

Beyond One-Hot Replay: Self-Distillation Reduces Target Variance in Continual Pre-Training

Peng Lu, Sijie Dong, Xiaoyun Feng, Xuanwei Hu, Xin Zhang*

Ant International

{yahao.ch, evan.zx}@ant-intl.com

Abstract

Continual pre-training (CPT) adapts a pre-trained language model to a target domain, but it often degrades general capabilities even when general-data replay is used to mitigate forgetting. We revisit this failure through replay supervision: prior CPT replay mainly tunes data mixtures and token budgets while applying one-hot next-token prediction (NTP) uniformly to domain and replay tokens. This target is natural for learning domain text, but on replay tokens it introduces target variance: it is a noisy, corpus-dependent proxy for preserving the frozen base model’s conditional behavior. We study *self-distilled replay*: domain tokens keep NTP, while replay tokens are trained to match the frozen base model distribution. Diagnostics show that this target substitution lowers target error, produces more directionally consistent replay gradients, and reduces base-behavior drift. Empirically, self-distilled replay improves the SmoLM3 domain-general trade-off over matched NTP replay across diverse token budgets and replay ratios. It also remains effective for Qwen3 models whose pretraining mixtures are unavailable: under the same mismatched, non-source replay setting, it improves both general score and PubMedQA in all eight matched comparisons. Overall, these results suggest supervising CPT replay to preserve base behavior, rather than fit empirical next tokens, as a way to reduce target variance.

1 Introduction

Continual pre-training (CPT) is a common paradigm to adapt a pretrained language model to specialized domains (e.g., biomedical, legal, or code) without training from scratch (Gururangan et al., 2020; Ke et al., 2023; Cheng et al., 2024; Jiang et al., 2025). However, distribution shift during CPT can induce catastrophic forgetting and degrade general capabilities (Li and Lee,

2024; Wang et al., 2023; Ibrahim et al., 2024; Li et al., 2025). This poses a fundamental stability-plasticity dilemma: an effective CPT recipe must learn the target distribution while retaining the behavior of the base model.

A widely used and scalable mitigation is replay. By mixing target-domain data with general data, the optimizer repeatedly revisits contexts representative of the original training distribution (Gupta et al., 2023; Ibrahim et al., 2024; Li et al., 2025; Abbes et al., 2026; Que et al., 2024; Ye et al., 2025). Much prior CPT replay work focuses on data mixtures and token budgets while applying the same next-token prediction (NTP) objective uniformly. We argue this overlooks a critical distinction in their purposes: domain tokens are used to *acquire* new knowledge, whereas replay tokens are used to *preserve* existing behavior. Under realistic compute and data budget constraints, we must maximize the preservation yield of every replay token.

Standard NTP replay is an indirect preservation mechanism. The behavior we aim to preserve is the frozen base model’s full next-token predictive distribution on replay contexts. However, NTP supervises the updated student model with a single observed token (a one-hot target) at each context. When the base model assigns probability mass to multiple plausible continuations (e.g., synonyms or equally valid reasonings), fitting a single sampled token discards the rich distributional information defining the base behavior. Consequently, one-hot replay introduces target variance: it is a noisy proxy for the true preservation target, and finite replay batches translate that variance into noisy gradient updates rather than a stable estimate of the base distribution.

To address this, we study **self-distilled replay** as an asymmetric yet minimal intervention for CPT supervision. Soft-label and distillation-based replay are well-established in continual learning

*Corresponding author.

(Li and Hoiem, 2018; Buzzega et al., 2020), and CPT methods such as Lifelong Pretraining and Mix-CPT already use distillation-style regularization (Jin et al., 2022; Jiang et al., 2025). We build on this line of work and study a more specific design choice: the objective assigned to each token should follow the distribution that token is meant to match. In our setup, the student model continues to use standard NTP on domain tokens to learn new distributions. On replay tokens, however, it replaces the high-variance one-hot target with a Kullback-Leibler (KL) divergence objective matching the frozen base model’s distribution. The teacher is simply the base checkpoint itself—requiring no external specialized models, with computational overhead limited by skipping the teacher forward pass on domain-only micro-batches.

Empirically, this target substitution improves the observed domain–general trade-off in matched CPT experiments. On SmoLM3-3B (Bakouch et al., 2025), across 11 matched comparisons spanning different domain–replay ratios and total token budgets, self-distilled replay improves PubMedQA in 10 comparisons and the general score in 9, with average gains of 4.33 and 1.16 points over matched NTP replay; NTP replay never dominates it on both axes. Crucially, because the supervision signal comes directly from the frozen base model rather than the observed tokens, the method reduces dependence on exact replay-corpus alignment. On Qwen3 (Qwen Team, 2025), where the original pretraining mixture is proprietary and unavailable, we compare both methods under the same mismatched, non-source replay setting; self-distilled replay improves both general score and PubMedQA in all eight matched comparisons over NTP replay, with average gains of 5.41 and 3.52 points.

In summary, our main contributions are twofold. (1) We show that one-hot NTP replay is a high-variance proxy for preserving the frozen base distribution: it supervises each replay context with a single observed token rather than the full base-model distribution, creating target error that propagates into less consistent replay gradients and larger behavior drift. Self-distilled replay replaces this proxy with a KL objective against the frozen base model, reducing target error, improving gradient consistency, and lowering drift. (2) We provide systematic matched comparisons between self-distilled replay and NTP replay under the

same training budgets, domain–replay ratios, and data/training settings. Across SmoLM3 source-replay sweeps and Qwen3 non-source replay settings, self-distilled replay improves the domain–general trade-off over NTP replay, with domain performance measured by PubMedQA and general capability measured by a broad benchmark suite, and remains effective when the original pretraining mixture is proprietary or unavailable.

2 Replay Supervision

In standard CPT, a student model p_θ (initialized from a frozen base checkpoint $p_{\theta_{\text{base}}}$) is optimized jointly on the target-domain corpus $\mathcal{D}$ and a general replay corpus $\mathcal{G}$. Let p_{mix} denote the mixed data distribution. The standard CPT objective applies NTP uniformly across the training stream:

$$\mathcal{L}_{\text{NTP}} = \mathbb{E}_{x \sim p_{\text{mix}}} [-\log p_\theta(x_t \mid x_{<t})].$$

In mixed CPT, domain tokens and replay tokens serve different roles. Domain tokens provide the new domain signal, whereas replay tokens are intended to preserve the behavior of the frozen base model on general text. This makes the replay target important: standard replay uses the observed one-hot next token as an indirect preservation signal. We instead keep the domain objective unchanged and replace only the replay target with the frozen base distribution $p_{\theta_{\text{base}}}(\cdot \mid x_{<t})$. We call this *self-distilled* replay because the teacher is the initialization of the student, requiring no external model.

2.1 Next-Token Replay as a Noisy Target

For replay tokens, the preservation target is the frozen base model’s next-token distribution. However, standard NTP replay observes only the next token in the replay corpus. For a replay context $x_{<t}$ with observed token $y = x_t$, it uses the one-hot label e_y as the target, rather than the full base distribution $p_{\theta_{\text{base}}}(\cdot \mid x_{<t})$. Thus, the hard label is an indirect proxy for the base behavior we want replay to preserve, and the gap becomes large when the base model assigns probability mass to multiple plausible next tokens.

To isolate the noise introduced by one-hot supervision, consider the most favorable case where the replay token is sampled from the base distribution itself. In this case, the one-hot label is an unbiased estimate of the preservation target. The problem is variance: each update sees a point mass rather

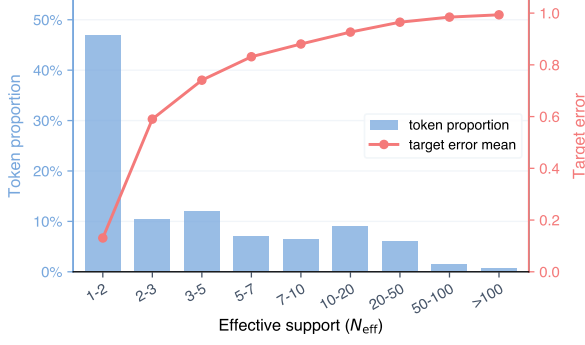


Figure 1: Empirical replay-token statistics. Bars show the fraction of replay tokens in each N_{eff} bin. The line shows the mean observed one-hot target error $\|p_{\theta_{\text{base}}}(\cdot | x_{<t}) - e_y\|_2^2$ in the same bins. By definition, larger N_{eff} corresponds to contexts where the base model assigns probability mass to more plausible next tokens; the line confirms that the observed one-hot target error rises with this support.

than the full distribution. For a fixed context, the covariance of a categorical one-hot sample gives the expected squared target error (The Book of Statistical Proofs, 2022):

$$\mathbb{E}_{y \sim p_{\theta_{\text{base}}}} [\|p_{\theta_{\text{base}}} - e_y\|_2^2] = 1 - \|p_{\theta_{\text{base}}}\|_2^2 = 1 - \frac{1}{N_{\text{eff}}(x_{<t})}.$$

Here, $N_{\text{eff}}(x_{<t}) = 1/\|p_{\theta_{\text{base}}}\|_2^2$ is the order-2 Hill number, also known as the inverse Simpson index or effective number (Hill, 1973; Jost, 2006). Figure 1 shows that this high-variance regime appears in replay data. Many replay contexts have nontrivial effective support under the base model, and the one-hot target error increases with this support.

Real replay corpora add a second source of error. The observed token y is sampled from the empirical data distribution rather than from the frozen base model. Thus, the one-hot target can differ from the preservation target for two reasons: intrinsic sampling variance, even when the data matches the base distribution, and data–base mismatch, when the corpus token is not representative of the base model’s conditional distribution. This makes NTP replay depend not only on the amount of replay data, but also on how closely the corpus matches the base model’s original behavior.

This motivates a direct alternative: use the preservation target itself rather than a sampled proxy. Self-distilled replay replaces this one-hot replay target with the frozen base distribution itself. For replay tokens, we use

$$\mathcal{L}_{\text{sd}}(x_{<t}; \theta) = \text{KL}(p_{\theta_{\text{base}}}(\cdot | x_{<t}) \| p_{\theta}(\cdot | x_{<t}))$$

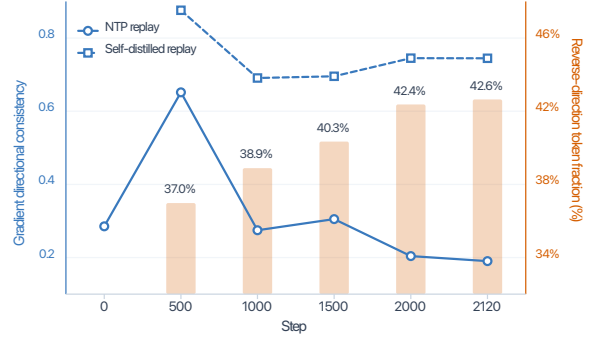


Figure 2: Gradient-variance diagnostics at NTP-trained checkpoints. Lines show gradient directional consistency $\bar{S}_{\text{dir}}$ for NTP and self-distilled replay on shared replay batches from $\mathcal{G}$; higher is better. Bars show the reverse-direction token fraction on these $\mathcal{G}$ tokens, where NTP increases the observed-token logit even though preservation would decrease it.

The mixed training rule is

$$\mathcal{L}(x) = \begin{cases} -\log p_{\theta}(x_t | x_{<t}), & x \in \mathcal{D}, \\ \mathcal{L}_{\text{sd}}(x_{<t}; \theta), & x \in \mathcal{G}, \end{cases}$$

This asymmetric design keeps NTP for domain learning, where the base model lacks target-domain knowledge, and uses distribution matching only for replay, where the goal is behavior preservation.

2.2 Gradient Variance and Behavior Drift

The target mismatch directly enters the replay gradient. For softmax cross-entropy, the logit-level gradient is $p_{\theta} - \text{target}$. Thus, standard NTP replay produces $p_{\theta} - e_y$, whereas self-distilled replay produces $p_{\theta} - p_{\theta_{\text{base}}}$. Their difference is exactly $p_{\theta_{\text{base}}} - e_y$, the one-hot target error analyzed above. This target error therefore becomes part of the update signal used to train the model.

This mismatch can make the NTP update oppose preservation on individual replay tokens. The NTP gradient fundamentally favors increasing the observed-token logit. When the student already assigns that token a higher probability than the base, i.e., $p_{\theta}(y | x_{<t}) > p_{\theta_{\text{base}}}(y | x_{<t})$, the preservation update would instead lower the model’s probability for the observed token. We call this a reverse-direction token, and define the reverse-direction token fraction as the fraction of replay tokens in $\mathcal{G}$ satisfying this condition. On such tokens, the NTP update pushes the student further away from the base model even though the overall NTP loss keeps decreasing; the fraction therefore measures how much of replay training locally works against preservation. The bars in Figure 2

report this fraction. Self-distilled replay instead uses $p_{\theta_{\text{base}}}$, so the model’s probability for the observed token is pulled back toward the base probability.

We summarize this stochastic effect with a directional-consistency diagnostic. For diagnostic purposes, let $g_i = \nabla_{\theta} \mathcal{L}_i(\theta)$ be the replay-loss gradient on batch i with respect to the student parameters, and let $\bar{g} = \frac{1}{K} \sum_i g_i$. We define

$$\bar{\mathcal{S}}_{\text{dir}} = \frac{1}{K} \sum_{i=1}^K \frac{\langle g_i, \bar{g} \rangle}{\|g_i\|_2 \|\bar{g}\|_2}.$$

Lower $\bar{\mathcal{S}}_{\text{dir}}$ means that different replay batches induce less consistent gradient directions, indicating higher directional gradient variance. The lines in Figure 2 show a clear gap: self-distilled replay gives much higher directional consistency than NTP on the same replay batches. This gap becomes especially large at later checkpoints, where the NTP consistency drops while the reverse-direction token fraction rises.

The final consequence of this gradient instability is behavior-level drift. We quantify this by measuring the expected divergence between a checkpoint p_{θ_s} and the frozen base model on a held-out general validation set $\mathcal{G}_{\text{val}}$:

$$\text{Drift}(\theta_s) = \mathbb{E}_{x \sim \mathcal{G}_{\text{val}}} [\text{KL}(p_{\theta_{\text{base}}}(\cdot | x_{<t}) \| p_{\theta_s}(\cdot | x_{<t}))].$$

Figure 3 shows self-distilled replay yields lower drift than NTP under the same token budget. Drift is measured only on held-out general contexts. Domain tokens keep standard NTP, so the model can still move away from the base where adaptation requires it; the shared-parameter tension remains, but it becomes role-aligned.

Together, these diagnostics link replay supervision to behavior preservation: self-distilled replay produces more consistent gradients on $\mathcal{G}$ and lower drift on $\mathcal{G}_{\text{val}}$. The experiments then test whether this preservation advantage improves the domain-general trade-off under the same token budget.

3 Experiments

This section evaluates self-distilled replay from four angles: the main domain-general trade-off under matched budgets, robustness to non-source replay data, sensitivity to distillation hyper-parameters, and comparison with related distillation-style CPT objectives. Appendix A.1 reports a supplementary finance domain check.

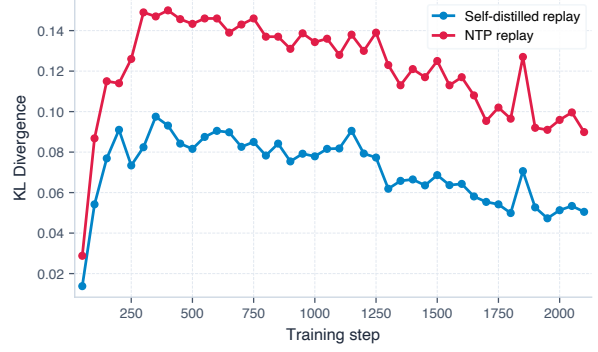


Figure 3: Base-behavior drift on held-out general validation contexts during domain CPT. Lower KL means closer preservation of the frozen base distribution.

3.1 Implementation

Models. Our main experiments use SmolLM3-3B, initialized from the released SmolLM3-3B-Base checkpoint, because it provides public training details, intermediate checkpoints, and the pre-training data mixture (Bakouch et al., 2025). We also evaluate Qwen3-4B/8B-Base (Qwen Team, 2025) to test whether the replay-supervision effect transfers beyond the main SmolLM3 setting.

Domain data. For domain data, we use abstracts from the PubMed annual baseline XML release (National Library of Medicine, 2025). We parse titles and abstracts into document-level examples, remove empty, short, and title-only records, normalize whitespace and control characters, and deduplicate the cleaned texts with exact hashes followed by MinHash near-duplicate removal. The cleaned corpus is shuffled, tokenized with the base-model tokenizer, and split into train and validation sets randomly; for SmolLM3, it contains 28.3M documents and 8.76B tokens, with 8.68B training tokens and 79.7M validation tokens.

Replay data. All experiments use replay data from the SmolLM3 Stage-3 general-data mixture (Bakouch et al., 2025).

Evaluation. For domain evaluation, we use the 1,000-example expert-labeled PubMedQA set (Jin et al., 2019). To reduce option-order artifacts, we evaluate all six permutations of the A/B/C answer options, giving 6×1000 scored examples. For general evaluation, we group benchmarks into five capability categories, as shown in Table 1. Non-coding benchmarks are evaluated with OpenCompass v0.5.2 (OpenCompass Contributors, 2023); coding benchmarks use EvalPlus (Liu et al., 2023) at commit 26d6d00. We average benchmark scores within each category, and aver-

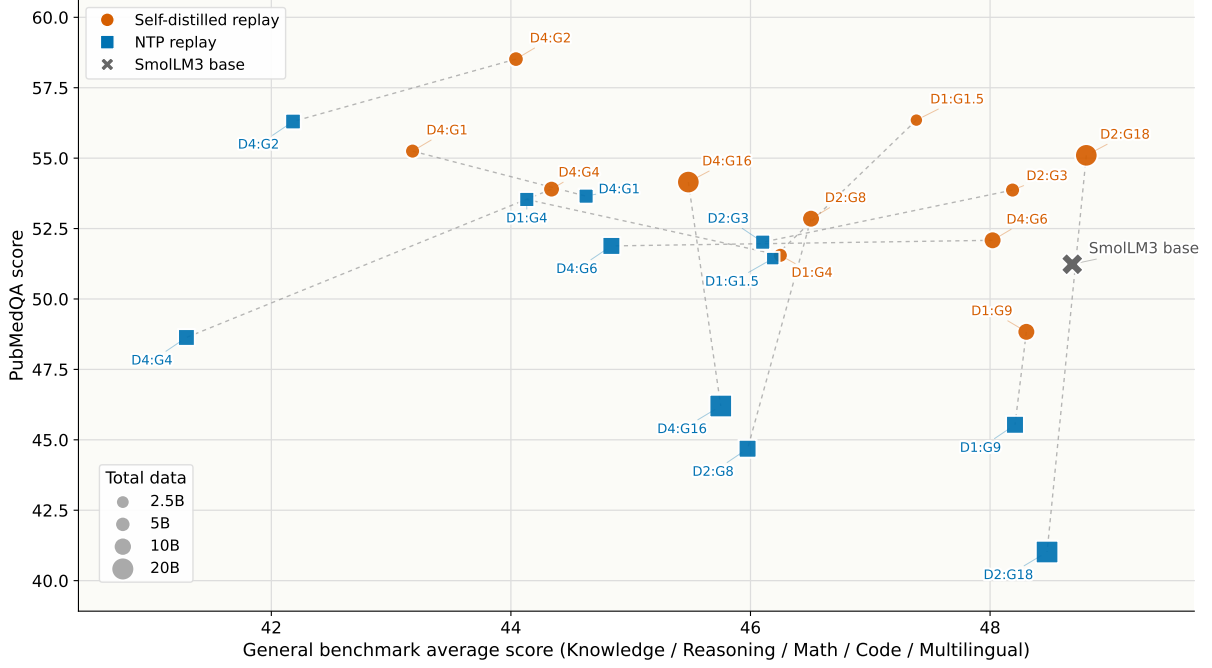


Figure 4: PubMedQA-general capability trade-off across domain:replay budgets; higher is better on both axes. Dashed lines connect matched NTP replay and self-distilled replay runs with the same token budget.

Category	Benchmarks
Knowledge	MMLU (Hendrycks et al., 2021), MMLU-Pro (Wang et al., 2024), C-Eval (Huang et al., 2023), GPQA-Diamond (Rein et al., 2024), SuperGPQA (M-A-P Team, 2025)
Reasoning	ARC-C/E (Clark et al., 2018), BBH (Suzgun et al., 2023), BoolQ (Clark et al., 2019), COPA (Roemmele et al., 2011), CommonsenseQA (Talmor et al., 2019), HellaSwag (Zellers et al., 2019), OpenBookQA (Mihaylov et al., 2018), PIQA (Bisk et al., 2020), WinoGrande (Sakaguchi et al., 2020)
Math	GSM8K (Cobbe et al., 2021), MATH-500 (Lightman et al., 2024)
Coding	HumanEval (Chen et al., 2021), HumanEval+ (Liu et al., 2023), MBPP (Austin et al., 2021), MBPP+ (Liu et al., 2023)
Multilingual	MGSM (Shi et al., 2023), MMMLU (downsampled) (OpenAI, 2024)

Table 1: General evaluation suite.

age the five category scores to obtain the general-capability score.

Training setup. We train with Nanotron (Hugging Face, 2024) commit bc1f5a6, using context length 4096 and batch size 576 per optimizer update, split into micro-batches of 4 sequences for gradient accumulation. Following SmolLM3’s Stage-3 decay-phase recipe (Bakouch et al., 2025) and re-warmup practice (Gupta et al., 2023), we warm up the learning rate to 2×10^{-4} and linearly decay it to zero. Experiments run on NVIDIA B200 GPUs; a representative 5B-token SmolLM3 CPT run uses about 56 B200 GPU-hours, and all training runs use approximately 9k B200 GPU-hours.

In each matched NTP replay and self-distilled replay comparison, we keep the initialization checkpoint, replay sources, domain/replay token budgets, mixture weights, sequence length, optimizer, learning-rate schedule, and evaluation pipeline fixed; only the replay supervision objective differs.

For self-distilled replay, we pack domain and replay sequences into separate micro-batches when possible. Domain-only micro-batches use only the student NTP forward pass, while replay-containing micro-batches also run the frozen base model to compute KL targets.

3.2 Domain-General Trade-off

This subsection tests whether changing only the replay supervision objective improves the trade-off between domain adaptation and general capability retention. We denote a domain:replay budget as $D_m:G_n$, meaning m billion (B) domain tokens and n billion general replay tokens. Our sweep includes replay-heavy budgets because preservation can be a first-class constraint in domain-oriented LLM training. This is common in large-scale recipes: e-Llama (Herold et al., 2025) uses only 50% e-commerce data, BloombergGPT (Wu et al., 2023) mixes 363B finance tokens with 345B general tokens, and long-context CPT (Gao et al., 2025) combines long data with high-quality short-

context data. We therefore treat high-replay settings as preservation-oriented points on the trade-off curve, not as universal production ratios.

Figure 4 plots the PubMedQA–general trade-off across domain and replay-token budgets. The most controlled comparison is each connected pair in the figure: both runs use the same domain tokens, replay tokens, data, and training setup, and differ only in the replay target. Under this matched-budget comparison, self-distilled replay strictly improves both PubMedQA and the general score in 8 of 11 pairs, while NTP replay never dominates it on both axes. Overall, self-distilled replay improves the general score in 9 pairs and PubMedQA in 10, with average gains of 1.16 and 4.33 points, respectively. These settings jointly vary domain volume, replay volume, ratio, and total budget; within each setting, the matched pair isolates the replay objective. We repeat two representative settings three times each (Appendix B); the General gap exceeds one sample standard deviation in both settings, while the PubMedQA differences are smaller.

Representative points show the same pattern at different replay scales. At D1:G1.5, self-distilled replay raises the general score from 46.19 to 47.38 and PubMedQA from 51.43 to 56.35. At D2:G3, it again improves both axes, from 46.10 to 48.19 on general score and from 52.02 to 53.87 on PubMedQA. At D2:G18, NTP replay recovers general capability but collapses to 41.02 on PubMedQA; self-distilled replay preserves similar general capability, 48.80 versus 48.48, while maintaining much higher PubMedQA at 55.10. Finance and legal results in Appendix A point in the same direction.

The main exception is the low-replay corner D4:G1: self-distilled replay still improves PubMedQA, but its general score drops by 1.45 points. Because replay accounts for only 20% of the tokens, the preservation signal is too small to offset dominant domain updates. Thus, when replay has sufficient budget to act, changing its target can improve both domain and general performance within a matched budget, rather than improving one only by sacrificing the other.

Training cost. Self-distilled replay adds a frozen-teacher forward pass on replay tokens, so overhead scales with the replay fraction rather than applying uniformly. As shown in Figure 5, even a conservative baseline that runs the teacher on every micro-batch retains 74.7% of

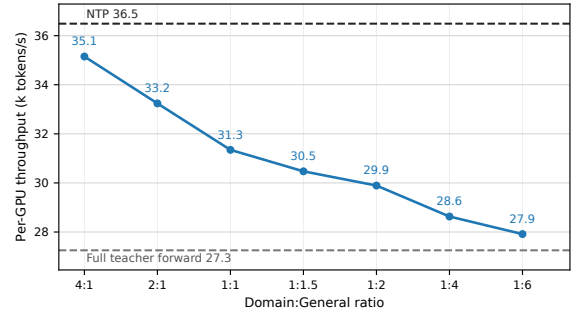


Figure 5: Training throughput across domain:replay mixtures. The full-teacher-forward baseline shows a conservative cost upper bound.

NTP throughput; at the 1:1.5 domain:replay ratio, for example, skipping the teacher on replay-free micro-batches reaches 83.5% of NTP throughput, equivalent to a 19.7% increase in training time per processed token. This cost is modest compared to the alternative of simply increasing replay token volume. In an 8-GPU B200 benchmark, peak per-GPU memory increases by 17.4% (170.48 vs. 145.15 GiB).

3.3 Benefit Beyond Original Replay Data

In self-distilled replay, replay tokens serve only as context, and supervision comes from the teacher’s output distribution, not from the tokens themselves. This decoupling means the method should be robust to replay data that differs from the original pretraining corpus. We test this hypothesis on Qwen3 models, whose original pretraining mixture is not publicly available.

We run Qwen3-4B-Base and Qwen3-8B-Base using SmolLM3’s replay mixture as non-source replay data, mismatched with Qwen’s proprietary pretraining mixture. With NTP replay, this mismatch affects general-domain preservation because the one-hot targets come from the replay source. Self-distilled replay instead uses soft targets from Qwen’s frozen base distribution, independent of the replay-token origin.

Figure 6 shows the PubMedQA–general trade-off for eight matched model-budget pairs, with the base checkpoints shown only as references. Self-distilled replay improves both general score and PubMedQA in all eight pairs, with average gains of 5.41 and 3.52 points over matched NTP replay. This consistent advantage across model sizes and mismatched replay data shows that the benefit is not tied to SmolLM3 or its original replay mixture. Appendix D provides the full benchmark-level scores. Relative to each base model, self-

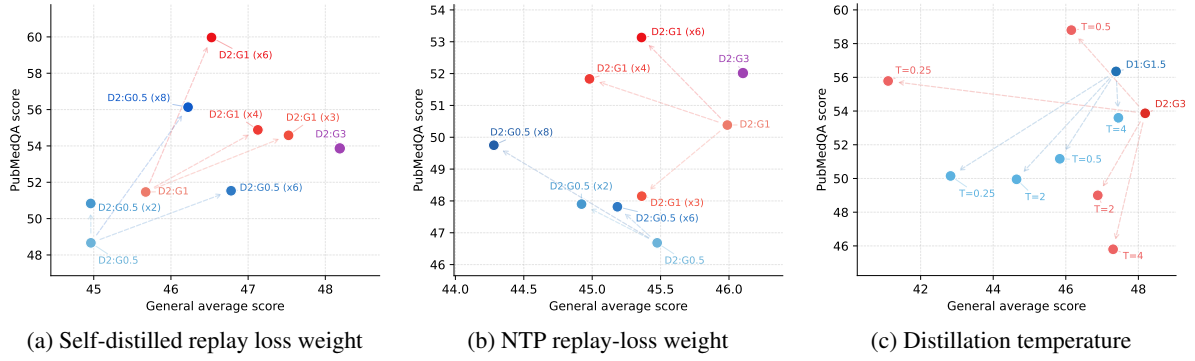


Figure 7: Hyper-parameter sensitivity on PubMedQA-general trade-off. (a) Replay-loss weight under self-distilled replay. (b) Replay-loss weight under NTP replay. In (a,b), labels such as $(\times 2)$ denote the replay-loss multiplier λ ; unlabeled points use $\lambda = 1$. (c) Distillation temperature.

distillation-based lifelong pretraining (Jin et al., 2022), Mix-CPT uses logit-swap self-distillation (Jiang et al., 2025), LongReD uses hidden-state restoration (Dong et al., 2025), and GeRe uses threshold-margin replay (Zhang et al., 2025).

This is a controlled objective-design comparison under one matched PubMed D2:G3 setting, not a fully tuned leaderboard for each prior method. All of these constraint/distillation variants score below pure NTP on PubMedQA, whereas self-distilled replay improves over pure NTP on both PubMedQA and general score. This separation suggests that self-distilled replay is not simply another instance of adding distillation regularization. Appendix C reports coefficient sweeps for these two methods on the same setting. Lifelong-PT reaches at most 48.08 General ($\lambda = 4$) or 50.57 PubMedQA ($\lambda = 0.5$); Mix-CPT peaks at 45.94 General and 51.72 PubMedQA ($\alpha = 0.90$). Self-distilled replay reaches 48.19 General and 53.87 PubMedQA.

The comparison points to two distinct design issues. On the replay side, logit L2, hidden cosine, and threshold-margin losses make extra assumptions about which representation of old behavior should be preserved. Such assumptions may be appropriate in their original settings, but they are not necessarily suitable for domain CPT replay. Self-distilled replay makes no feature- or logit-level proxy assumption; it directly constrains the student to match the frozen-base next-token distribution. On the domain side, methods that add distillation-style regularization to domain tokens do not improve domain adaptation here, consistent with the need to keep the domain objective plastic.

Self-distilled replay avoids both changes made by the more complex objectives. It leaves domain

tokens on NTP and only softens the replay target from a one-hot label to the frozen-base distribution. Since this minimal target substitution improves over pure NTP while the heavier constraint variants do not, the useful intervention is target replacement rather than additional distillation machinery.

4 Related Work

Replay in domain CPT and continual adaptation. Continual pre-training (CPT) adapts a pre-trained language model to new domains by continuing language modeling on new text (Gururangan et al., 2020; Gupta et al., 2023; Ibrahim et al., 2024), but preserving general ability under domain shift remains difficult (Wang et al., 2023; Li and Lee, 2024; Li et al., 2025). Replay is a common mitigation because it mixes old or general text into the new-data stream without changing the model architecture (Abbes et al., 2026; Lu et al., 2026), while other work explores regularization, model merging, or capacity isolation (Kirkpatrick et al., 2017; Ke et al., 2023; Šliogeris et al., 2025; Liu and Chang, 2026; Yadav et al., 2023; Siriwardhana et al., 2024; Ueda et al., 2025; Zhang et al., 2026). Prior CPT replay work has largely framed replay as a question of data mixtures and token budgets, while applying one-hot next-token prediction (NTP) uniformly to both domain and replay tokens. This implicitly treats replay labels as ordinary empirical samples, leaving underexamined the target variance and corpus dependence introduced when sampled tokens are used to preserve base-model behavior.

Distillation and soft-label replay. Knowledge distillation transfers a soft teacher distribution to a student (Hinton et al., 2014), and continual learn-

Method	Domain obj.	Replay obj.	PubMedQA	General	Know.	Reason.	Math	Code	Multi.
Base	–	–	51.23	48.69	33.96	70.53	46.58	47.73	44.64
DER	NTP	logit L2	42.73	40.86	32.18	67.12	23.65	42.05	39.31
DER++	NTP	NTP + logit L2	43.28	45.49	31.80	68.42	46.26	40.70	40.25
SER	NTP + logit L2	NTP + logit L2	45.80	45.94	32.39	67.76	48.07	41.00	40.49
Lifelong PT	NTP + KD	NTP + KD	45.35	48.02	36.97	69.07	50.03	41.53	42.51
Mix-CPT	NTP + LSSD	NTP + LSSD	48.72	46.20	32.00	68.04	48.69	41.38	40.91
LongReD	NTP	hidden cosine	50.52	43.46	33.92	67.32	33.78	40.55	41.74
GeRe	NTP	NTP + TM	46.82	45.61	38.76	66.55	41.19	40.23	41.32
Pure NTP	NTP	NTP	52.02	46.10	35.43	66.37	49.35	39.77	39.59
SDR	NTP	KL	53.87	48.19	37.62	69.23	48.23	43.83	42.04

Table 3: Controlled comparison of pure NTP, self-distilled replay, and distillation-style CPT objectives under matched PubMed D2:G3 training. Base is the frozen checkpoint before CPT. All trained methods use the same model, data, token budget, and training setup; only the objective design differs. General averages Knowledge, Reasoning, Math, Code, and Multilingual scores. KD, LSSD, and TM denote knowledge distillation, logit-swap self-distillation, and threshold-margin loss, respectively; hidden cosine matches frozen-base last-layer hidden states with cosine loss. **Bold** marks the best *trained* method in each column.

ing has long used such signals to preserve previous behavior (Li and Hoiem, 2018). Related methods include logit replay in DER/DER++ (Buzzega et al., 2020), replay with distillation-style consistency in SER (Zhuo et al., 2023), distillation-based lifelong pretraining in Lifelong Pretraining (Jin et al., 2022), logit-swap self-distillation in Mix-CPT (Jiang et al., 2025), hidden-state restoration in LongReD (Dong et al., 2025), and threshold-margin replay in GeRe (Zhang et al., 2025). For episodic supervised methods such as DER and SER, soft labels mainly compensate for memory scarcity by storing richer information in each replay example. Language-model variants use distillation as an auxiliary constraint, often over all tokens; for example, Lifelong Pretraining uses *MLM + KD* across current and memory batches, while Mix-CPT combines NTP with logit-swap self-distillation over a unified knowledge mixture. Domain CPT has a different bottleneck: replay text can be available at scale, but each one-hot next-token label is a high-variance target for preserving the base model’s conditional behavior. Self-distilled replay is narrower: it leaves domain-token NTP unchanged and only substitutes the replay target, replacing the sampled one-hot label with the frozen base model’s next-token distribution. Thus, our focus is not distillation machinery, but the asymmetric supervision role of replay tokens in domain CPT.

5 Conclusion

This paper revisits domain-CPT replay as a supervision-target problem: self-distilled replay

keeps domain-token NTP unchanged and replaces one-hot replay supervision with KL matching to the frozen base model. This minimal target substitution addresses the fact that sampled replay tokens are noisy and corpus-dependent proxies for the base behavior replay is meant to preserve. Our analysis links this change to lower target error, more stable replay gradients, and reduced base-behavior drift. Empirically, it improves the domain–general trade-off over NTP replay in matched settings, remains effective with non-source replay data for models such as Qwen whose pretraining mixtures are unavailable, and outperforms more complex distillation-style variants in the evaluated CPT setting. Together, these findings show that moving beyond one-hot replay to base-behavior preservation reduces target variance and yields a simple, data-flexible path to improve domain CPT.

Limitations

Our experiments focus on replay-based domain CPT for decoder-only language models, with biomedical adaptation as the main domain and benchmark averages as general-capability diagnostics. They do not establish the same trade-off for all domains, model scales, languages, or instruction-tuned chat models. We use biomedical adaptation as a high-value, distribution-shifted testbed; the target-noise and gradient-variance mechanisms arise from the replay objective itself. We include supplementary finance and legal checks in Appendix A, but neither is a full sweep over domains, ratios, or hyper-parameters; broader

empirical validation remains future work.

Self-distilled replay also requires access to the frozen base model during training: batches containing replay examples need an additional teacher forward pass, even though domain-only batches skip it. Although the throughput overhead is modest at the scales tested (Figure 5), retaining the frozen teacher also increases the parameter-memory footprint. Scaling this design to very large models, such as 100B+ checkpoints, may require systems optimizations such as teacher offloading, cached teacher logits, or careful tensor-parallel scheduling.

Self-distilled replay also assumes that the frozen base model’s behavior is worth preserving: if the base is undertrained or systematically biased, SDR can preserve those deficiencies, and abilities absent from the replay corpus are not directly constrained.

Finally, our compute budget concentrates the broadest sweep on the main matched-budget comparison between NTP replay and self-distilled replay. The non-source replay and hyper-parameter studies are smaller robustness and sensitivity checks, and the distillation-style baselines are compared under a single matched PubMed D2:G3 setting rather than exhaustively tuned per method. Thus, Table 3 should be read as a controlled comparison of objective designs under one setting, not as a fully optimized leaderboard for all prior methods. Broader sweeps over replay sources, hyper-parameters, and baseline-specific tuning remain useful future work.

Ethical Considerations

We use publicly released models, corpora, benchmarks, and software frameworks under their original licenses and terms of use, together with the supplementary finance data described in Appendix A.1, for research training and evaluation only. This includes the released artifacts used in our experiments, such as SmolLM3, Qwen3, Nanotron, OpenCompass, PubMed, PubMedQA, and the replay data source. We do not redistribute original model weights or raw corpora, and derived checkpoints are not intended for clinical decision making or production deployment.

We do not collect new human-subject data or recruit annotators. For personally identifying or offensive content risks, we rely on the release and curation practices of public source artifacts; we do

not conduct an additional exhaustive per-example audit of the web-derived finance corpus or internal documents. Because self-distilled replay preserves frozen-base behavior, it may also retain biases, factual errors, or unsafe tendencies present in the base model.

References

- Istabraq Abbas, Gopesh Subbaraj, Matthew Riemer, Nizar Islah, Tsuguchika Tabaru, Hiroaki Kingetsu, Sarath Chandar, and Irina Rish. 2026. [Revisiting replay and gradient alignment for continual pre-training of large language models](#). In *Proceedings of The 4th Conference on Lifelong Learning Agents*, volume 330 of *Proceedings of Machine Learning Research*, pages 465–486. PMLR.
- Jacob Austin, Augustus Odena, Maxwell Nye, Maarten Bosma, Henryk Michalewski, David Dohan, Ellen Jiang, Carrie Cai, Michael Terry, Quoc Le, and Charles Sutton. 2021. [Program synthesis with large language models](#). *Preprint*, arXiv:2108.07732.
- Elie Bakouch, Loubna Ben Allal, Anton Lozhkov, Nouamane Tazi, Lewis Tunstall, Carlos Miguel Patiño, Edward Beeching, Aymeric Roucher, Akssel Joonas Reedi, Quentin Gallouédec, Kashif Rasul, Nathan Habib, Clémentine Fourrier, Hynek Krdlicek, Guilherme Penedo, Hugo Larcher, Mathieu Morlon, Vaibhav Srivastav, Joshua Lochner, and 4 others. 2025. SmolLM3: smol, multilingual, long-context reasoner. <https://huggingface.co/blog/smollm3>.
- Yonatan Bisk, Rowan Zellers, Ronan Le Bras, Jianfeng Gao, and Yejin Choi. 2020. PIQA: Reasoning about physical commonsense in natural language. In *Proceedings of the AAAI Conference on Artificial Intelligence*.
- Pietro Buzzega, Matteo Boschini, Angelo Porrello, Davide Abati, and Simone Calderara. 2020. Dark experience for general continual learning: a strong, simple baseline. In *Advances in Neural Information Processing Systems*.
- Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, Alex Ray, Raul Puri, Gretchen Krueger, Michael Petrov, Heidy Khlaaf, Girish Sastry, Pamela Mishkin, Brooke Chan, Scott Gray, and 39 others. 2021. [Evaluating large language models trained on code](#). *Preprint*, arXiv:2107.03374.
- Daixuan Cheng, Shaohan Huang, and Furu Wei. 2024. [Adapting large language models to domains via reading comprehension](#). In *International Conference on Learning Representations*.
- Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina

- Toutanova. 2019. BoolQ: Exploring the surprising difficulty of natural yes/no questions. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*.
- Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. [Think you have solved question answering? try ARC, the AI2 reasoning challenge](#). Preprint, arXiv:1803.05457.
- Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. [Training verifiers to solve math word problems](#). Preprint, arXiv:2110.14168.
- Zican Dong, Junyi Li, Jinhao Jiang, Mingyu Xu, Wayne Xin Zhao, Bingning Wang, and Weipeng Chen. 2025. [LongReD: Mitigating short-text degradation of long-context large language models via restoration distillation](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10687–10707. Association for Computational Linguistics.
- Tianyu Gao, Alexander Wettig, Howard Yen, and Danqi Chen. 2025. [How to train long-context language models \(effectively\)](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7376–7399. Association for Computational Linguistics.
- Neel Guha, Julian Nyarko, Daniel Ho, Christopher Ré, Adam Chilton, Aditya K, Alex Chohlas-Wood, Austin Peters, Brandon Waldon, Daniel Rockmore, Diego Zambrano, Dmitry Talisman, Enam Hoque, Faiz Surani, Frank Fagan, Galit Sarfaty, Gregory Dickinson, Haggai Porat, Jason Hegland, and 21 others. 2023. [Legalbench: A collaboratively built benchmark for measuring legal reasoning in large language models](#). Preprint, arXiv:2308.11462.
- Kshitij Gupta, Benjamin Thérien, Adam Ibrahim, Mats L. Richter, Quentin Anthony, Eugene Belilovsky, Irina Rish, and Timothée Lesort. 2023. [Continual pre-training of large language models: How to \(re\)warm your model?](#) Preprint, arXiv:2308.04014.
- Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. [Don’t stop pretraining: Adapt language models to domains and tasks](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360. Association for Computational Linguistics.
- Peter Henderson, Mark Krass, Lucia Zheng, Neel Guha, Christopher D. Manning, Dan Jurafsky, and Percy Liang. 2022. [Pile of law: Learning responsible data filtering from the law and a 256gb open-source legal dataset](#). In *Advances in Neural Information Processing Systems Datasets and Benchmarks Track*.
- Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. [Measuring massive multitask language understanding](#). In *International Conference on Learning Representations*.
- Christian Herold, Michael Kozielski, Tala Bazazo, Pavel Petrushkov, Patrycja Cieplicka, Dominika Basaj, Yannick Versley, Seyyed Hadi Hashemi, and Shahram Khadivi. 2025. [Domain adaptation of foundation LLMs for e-commerce](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 6: Industry Track)*, pages 1039–1049. Association for Computational Linguistics.
- M. O. Hill. 1973. [Diversity and evenness: A unifying notation and its consequences](#). *Ecology*, 54(2):427–432.
- Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2014. [Distilling the knowledge in a neural network](#). In *NIPS 2014 Deep Learning Workshop*.
- Yuzhen Huang, Yuzhuo Bai, Zhihao Zhu, Junlei Zhang, Jinghan Zhang, Tangjun Su, Junteng Liu, Chuancheng Lv, Yikai Zhang, Jiayi Lei, Yao Fu, Maosong Sun, and Junxian He. 2023. [C-Eval: A multi-level multi-discipline chinese evaluation suite for foundation models](#). In *Advances in Neural Information Processing Systems*.
- Hugging Face. 2024. [Nanotron: Minimalistic large language model 3d-parallelism training](#). <https://github.com/huggingface/nanotron>.
- Adam Ibrahim, Benjamin Thérien, Kshitij Gupta, Mats L. Richter, Quentin Anthony, Timothée Lesort, Eugene Belilovsky, and Irina Rish. 2024. [Simple and scalable strategies to continually pre-train large language models](#). *Transactions on Machine Learning Research*.
- Jinhao Jiang, Junyi Li, Xin Zhao, Yang Song, Tao Zhang, and Ji-Rong Wen. 2025. [Mix-CPT: A domain adaptation framework via decoupling knowledge learning and format alignment](#). In *International Conference on Learning Representations*.
- Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. 2019. [PubMedQA: A dataset for biomedical research question answering](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2567–2577.
- Xisen Jin, Dejiao Zhang, Henghui Zhu, Wei Xiao, Shang-Wen Li, Xiaokai Wei, Andrew Arnold, and

- Xiang Ren. 2022. Lifelong pretraining: Continually adapting language models to emerging corpora. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4764–4780. Association for Computational Linguistics.
- Lou Jost. 2006. Entropy and diversity. *Oikos*, 113(2):363–375.
- Zixuan Ke, Yijia Shao, Haowei Lin, Tatsuya Konishi, Gyuhak Kim, and Bing Liu. 2023. Continual pre-training of language models. In *International Conference on Learning Representations*.
- 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. 2017. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13):3521–3526.
- Chen-An Li and Hung-Yi Lee. 2024. Examining forgetting in continual pre-training of aligned large language models. *Preprint*, arXiv:2401.03129.
- Jeffrey Li, Mohammadreza Armandpour, Seyed Iman Mirzadeh, Sachin Mehta, Vaishaal Shankar, Raviteja Vemulapalli, Samy Bengio, Oncel Tuzel, Mehrdad Farajtabar, Hadi Pouransari, and Fartash Faghri. 2025. TIC-LM: A web-scale benchmark for time-continual LLM pretraining. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 32231–32273. Association for Computational Linguistics.
- Zhizhong Li and Derek Hoiem. 2018. Learning without forgetting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(12):2935–2947.
- Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. 2024. Let’s verify step by step. In *International Conference on Learning Representations*.
- Jiawei Liu, Chunqiu Steven Xia, Yuyao Wang, and Lingming Zhang. 2023. Is your code generated by ChatGPT really correct? rigorous evaluation of large language models for code generation. In *Advances in Neural Information Processing Systems*, volume 36.
- Xuan Liu and Xiaobin Chang. 2026. Elastic weight consolidation done right for continual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.
- Yiyang Lu, Yu He, Jianlong Chen, and Hongyuan Zha. 2026. MSSR: Memory-aware adaptive replay for continual LLM fine-tuning. *Preprint*, arXiv:2603.09892.
- M-A-P Team. 2025. SuperGPQA: Scaling LLM evaluation across 285 graduate disciplines. In *Advances in Neural Information Processing Systems Datasets and Benchmarks Track*.
- Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. 2018. Can a suit of armor conduct electricity? a new dataset for open book question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*.
- National Library of Medicine. 2025. Download PubMed data. Accessed 2026-05-20.
- OpenAI. 2024. Multilingual massive multitask language understanding (MMMLU). <https://huggingface.co/datasets/openai/MMMLU>.
- OpenCompass Contributors. 2023. OpenCompass: A universal evaluation platform for foundation models. <https://github.com/open-compass/opencompass>.
- Haoran Que, Jiaheng Liu, Ge Zhang, Chenchen Zhang, Xingwei Qu, Yinghao Ma, Feiyu Duan, Zhiqi Bai, Jiakai Wang, Yuanxing Zhang, Xu Tan, Jie Fu, Jiamang Wang, Lin Qu, Wenbo Su, and Bo Zheng. 2024. D-CPT Law: Domain-specific continual pre-training scaling law for large language models. In *Advances in Neural Information Processing Systems*, volume 37.
- Qwen Team. 2025. Qwen3 Technical Report. *Preprint*, arXiv:2505.09388.
- David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. 2024. GPQA: A graduate-level google-proof Q&A benchmark. In *First Conference on Language Modeling*.
- Melissa Roemmele, Cosmin Adrian Bejan, and Andrew S. Gordon. 2011. Choice of plausible alternatives: An evaluation of commonsense causal reasoning. In *AAAI Spring Symposium on Logical Formalizations of Commonsense Reasoning*.
- Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2020. WinoGrande: An adversarial winograd schema challenge at scale. In *Proceedings of the AAAI Conference on Artificial Intelligence*.
- Freda Shi, Mirac Suzgun, Markus Freitag, Xuezhi Wang, Suraj Srivats, Soroush Vosoughi, Hyung Won Chung, Yi Tay, Sebastian Ruder, Denny Zhou, Dipanjan Das, and Jason Wei. 2023. Language models are multilingual chain-of-thought reasoners. In *The Eleventh International Conference on Learning Representations*.
- Shamane Siriwardhana, Mark McQuade, Thomas Gauthier, Lucas Atkins, Fernando Fernandes Neto, Luke Meyers, Anneketh Vij, Tyler Odenthal, Charles Goddard, Mary MacCarthy, and Jacob Solawetz. 2024. Domain adaptation of Llama3-70B-Instruct through

- continual pre-training and model merging: A comprehensive evaluation. *Preprint*, arXiv:2406.14971.
- Vytėnė Šliogeris, Povilas Daniušis, and Artūras Nakvošas. 2025. Full-parameter continual pretraining of Gemma2: Insights into fluency and domain knowledge. *Preprint*, arXiv:2505.05946.
- Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc V. Le, Ed H. Chi, Denny Zhou, and Jason Wei. 2023. Challenging BIG-Bench tasks and whether chain-of-thought can solve them. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13003–13051. Association for Computational Linguistics.
- Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. CommonsenseQA: A question answering challenge targeting commonsense knowledge. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*.
- The Book of Statistical Proofs. 2022. Proof: Covariance matrix of the categorical distribution. Accessed 2026-05-19.
- Kentaro Ueda, François Portet, Hirohiko Suwa, and Keiichi Yasumoto. 2025. Merging continual pre-training models for domain-specialized LLMs: A case study in finance. *Preprint*, arXiv:2511.02451.
- 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. 2023. TRACE: A comprehensive benchmark for continual learning in large language models. *Preprint*, arXiv:2310.06762.
- Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyang Jiang, Tianle Li, Max Ku, Kai Wang, Alex Zhuang, Rongqi Fan, Xiang Yue, and Wenhui Chen. 2024. MMLU-Pro: A more robust and challenging multi-task language understanding benchmark. In *Advances in Neural Information Processing Systems*.
- Shijie Wu, Ozan Irsoy, Steven Lu, Vadim Dabravolski, Mark Dredze, Sebastian Gehrmann, Prabhjan Kambadur, David Rosenberg, and Gideon Mann. 2023. BloombergGPT: A large language model for finance. *Preprint*, arXiv:2303.17564.
- Prateek Yadav, Derek Tam, Leshem Choshen, Colin Raffel, and Mohit Bansal. 2023. TIES-Merging: Resolving interference when merging models. In *Advances in Neural Information Processing Systems*, volume 36.
- Jiasheng Ye, Peiju Liu, Tianxiang Sun, Jun Zhan, Yunhua Zhou, and Xipeng Qiu. 2025. Data mixing laws: Optimizing data mixtures by predicting language modeling performance. In *International Conference on Learning Representations*.
- Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. HellaSwag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*.
- Jinyang Zhang, Yue Fang, Hongxin Ding, Weibin Liao, Muyang Ye, Junfeng Zhao, Yasha Wang, and Xu Chu. 2026. ADEPT: Continual pretraining via adaptive expansion and dynamic decoupled tuning. In *International Conference on Learning Representations*.
- Yunan Zhang, Shuoran Jiang, Mengchen Zhao, Yuefeng Li, Yang Fan, Xiangping Wu, and Qingcai Chen. 2025. GeRe: Towards efficient anti-forgetting in continual learning of LLM via general samples replay. *Preprint*, arXiv:2508.04676.
- Lucia Zheng, Neel Guha, Brandon R. Anderson, Peter Henderson, and Daniel E. Ho. 2021. When does pretraining help? assessing self-supervised learning for law and the CaseHOLD dataset. In *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Law*, pages 159–168.
- Tao Zhuo, Zhiyong Cheng, Zan Gao, Hehe Fan, and Mohan Kankanhalli. 2023. Continual learning with strong experience replay. *Preprint*, arXiv:2305.13622.

A Additional Domain Results

A.1 Finance

Scope. We include an additional cross-domain check in finance using professional exam-style benchmarks. This experiment is intended as supplementary evidence beyond the biomedical main-text setting, not as a fully tuned sweep over domains, replay ratios, or hyper-parameters.

Domain evaluation. For domain evaluation, we use exam-style practice questions aligned with the syllabi of CAMS, CDCS, CertICM, CertPay, CITF, and FRM. We also report perplexity on held-out internal documents, where lower values indicate better domain fit.

Training data. The domain corpus is built from large web corpora, using keyword filtering followed by judge-model filtering for finance relevance. The replay corpus is the same general replay mixture used in the main experiments. This keeps the replay setting aligned with the main results while changing the target domain and domain evaluation.

Results. Figure 8 reports average score differences from the baseline. The 11B NTP-replay-only run is not a neutral update: even without domain data, it lowers the domain exam average by 3.05 points, increases document PPL by 0.91, and reduces four of five general-capability categories. This domain drop may partly reflect general-capability erosion, but it is also consistent with our premise that one-hot replay is a noisy preservation target rather than a stable anchor. Adding 4B domain NTP with 11B NTP replay reverses the domain trend by improving the exam average by 3.73 points and lowering PPL by 0.85, but still produces large general drops, especially 7.30 points in Code and 5.30 in Multilingual. With self-distilled replay, the same domain-adaptation setup gives the strongest domain result: the exam average improves by 5.76 points and PPL drops by 1.11, with smaller drops in every general category.

Pairwise, both domain-adapted runs improve the domain metrics over replay-only training, showing that the filtered finance corpus provides the needed adaptation signal. More importantly, replacing NTP replay with self-distilled replay improves the domain-exam gain by 2.03 points and further lowers PPL, while preserving general ability better in every category: Knowledge drops by 1.04 points rather than 1.70, Reasoning by 1.66

rather than 4.18, Math by 1.01 rather than 1.11, Code by 5.58 rather than 7.30, and Multilingual by 2.91 rather than 5.30. Thus, the finance check supports the same conclusion as the main biomedical experiments: self-distilled replay turns replay into a more useful preservation signal, improving domain adaptation with less general-capability erosion than standard NTP replay.

A.2 Legal

We include a third domain as a direction check only. We train SmolLM3 on 4B tokens of the public Pile of Law (Henderson et al., 2022) with 12B tokens of the general replay mixture, and evaluate on CaseHOLD (Zheng et al., 2021) and LegalBench (Guha et al., 2023). This public-data setup does not produce a stable absolute domain improvement over the frozen base model, so we do not use it as a standalone validation of legal CPT; we report the matched NTP-SDR comparison only as evidence that the direction of the effect is not biomedical-specific. Self-distilled replay is 1.67 points higher on General and 2.94 points higher on Legal than NTP replay.

Method	General	Legal
Frozen base	48.69	51.00
NTP replay	42.18	45.09
Self-distilled replay	43.85	48.03

Table 4: Legal-domain matched comparison. General averages the same five capability categories as the main text; Legal averages CaseHOLD and LegalBench. Both methods use the same 4B domain / 12B replay budget.

B Run-to-Run Variance

We reran two representative settings three times each with different data-order seeds. Table 5 reports the mean and sample standard deviation; the General gap exceeds one standard deviation in both settings.

Setting	Method	General	PubMedQA
D1:G4	NTP	43.66 $\pm$ 0.57	52.33 $\pm$ 1.07
	SDR	47.24 $\pm$ 0.95	51.42 $\pm$ 0.80
D2:G3	NTP	45.37 $\pm$ 0.64	51.15 $\pm$ 0.93
	SDR	47.95 $\pm$ 0.60	53.45 $\pm$ 0.87

Table 5: Run-to-run variance for two representative matched settings. Scores are mean $\pm$ sample standard deviation over three runs with different data-order seeds. SDR denotes self-distilled replay.

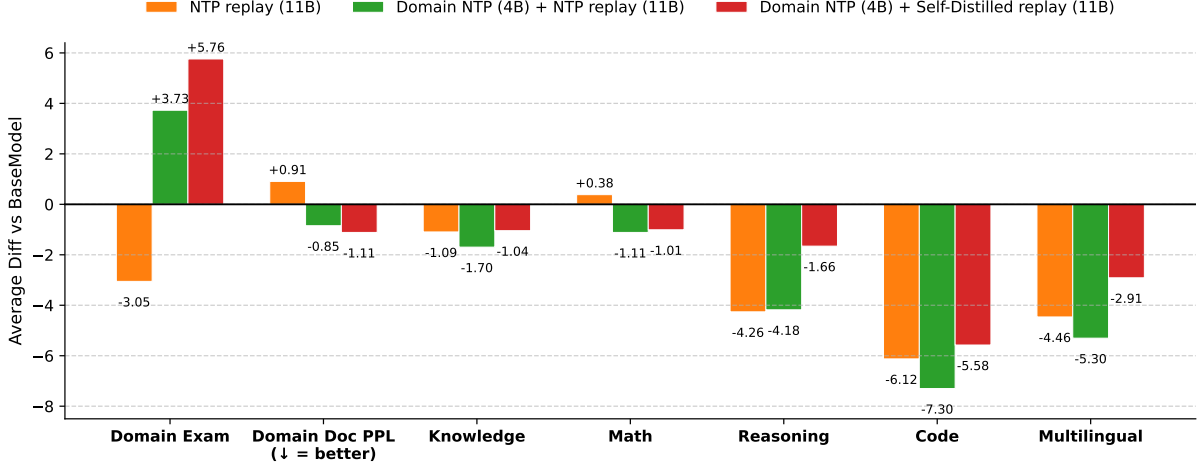


Figure 8: Finance domain results. Bars show average differences from the baseline across domain and general categories. The domain exam score averages CAMS, CDCS, CertICM, CertPay, CITE, and FRM; domain-document PPL is measured on held-out internal documents, where lower is better.

C Baseline Coefficient Sweeps

The main-text comparison probes objective designs under one coefficient per method. To check whether the two closest all-token regularizer-style methods would match self-distilled replay under their own loss weights, we run single-seed coefficient grids on the same PubMed D2:G3 setting. These runs are controlled objective probes, not full reproductions of the original training systems.

Method	General	PubMedQA
Frozen base	48.69	51.23
Pure NTP	46.10	52.02
SDR	48.19	53.87
Lifelong-PT, $\lambda = 0.25$	46.01	48.07
Lifelong-PT, $\lambda = 0.5$	46.87	50.57
Lifelong-PT, $\lambda = 1$	48.02	45.35
Lifelong-PT, $\lambda = 2$	46.09	47.30
Lifelong-PT, $\lambda = 4$	48.08	48.67
Mix-CPT, $\alpha = 0.60$	45.32	45.80
Mix-CPT, $\alpha = 0.80$	45.87	47.83
Mix-CPT, $\alpha = 0.90$	45.94	51.72
Mix-CPT, $\alpha = 0.95$	45.22	42.75

Table 6: Coefficient sensitivity for the two closest regularizer-style objective probes on PubMed D2:G3. λ scales the Lifelong-PT-style all-token KL term; α is the NTP weight in the Mix-CPT-style logit-swap objective; Mix-CPT rows use reverse KL. SDR denotes self-distilled replay. All runs use the same data, budget, and seed.

No tested coefficient point reaches self-distilled replay on both axes. The Lifelong-PT-inspired grid has a Pareto frontier split between $\lambda = 0.5$ (best PubMedQA) and $\lambda = 4$ (best General), while the Mix-CPT-inspired grid peaks at $\alpha = 0.90$ and remains below pure NTP on General. These

single-seed grids provide sensitivity evidence, not a claim that the original methods are fully tuned and decisively surpassed.

D Exact Matched Comparison Tables

Tables 7–10 report the full benchmark-level matched comparisons summarized in the main text, including base-checkpoint reference columns. SDR denotes self-distilled replay. General avg. is the average over the five general-capability categories. PubMedQA uses the answer-permutation score used in the corresponding trade-off plot.

Category-level retention. Across the reported settings, both NTP and SDR largely preserve Knowledge for SmolLM3 (-0.24 and $+0.20$ points relative to Base), while Code and Multilingual decline more; SDR reduces the mean drop in all five categories. For Qwen, NTP loses 6.00–10.05 points across the five categories, whereas SDR preserves Math ($+0.07$), limits the Reasoning loss to -1.28 , and reduces the losses in Knowledge, Code, and Multilingual to -3.44 , -3.93 , and -2.66 points.

Benchmark	Base	D1:G1.5		D1:G4		D1:G9		D2:G3		D2:G8		D2:G18	
		NTP	SDR	NTP	SDR	NTP	SDR	NTP	SDR	NTP	SDR	NTP	SDR
General avg.	48.69	46.19	47.38	44.13	46.25	48.21	48.30	46.10	48.19	45.98	46.51	48.48	48.80
PubMedQA	51.23	51.43	56.35	53.53	51.55	45.53	48.83	52.02	53.87	44.68	52.85	41.02	55.10
Knowledge avg.	33.96	31.23	32.44	37.07	31.06	32.76	36.84	35.43	37.62	32.72	33.19	37.03	35.69
MMLU	62.08	59.60	60.84	59.58	60.35	59.82	61.42	57.92	60.27	59.75	60.75	60.22	61.43
MMLU-Pro	7.71	7.91	9.26	31.14	3.66	5.47	29.78	33.92	35.70	11.69	11.23	31.13	23.76
C-Eval	50.68	46.35	47.28	47.83	46.47	47.17	47.70	44.52	46.83	46.34	49.19	49.00	48.45
GPQA-Diamond	32.32	27.27	29.80	31.82	29.80	35.35	29.29	26.77	30.30	30.81	28.79	29.80	28.79
SuperGPQA	17.00	15.00	15.00	15.00	15.00	16.00	16.00	14.00	15.00	15.00	16.00	15.00	16.00
Reasoning avg.	70.53	67.09	68.83	67.18	68.74	67.96	69.16	66.37	69.23	66.69	68.82	68.71	69.99
ARC-C	76.27	65.42	70.85	63.73	69.49	68.81	71.86	67.80	71.19	67.46	72.20	71.19	73.90
ARC-E	86.07	82.89	82.54	81.48	84.48	86.24	85.54	82.89	84.13	82.89	84.13	86.77	84.83
BBH	51.70	47.50	48.12	48.82	48.51	47.89	49.92	45.91	51.67	46.63	49.33	47.96	50.61
BoolQ	72.11	61.47	68.26	65.75	68.47	66.36	69.42	58.23	67.61	62.69	64.53	65.23	70.21
COPA	67.00	68.00	70.00	66.00	69.00	69.00	65.00	67.00	68.00	68.00	67.00	69.00	68.00
CommonsenseQA	70.11	67.16	68.96	67.08	67.65	67.32	69.86	67.57	67.90	66.91	69.78	67.90	69.86
HellaSwag	73.98	71.36	72.22	71.63	72.11	71.62	72.66	70.41	71.76	71.02	72.45	71.57	72.74
OpenBookQA	63.80	63.80	64.00	65.70	64.30	59.70	63.80	61.50	67.40	58.90	64.90	64.50	66.50
PIQA	78.56	78.02	78.13	77.69	78.67	78.84	78.62	77.58	77.80	78.07	78.02	77.75	78.07
WinoGrande	65.75	65.27	65.27	63.93	64.72	63.77	64.96	64.80	64.80	64.33	65.90	65.19	65.19
Math avg.	46.58	49.48	47.38	32.45	46.87	52.88	45.69	49.35	48.23	47.72	42.20	52.85	49.84
GSM8K	72.55	64.97	66.57	27.90	65.73	67.55	61.18	62.70	67.85	56.63	64.59	67.10	69.67
MATH-500	20.60	34.00	28.20	37.00	28.00	38.20	30.20	36.00	28.60	38.80	19.80	38.60	30.00
Code avg.	47.73	42.25	46.30	42.58	42.45	45.70	45.52	39.77	43.83	41.58	45.42	42.15	44.70
HumanEval	41.50	30.50	39.00	34.10	32.90	37.80	36.00	29.30	34.80	29.30	37.80	34.10	34.10
HumanEval+	34.80	27.40	33.50	27.40	27.40	32.30	30.50	25.60	30.50	23.80	31.70	28.70	29.90
MBPP	63.00	61.40	61.10	59.30	61.10	61.60	63.00	56.10	60.30	61.90	60.60	57.70	62.20
MBPP+	51.60	49.70	51.60	49.50	48.40	51.10	52.60	48.10	49.70	51.30	51.60	48.10	52.60
Multilingual avg.	44.64	40.88	41.97	41.38	42.13	41.75	44.30	39.59	42.04	41.18	42.89	41.64	43.80
MGSM	43.40	38.08	39.60	39.40	40.12	39.96	43.56	37.68	39.64	39.68	41.44	40.68	43.08
MMMLU	45.87	43.68	44.34	43.36	44.15	43.54	45.04	41.50	44.44	42.68	44.34	42.61	44.53

Table 7: Full SmolLM3 source-replay matched comparisons for D1 and D2 settings, with the SmolLM3 base checkpoint as reference. Category averages and General avg. use the same aggregation as the main text.

Benchmark	Base	D4:G1		D4:G2		D4:G4		D4:G6		D4:G16	
		NTP	SDR	NTP	SDR	NTP	SDR	NTP	SDR	NTP	SDR
General avg.	48.69	44.63	43.18	42.18	44.04	41.29	44.34	44.84	48.02	45.75	45.48
PubMedQA	51.23	53.65	55.25	56.30	58.52	48.63	53.90	51.88	52.08	46.20	54.15
Knowledge avg.	33.96	32.78	31.52	30.84	30.72	34.11	33.95	34.90	36.79	32.08	35.99
MMLU	62.08	57.99	57.60	56.57	57.69	58.87	58.44	58.65	60.30	59.07	59.20
MMLU-Pro	7.71	17.43	11.73	12.40	10.00	27.36	23.59	26.32	30.61	7.80	27.28
C-Eval	50.68	45.18	43.48	45.47	42.11	44.58	44.45	45.21	48.74	45.22	47.66
GPQA-Diamond	32.32	29.29	30.81	26.77	29.80	24.75	28.28	29.29	29.29	33.33	29.80
SuperGPQA	17.00	14.00	14.00	13.00	14.00	15.00	15.00	15.00	15.00	15.00	16.00
Reasoning avg.	70.53	65.11	66.95	64.56	67.83	66.53	67.78	65.71	68.04	66.59	66.66
ARC-C	76.27	60.00	66.10	60.00	68.14	63.05	68.14	65.42	67.80	66.44	68.14
ARC-E	86.07	81.13	80.78	78.84	82.36	81.31	82.01	82.54	83.60	82.89	84.13
BBH	51.70	47.88	42.94	42.43	47.84	45.93	45.33	46.31	47.53	46.36	46.66
BoolQ	72.11	57.58	69.02	65.02	64.74	64.28	67.06	59.42	67.22	58.17	57.92
COPA	67.00	67.00	65.00	64.00	71.00	68.00	71.00	65.00	67.00	67.00	63.00
CommonsenseQA	70.11	65.03	66.26	64.70	66.18	66.26	67.08	67.24	67.57	68.39	67.73
HellaSwag	73.98	69.37	69.32	68.34	70.04	70.57	70.42	70.40	71.56	70.46	71.18
OpenBookQA	63.80	63.40	70.10	62.90	67.70	65.60	64.90	59.20	65.60	64.10	66.50
PIQA	78.56	76.33	76.71	76.71	77.53	77.20	77.09	77.09	77.37	77.53	76.99
WinoGrande	65.75	63.38	63.30	62.67	62.75	63.14	64.72	64.48	65.19	64.56	64.33
Math avg.	46.58	46.53	40.95	41.31	43.69	25.11	43.09	45.27	48.98	49.84	42.77
GSM8K	72.55	59.67	58.91	58.23	61.18	16.22	61.18	56.33	66.57	62.47	63.15
MATH-500	20.60	33.40	23.00	24.40	26.20	34.00	25.00	34.20	31.40	37.20	22.40
Code avg.	47.73	40.85	40.90	38.90	39.80	40.90	38.42	40.10	44.02	41.38	42.20
HumanEval	41.50	31.10	32.90	32.90	32.30	31.10	28.70	32.30	34.10	32.90	32.30
HumanEval+	34.80	26.80	28.00	27.40	27.40	26.20	25.00	26.80	29.90	26.80	28.00
MBPP	63.00	58.70	55.30	51.90	54.00	57.40	53.70	55.80	61.60	58.20	59.30
MBPP+	51.60	46.80	47.40	43.40	45.50	48.90	46.30	45.50	50.50	47.60	49.20
Multilingual avg.	44.64	37.87	35.56	35.29	38.17	39.80	38.45	38.22	42.26	38.87	39.78
MGSM	43.40	34.24	29.76	29.80	34.64	37.28	34.56	35.96	40.40	35.72	36.44
MMMLU	45.87	41.49	41.36	40.78	41.71	42.32	42.35	40.49	44.13	42.03	43.13

Table 8: Full SmolLM3 source-replay matched comparisons for D4 settings, with the SmolLM3 base checkpoint as reference. Category averages and General avg. use the same aggregation as the main text.

Benchmark	Base	D1:G9		D2:G3		D2:G8		D4:G6	
		NTP	SDR	NTP	SDR	NTP	SDR	NTP	SDR
General avg.	65.07	59.18	63.62	60.13	63.78	58.38	63.89	52.02	62.72
PubMedQA	66.92	63.80	70.03	63.75	70.17	63.33	68.35	67.15	68.98
Knowledge avg.	54.97	43.28	52.57	46.89	51.30	42.83	52.16	44.21	48.83
MMLU	75.21	71.72	74.61	71.88	73.64	71.28	74.13	71.52	73.07
MMLU-Pro	51.87	17.67	46.86	35.87	43.96	17.55	47.84	25.48	34.82
C-Eval	79.38	71.19	76.47	71.87	74.51	70.50	73.95	69.24	73.86
GPQA-Diamond	41.41	32.83	39.90	31.82	39.39	31.82	39.90	32.83	38.38
SuperGPQA	27.00	23.00	25.00	23.00	25.00	23.00	25.00	22.00	24.00
Reasoning avg.	78.41	72.61	77.78	72.73	76.93	72.31	77.27	71.47	77.05
ARC-C	90.85	84.41	89.49	84.07	88.14	82.71	89.15	83.39	88.81
ARC-E	96.65	94.71	96.12	92.42	95.41	94.53	95.24	93.65	95.77
BBH	68.35	63.29	64.10	63.76	64.71	62.95	64.06	62.41	63.33
BoolQ	78.62	58.35	81.77	64.01	80.34	61.01	80.83	55.78	77.13
COPA	78.00	70.00	76.00	66.00	75.00	69.00	75.00	63.00	78.00
CommonsenseQA	70.43	67.73	69.86	69.04	69.04	67.65	70.19	67.98	69.29
HellaSwag	71.27	69.96	70.91	69.20	70.33	69.27	70.64	68.94	70.07
OpenBookQA	85.80	77.30	85.30	79.70	83.80	76.10	85.10	78.90	85.70
PIQA	77.75	76.93	78.62	76.71	77.80	76.66	77.86	76.99	77.97
WinoGrande	66.38	63.46	65.67	62.35	64.72	63.22	64.64	63.69	64.40
Math avg.	67.28	67.50	68.10	68.56	72.06	65.92	68.87	34.01	70.77
GSM8K	74.15	79.61	81.80	79.91	82.71	75.44	82.34	15.01	80.74
MATH-500	60.40	55.40	54.40	57.20	61.40	56.40	55.40	53.00	60.80
Code avg.	64.28	56.02	60.85	57.12	60.27	54.40	62.17	55.17	58.58
HumanEval	63.40	49.40	59.80	52.40	57.90	47.60	62.80	49.40	53.70
HumanEval+	56.10	42.70	52.40	47.00	49.40	40.90	51.80	42.70	45.70
MBPP	74.60	71.40	71.40	70.10	73.50	70.10	73.00	70.90	74.10
MBPP+	63.00	60.60	59.80	59.00	60.30	59.00	61.10	57.70	60.80
Multilingual avg.	60.43	56.45	58.79	55.37	58.34	56.46	58.99	55.25	58.39
MGSM	61.89	58.36	59.71	56.22	60.00	57.93	60.73	56.44	60.36
MMMLU	58.97	54.54	57.87	54.53	56.67	54.99	57.26	54.07	56.41

Table 9: Full Qwen3-4B non-source replay matched comparisons, with the base checkpoint as reference. Both methods use the same non-source replay corpus at each token budget.

Benchmark	Base	D1:G9		D2:G3		D2:G8		D4:G6	
		NTP	SDR	NTP	SDR	NTP	SDR	NTP	SDR
General avg.	69.89	61.94	66.34	63.29	68.96	61.98	65.56	61.64	67.00
PubMedQA	71.98	69.27	73.50	69.98	73.05	72.80	73.33	70.98	71.85
Knowledge avg.	57.87	50.71	54.23	52.37	56.03	49.59	53.99	52.32	54.74
MMLU	78.59	74.66	76.24	76.05	77.82	75.18	76.69	73.87	76.26
MMLU-Pro	53.39	39.68	48.72	43.37	51.63	34.88	48.49	43.06	49.37
C-Eval	83.44	74.84	78.78	77.05	81.28	76.53	79.34	75.72	78.15
GPQA-Diamond	43.94	39.39	40.40	40.40	42.42	36.36	40.40	44.95	42.93
SuperGPQA	30.00	25.00	27.00	25.00	27.00	25.00	25.00	24.00	27.00
Reasoning avg.	81.92	75.99	80.89	76.15	80.82	75.49	80.04	74.50	80.29
ARC-C	91.53	89.49	90.51	89.83	90.17	87.46	91.19	86.44	91.53
ARC-E	97.18	95.59	97.18	95.94	96.30	95.77	95.24	95.24	97.18
BBH	74.31	69.15	69.65	67.00	71.38	69.41	67.76	67.60	69.66
BoolQ	80.18	60.83	82.11	66.21	81.04	60.31	78.04	62.81	77.74
COPA	87.00	72.00	85.00	74.00	81.00	72.00	81.00	70.00	82.00
CommonsenseQA	74.12	71.74	73.46	69.78	73.55	70.27	73.05	69.53	72.24
HellaSwag	76.12	72.17	75.01	70.89	75.42	72.51	74.68	71.22	75.08
OpenBookQA	89.10	83.10	87.20	85.10	89.70	83.30	89.40	79.20	88.20
PIQA	79.76	79.00	79.38	78.02	80.41	78.40	80.36	77.48	79.87
WinoGrande	69.93	66.85	69.38	64.72	69.22	65.51	69.69	65.51	69.38
Math avg.	75.81	69.11	72.00	71.09	75.61	69.34	72.35	68.07	73.14
GSM8K	86.43	80.21	82.79	82.18	84.23	80.89	81.50	80.14	84.69
MATH-500	65.20	58.00	61.20	60.00	67.00	57.80	63.20	56.00	61.60
Code avg.	67.95	55.12	63.08	58.08	68.38	56.70	60.00	55.88	64.17
HumanEval	69.50	50.60	64.00	56.70	68.90	52.40	56.10	50.00	64.00
HumanEval+	63.40	43.90	55.50	49.40	61.00	45.10	50.00	41.50	54.90
MBPP	74.90	68.30	72.20	70.40	77.50	70.60	72.00	73.30	75.40
MBPP+	64.00	57.70	60.60	55.80	66.10	58.70	61.90	58.70	62.40
Multilingual avg.	65.89	58.77	61.49	58.77	63.94	58.76	61.43	57.45	62.67
MGSM	69.85	61.05	63.38	59.38	66.76	60.11	62.65	58.62	65.35
MMMLU	61.93	56.48	59.60	58.16	61.11	57.41	60.21	56.29	59.99

Table 10: Full Qwen3-8B non-source replay matched comparisons, with the base checkpoint as reference. Both methods use the same non-source replay corpus at each token budget.