a qualitative stability–plasticity check, not a scaling law.

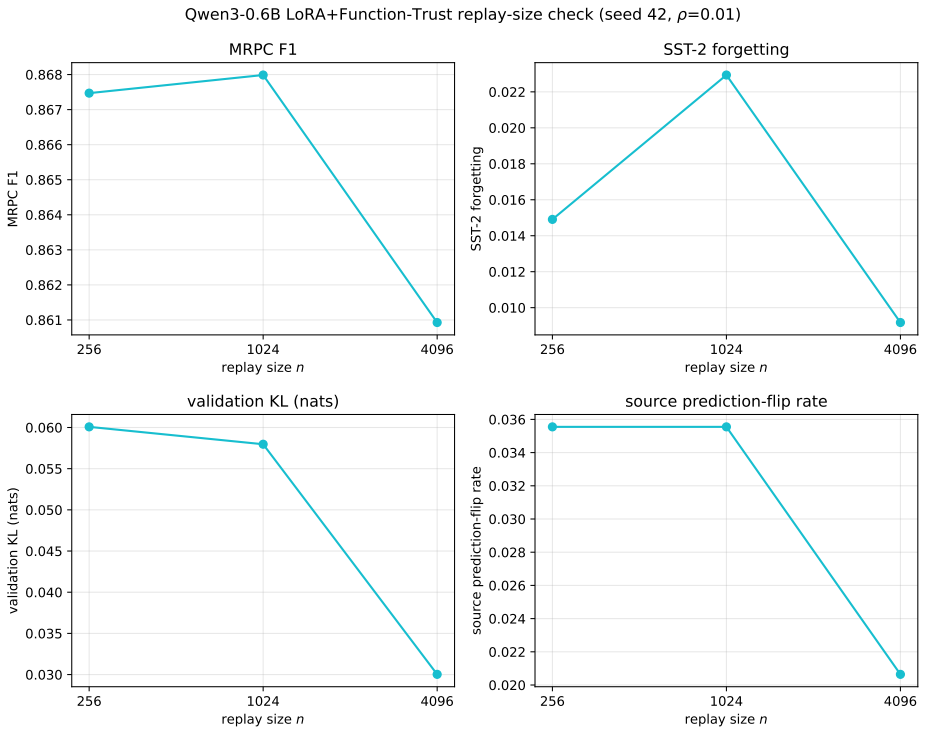


Figure 8: Qwen3-0.6B LoRA+Trust replay-size diagnostic: new $n = 256$ and 4096 points are seed 42, while $n = 1024$ is the three-seed canonical anchor. The larger replay cache reduces measured source drift but incurs a target-score trade-off; the result is cross-model qualitative evidence only.

BridgeOFT sensitivity. In seed-42 TinyBERT sensitivity runs, block sizes $8, 16, 32, 64$ yield MRPC F1 $0.8670, 0.8703, 0.8639, 0.8209$, forgetting $0.0229, 0.0195, 0.0310, 0.0619$, validation KL $0.0576, 0.0854, 0.1317, 0.2928$, and trainable parameters $23{,}026, 42{,}290, 80{,}434, 155{,}186$. Mean condition diagnostics increase from 13.9 to 89.8 across these configurations. Larger blocks coincide with poorer conditioning and greater functional drift in this sensitivity study; this is an observed association, not a causal theorem, and does not establish block size 8 as universally optimal. Calibration-size runs $(256, 1024, 4096)$ show no monotone effect and remain exploratory.

C Budget and fixed-penalty sweeps

Table 3: Exploratory seed-42 budget and fixed-penalty sweep at target-selected checkpoints. Validation KL is measured on the full source validation set, not the replay cache.

Solver	Control	MRPC F1↑	Forgetting↓	Val. KL↓
AL	$\rho_{\mathrm{KL}} = 0.001$	0.8492	0.0424	0.1431
AL	$\rho_{\mathrm{KL}} = 0.003$	0.8482	0.0390	0.1628
AL	$\rho_{\mathrm{KL}} = 0.010$	0.8639	0.0310	0.1317
AL	$\rho_{\mathrm{KL}} = 0.030$	0.8750	0.0459	0.1776
AL	$\rho_{\mathrm{KL}} = 0.100$	0.8576	0.0631	0.1738
Fixed	$\lambda = 0.01$	0.8713	0.0505	0.1614
Fixed	$\lambda = 0.1$	0.8756	0.0482	0.1408
Fixed	$\lambda = 1$	0.8571	0.0310	0.1123
Fixed	$\lambda = 10$	0.8633	0.0378	0.1512

The requested AL budget is not a certificate. At the end of training, the logged minibatch KL is 0.0044 for requested $\rho_{\mathrm{KL}} = 0.001$, so this run remains infeasible at the measured step; the default $\rho_{\mathrm{KL}} = 0.01$ run logs 0.0069. Penalty saturation and checkpoint selection can further separate a requested budget from achieved replay and held-out source drift. This is why the main paper reports ξ and held-out source audits rather than treating ρ_{KL} as an automatically satisfied guarantee.

D Geometry-by-Trust factorial

Table 4: Seed-42 geometry factorial with matched trainable parameter count. Covariance geometry alone is not the retention mechanism.

Basis	Constraint	MRPC F1↑	Forgetting↓
Identity	none	0.8688	0.1032
Random orthogonal	none	0.8618	0.1204
Covariance	none	0.8662	0.1950
Identity	Trust	0.8576	0.0344
Random orthogonal	Trust	0.8558	0.0401
Covariance	Trust	0.8639	0.0310

E Confirmatory fixed-KL and second-task extensions

F Additional regime checks

Reverse task. In the seed-42 MRPC→SST-2 pilot, "corrected" denotes feasibility-aware checkpoint selection rather than a change to the task or metric. Head-only reaches target accuracy 0.8088 with no source forgetting; LoRA reaches 0.8039 with 0.0049 forgetting, BridgeOFT +Trust 0.7966 with 0.0123, and full fine-tuning 0.7917 with 0.0172. The forward-direction target advantage does not transfer.

Table 5: Three-seed extensions. Target is MRPC F1 for fixed-KL and RTE accuracy for SST-2→RTE.

Experiment	Method	n	Target↑	Forgetting↓
Fixed-KL	adaptive AL	3	0.8589	0.0348
Fixed-KL	fixed $\lambda = 0.1$	3	0.8618	0.0489
Fixed-KL	fixed $\lambda = 1.0$	3	0.8664	0.0359
Fixed-KL	no Trust	3	0.8671	0.1674
SST-2→RTE	head-only	3	0.5355	0.0000
SST-2→RTE	LoRA	3	0.6221	0.0065
SST-2→RTE	LoRA+Trust	3	0.6197	0.0061
SST-2→RTE	BridgeOFT	3	0.5836	0.1070
SST-2→RTE	BridgeOFT+Trust	3	0.5981	0.0245
SST-2→RTE	Full FT	3	0.6450	0.0248
SST-2→RTE	Full FT+Trust	3	0.6522	0.0096

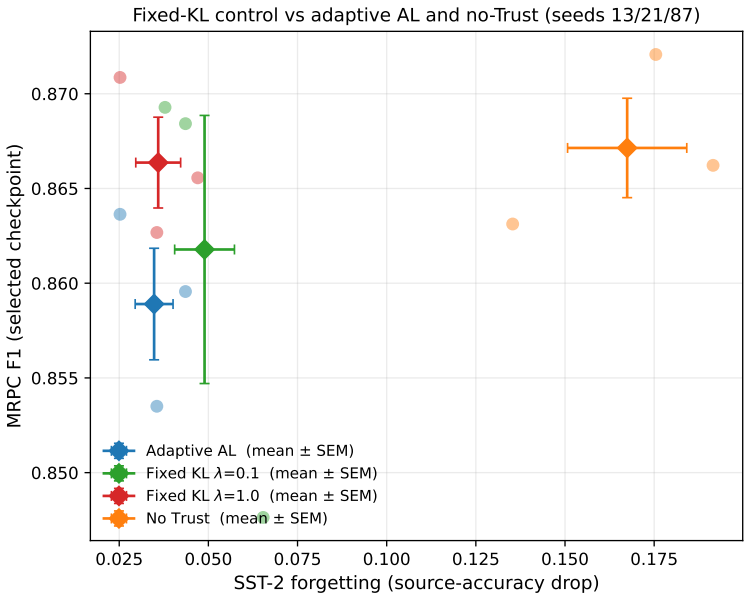


Figure 9: Three-seed fixed-penalty control. Points are seeds; diamonds show mean ± SEM. Fixed $\lambda = 1$ and adaptive AL have similar forgetting, while the fixed penalty has higher target F1 in this small comparison.

Qwen generative transfer. We further adapt Qwen3-0.6B from GSM8K mathematics to English–German translation. Each trainable method uses 126,976 adapter parameters with the backbone frozen, 1,000 optimizer steps at learning rate 3×10^{-4}, effective batch size 16, maximum sequence length 512, and 50,000 target-training examples. Both capabilities use fixed 256-example validation subsets. Pure OT is the deterministic target-to-source transport endpoint; hard BridgeOFT adds its trainable blockwise Cayley gauge; metric OFT learns the same gauge in the source metric without target-to-source transport; and Euclidean OFT replaces that metric by the identity. The last is an activation-space comparator, not weight-space OFTv2. Table 6 reports mean ± SEM over seeds $\{13, 21, 42\}$. Using the displayed means, Euclidean and metric OFT improve target chrF by 15.83 and 15.57 points over the frozen identity model. They also increase source exact match by 2.99 and 4.69 percentage points, respectively. This positive transfer may reflect a regularizing effect of the restricted gauge or evaluation variance; because there are only three seeds and 256 validation examples, we do not interpret it as evidence that translation training improves mathematics. Hard BridgeOFT improves target chrF by 13.88 points but loses 1.69 source points, while pure OT degrades both metrics. The 0.27 chrF gap between Euclidean and metric OFT is

Table 6: Qwen3-0.6B GSM8K→English–German transfer (mean ± SEM, three seeds). Source Δ is the change in GSM8K exact match relative to the frozen identity model.

Method	Translation chrF↑	GSM8K EM↑	Source Δ (pp)↑
Identity	19.97 ± 0.00	18.36 ± 0.00	0.00
Pure OT	16.95 ± 0.00	15.62 ± 0.00	−2.73
Euclidean OFT	$\mathbf{35.80} \pm 0.43$	21.35 ± 0.57	+2.99
Metric OFT	35.54 ± 0.28	$\mathbf{23.05} \pm 0.90$	**+4.69**
Hard BridgeOFT	33.85 ± 0.34	16.67 ± 0.13	−1.69

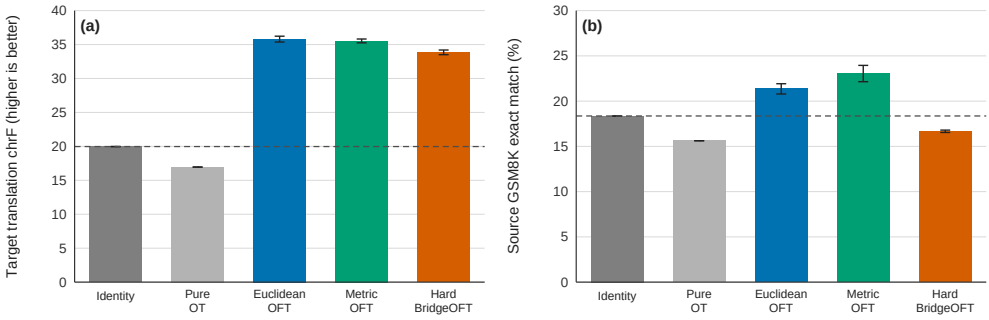


Figure 10: Qwen3-0.6B generative transfer from GSM8K mathematics to English–German translation. Bars show mean ± SEM over three seeds; dashed lines mark the frozen identity baseline. The learned Euclidean and metric gauges improve target translation while preserving source mathematics in this validation protocol, whereas deterministic OT alone degrades both capabilities.

small relative to their uncertainty. These validation results show that a learned gauge is necessary in this setting; they do not establish a universal winner or superiority over LoRA, which was not included in this matched protocol.

G Synthetic BridgeOFT validation

Synthetic experiments isolate BridgeOFT geometry from language-model confounds. The suite contains 28 canonical hard-BridgeOFT runs, 11 block-size runs, 5 calibration-size runs, 5 schedule runs, 24 low-dimensional runs, and 4 approximation guards. The maximum endpoint effective-metric residual is 6.85×10^{-15}. The median held-out task-loss improvement of learned BridgeOFT over the $Q = I$ bridge control is 0.3936, while the minimum is 0, so the learned gauge is useful in typical synthetic instances but not guaranteed to improve every case.

The canonical generator samples noncommuting source and target Gaussian covariances in $d = 128$ with condition number 10^3, target Bures distance 3, and 1,000 samples for covariance estimation. A frozen linear readout receives target-distributed inputs; regression labels are produced by a teacher transform $Z_s Q_\star Z_t^{-1}$ and additive Gaussian training noise of standard deviation 0.05. Runs use 4,096 training and 2,048 held-out examples, block width 32, and mean-squared error, so only the packed Cayley gauge is optimized. The block-size, calibration-size, schedule, low-dimensional, and numerical-approximation suites vary one of these ingredients while retaining the corresponding controls.

Table 7: Mean paired effect of enabling Function Trust. Positive forgetting reduction means better retention.

Model	Family	n	ΔF1 pp	Forgetting red. pp	KL red.
TinyBERT	LoRA	5	-0.09	2.94	0.060
TinyBERT	BridgeOFT	5	-0.24	13.85	0.282
TinyBERT	Full FT	5	-0.08	5.78	0.150
Qwen3-0.6B	LoRA	3	0.45	6.46	0.251
Qwen3-0.6B	BridgeOFT	3	-0.95	0.50	0.007

H Per-family Trust effects

I Additional diagnostics

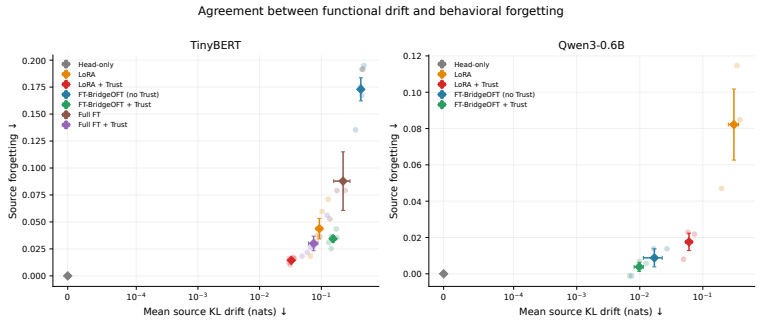


Figure 11: Full-validation source KL versus source-accuracy loss across canonical checkpoints. Their qualitative alignment supports KL as a continuous diagnostic of functional drift; it does not make KL a calibrated surrogate for every source task.

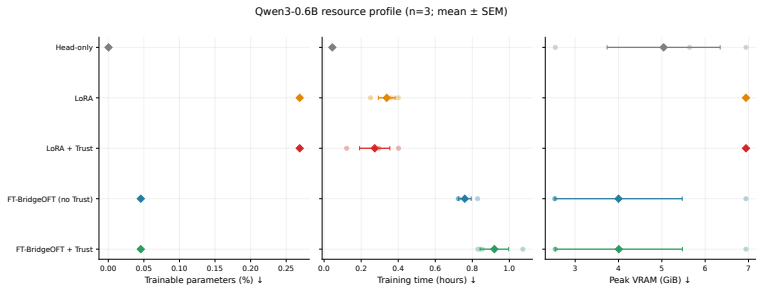


Figure 12: Qwen3-0.6B trainable fraction, wall-clock time, and recorded peak VRAM. Values are system-dependent and are included only as within-environment implementation diagnostics.