

PRISM: Private Rank-Invariant Subspace Memory for Federated Continual Fine-Tuning under Local Differential Privacy

Ziyi Wang¹, Liang Liu², Zhu Meng^{1,3}, Zhicheng Zhao^{1,3}

¹School of Artificial Intelligence, Beijing University of Posts and Telecommunications

²Beijing Academy of Blockchain and Edge Computing

³Key Laboratory of Interactive Technology and Experience System,
Ministry of Culture and Tourism

Correspondence: zhaozc@bupt.edu.cn

Abstract

Privacy-preserving adaptation of pretrained language models requires continual fine-tuning as private client distributions shift, while protecting individual samples under local differential privacy (LDP). In streaming differentially private federated learning, full-rank LoRA adapters will expose many trainable directions to DP noise, and later domains may overwrite directions relevant to earlier domains. We propose *Private Rank-Invariant Subspace Memory (PRISM)*, a fixed-rank LoRA framework that schedules the active plastic rank over the stream and applies shrinkage-gated memory writes to DP-protected aggregate updates. Server-side consolidation retains high-SNR directions in an orthogonal memory and performs pure post-processing on LDP updates. On MNLI genre and Sentiment3, PRISM stabilizes adaptation under strict privacy budgets while preserving an effective public initialization. At $\epsilon = 2$, it reaches 79.8% and 92.3% final average accuracy, compared with 52.3% and 85.0% for full-rank DP-FedSVD. Ablations indicate that rank scheduling accounts for most strict-budget recovery; shrinkage-gated memory contributes a slight retention gain and modestly outperforms the best static low-rank method on MNLI.

1 Introduction

Adapting pretrained language models in privacy-preserving deployment scenarios typically requires continuous learning from private client data. In many deployments, these data are distributed across clients and arrive over time, rather than as a centrally collected static training set. We study federated continual fine-tuning under record-level local differential privacy (LDP). Each client observes a sequence of related data distributions for a single prediction task. Each local update is produced through differentially private stochastic gradient descent (DP-SGD), and a task-aligned public checkpoint supplies a frozen backbone and classifier head. The setting is domain-incremental rather

than class-incremental: the label interface is fixed by public supervision, while private input distributions continue to shift. The core challenge is to adapt to each new private distribution while maintaining performance on historical distributions under one privacy budget accumulated over the entire stream. This formulation differs from static private adaptation in two operational constraints: the server cannot use unprotected validation signals to select directions for retention, and the fixed adapter rank leaves no task-specific capacity that is allowed to grow with the stream. The allocation rule is therefore part of the training protocol, and all boundary decisions remain post-processing of already privatized updates.

The mainstream paradigm for private language-model adaptation combines federated optimization with parameter-efficient fine-tuning. DP-FedAvg protects local records but updates a large parameter set under DP noise, which substantially reduces utility under strict budgets (McMahan et al., 2017a,b; Kairouz et al., 2021). Low-Rank Adaptation (LoRA)-based DP fine-tuning reduces the trainable subspace by freezing the pretrained backbone (Hu et al., 2021; Yu et al., 2022). FedSVD further freezes one LoRA factor during local training and uses server-side SVD re-parameterization of the DP-aggregated adapter update, thereby improving performance in static private federated adaptation (Lee et al., 2025).

A crucial limitation remains: static low-rank private adaptation methods do not allocate feature directions between plastic adaptation and cross-boundary retention. Under strict DP, this issue creates two coupled failure modes. First, an excessively large active adapter subspace amplifies privacy noise relative to task-relevant signal. Second, SVD-derived directions from later distributions can overwrite representation directions that remain beneficial to earlier distributions. In our experiments, these failures appear as retention decay on previous

domains and, at the most restrictive budgets, single-class prediction collapse. Static rank reduction is a strong regularizer and an important baseline. Our objective is to preserve this DP benefit while keeping total adapter capacity fixed and filtering boundary directions before they become persistent state.

To address this problem, we propose *PRISM*, a fixed-rank LoRA framework for private federated continual adaptation. *PRISM* treats the fixed LoRA rank as a capacity budget allocated over time. It separates each adapter into memory and plastic subspaces, trains only the current plastic factor on clients, and applies server-side boundary consolidation to retain high-SNR feature directions while discarding low-SNR components. The rank schedule is the primary noise-control mechanism, and shrinkage-gated memory provides an additional retention mechanism for directions that survive the DP noise floor. Because boundary consolidation uses only DP-protected aggregates and public or synthetic noise estimates, it introduces no additional privacy-consuming signals beyond client-side DP-SGD.

Contributions. We make three contributions.

(1) We propose *PRISM*, a fixed-rank LoRA framework for federated continual learning under record-level LDP. It decomposes the adapter capacity into plastic and memory subspaces, allowing the same rank budget to adapt to new private domains while preserving valuable feature directions from earlier ones.

(2) We introduce a stream-boundary consolidation procedure for DP-protected aggregates. The procedure maintains plastic updates in the existing memory complement and, after each memory write, confines the next plastic basis to the updated complement. It also applies noise-floor singular-value shrinkage and uses only retained directions to update memory and initialize the next plastic subspace.

(3) Experiments on MNLI genre (Williams et al., 2018) and Sentiment3 show that *PRISM* consistently improves accuracy and retention under strict DP over full-rank DP-FedSVD and an O-LoRA-style orthogonal-adapter baseline. At $\epsilon = 2$, it raises final average accuracy over DP-FedSVD from 52.3% to 79.8% on MNLI genre and from 85.0% to 92.3% on Sentiment3, while reducing client-to-server uplink by 42.5%.

2 Related Work

Private and federated parameter-efficient adaptation. DP-SGD (Abadi et al., 2016) and Rényi-DP (RDP) accounting (Mironov, 2017) provide the privacy mechanism used in private fine-tuning. Full-model private fine-tuning, however, suffers from high computational cost and substantial DP-noise interference under strict budgets. Parameter-efficient private fine-tuning alleviates this issue by updating a small number of parameters in a pre-trained model, including LoRA (Hu et al., 2021), DP language-model fine-tuning (Yu et al., 2022), DP-BiTFFiT (Bu et al., 2022), and DP-LoRA variants (Liu et al., 2025; Xu et al., 2026, 2025). In federated parameter-efficient adaptation, FLoRA supports heterogeneous client LoRA ranks through stacking-based aggregation (Wang et al., 2024), FedEx-LoRA corrects inexact LoRA aggregation through a residual update to the frozen backbone (Singhal et al., 2025), and FedSVD uses server-side SVD re-parameterization so that clients train only one LoRA factor (Lee et al., 2025). Although these methods reduce the parameter dimensionality affected by DP noise or improve federated aggregation, they primarily target single-task or static adaptation. In contrast, *PRISM* targets a continual private federated stream and allocates the trainable subspace among successive private distributions under a globally accumulated privacy budget.

Continual subspace memory under privacy constraints. Continual learning methods use regularization, replay, gradient projection, prompts, or adapters to reduce forgetting (Kirkpatrick et al., 2017; Saha et al., 2021; Wang et al., 2022; Smith et al., 2023). Among subspace-based parameter-efficient methods, O-LoRA is the most relevant one: it assigns sequential language tasks to orthogonal LoRA subspaces (Wang et al., 2023). Federated continual learning methods such as FedWeIT (Yoon et al., 2021), federated class-incremental learning (Dong et al., 2022), and FedET (Liu et al., 2023) address forgetting through inter-client transfer, distillation, replay, or compact task modules. These approaches generally target non-private or class-incremental settings and often rely on stored examples, private logits, or growing task-specific capacity. In contrast, *PRISM* assumes a fixed label interface, shifting private domains, record-level LDP, and memory writes that post-process DP-protected aggregates. Appendix F provides extended comparisons with these related methods.

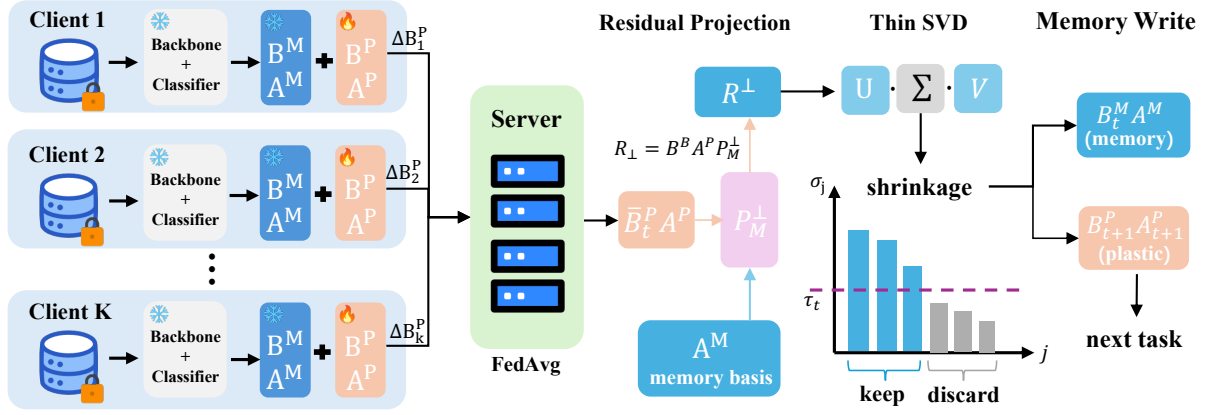


Figure 1: **PRISM workflow.** The four numbered stages correspond to Section 3. (1) FedAvg includes public task-aligned initialization, client-side plastic training with DP-SGD, and aggregation of DP-protected plastic updates into $\bar{B}_t^P A^P$. (2) Residual projection expresses the plastic update in the memory complement. (3) Thin SVD applies noise-floor shrinkage to keep high-SNR directions and discard low-SNR directions. (4) Memory write stores retained directions in $B_t^M A^M$ and initializes $B_{t+1}^P A_{t+1}^P$ for the next task.

3 Method

3.1 Problem Formulation

Consider K clients and a predefined stream of M distribution segments $\mathcal{T} = (T_1, \dots, T_M)$ for a single prediction task. All segments share the same output interface and label semantics, but they differ in input distribution or domain. For segment t , client c owns local records $D_{c,t} = \{(x_i, y_i) : y_i \in \mathcal{Y}\}$. Raw examples, private logits, and per-example gradients are not transmitted to the server under the LDP protocol. For consistency with continual-learning notation, we refer to segments as tasks hereafter. They are domain segments within one prediction task, rather than separate tasks with distinct label sets. Transition boundaries between segments are predefined. A public initialization provides a frozen backbone θ_0 and a frozen classifier head h .

Training is restricted to LoRA adapters. For each adapted layer l , the update applied to the frozen weight is $\Delta W_l = B_l A_l$, where $A_l \in \mathbb{R}^{r \times d_{in,l}}$, $B_l \in \mathbb{R}^{d_{out,l} \times r}$, and the total rank r is fixed over the whole stream. During segment t , the ideal federated objective is

$$\min_{\Phi_t} \sum_{c=1}^K p_{c,t} \mathbb{E}_{(x,y) \in D_{c,t}} \ell(h(f_{\theta_0, \Phi_t}(x)), y), \quad (1)$$

where $\Phi_t = \{A_l, B_l\}_l$ and $p_{c,t}$ is the client weighting used in FedAvg. After each segment, the protocol returns a single shared model for the task, rather than a domain-specific expert. The model is evaluated across all previously observed domains

in terms of final average accuracy and forgetting metrics. The privacy unit is a single data record belonging to a fixed client, and the privacy budget accumulates over the entire stream.

3.2 Overview of PRISM

As shown in Figure 1, PRISM allocates the fixed LoRA rank between retention and adaptation over time through four stages. The first stage, FedAvg, covers the public task-aligned initialization, client-side plastic training with DP-SGD, and aggregation of privatized plastic updates. At stream boundaries, the server then applies residual projection, thin SVD with shrinkage, and memory writing before initializing the next plastic subspace. For segment t , each LoRA layer is decomposed into a frozen auxiliary memory component and an active plastic component,

$$\Delta W_l = B_l^M A_l^M + B_l^P A_l^P, \quad (2)$$

where A_l^M, B_l^M store directions consolidated from previous segments, while A_l^P, B_l^P are reserved for the current segment. The rows of A_l^M and A_l^P are jointly orthonormal. Within a segment, clients only train B_l^P ; A_l^M, B_l^M , and A_l^P remain fixed. For inference, the single adapter uses both terms in Equation 2. After the final segment, the latest plastic component is retained without reset.

3.3 Client-Side Private Plastic Training

At the start of segment t , the server broadcasts the frozen memory factors and the current plastic basis $\{A_l^P\}_l$. A participating client optimizes

only $\{B_l^P\}_l$ on its local records. For each mini-batch, per-example gradients with respect to B^P are clipped to norm C , Gaussian noise with multiplier σ is added, and the client executes the local optimization updates. The number of rounds, local steps, clipping values, and batch sizes are specified in Section 4.

After local training, client c sends only its privatized plastic factor, or an equivalent plastic update, to the server. The server performs weighted FedAvg,

$$\bar{B}_{l,t}^P = \sum_{c \in S_t} w_{c,t} B_{l,c,t}^P, \quad (3)$$

over the selected clients S_t . Since A_l^P is fixed during the segment, the aggregated plastic adapter for layer l is $\bar{B}_{l,t}^P A_l^P$. Note that all subsequent server operations are pure post-processing of this DP-protected aggregate, allowing boundary filtering and memory updates without additional privacy loss.

3.4 Server-Side Boundary Consolidation

At a boundary after segment $t < M$, PRISM expresses the aggregated plastic update in the old memory complement and then writes shrinkage-retained directions into auxiliary memory. Let $P_M^\perp = I - (A_l^M)^\top A_l^M$ denote this complement. The residual form $R_l = \bar{B}_{l,t}^P A_l^P P_M^\perp$ explicitly expresses the plastic update in this complement; joint orthogonality gives $A_l^P P_M^\perp = A_l^P$. The server computes a thin SVD $R_l = U \Sigma V^\top$. Following the shrinkage test in Section 3.5, it writes the top k retained high-SNR directions into memory:

$$\begin{aligned} A_l^M &\leftarrow [A_l^M; V_k^\top], \\ B_l^M &\leftarrow [B_l^M; U_k \text{diag}(\tilde{s}_{1:k})]. \end{aligned} \quad (4)$$

The boundary procedure then recomputes the complement after the memory write and samples the next A_l^P inside it. This construction prevents future plastic directions from immediately overwriting consolidated memory. Thus, consolidated directions remain protected while the remaining rank is left available for future adaptation. Algorithm 1 presents the layer-wise procedure.

3.5 Rank Scheduling and Noise-Floor Shrinkage

Let the total rank be fixed at $r = r_M + r_P$, and reserve at least $r_P^{\min}$ plastic directions for every non-final segment. Before training T_1 , $r_M = 0$

Algorithm 1: PRISM stream-boundary processing.

Input: layer factors $A_l^M, B_l^M, A_l^P, \bar{B}_{l,t}^P$;
per-boundary write capacity $m_t^{\max}$; calibrated noise scale σ , total rank r ; shrinkage coefficient γ .

Output: updated A_l^M, B_l^M and initialized A_l^P, B_l^P for the next segment.

```

1  $P_M^\perp \leftarrow I - (A_l^M)^\top A_l^M$ ; // old memory complement
2  $R_l \leftarrow \bar{B}_{l,t}^P A_l^P P_M^\perp$ ;
3  $U, \Sigma, V \leftarrow \text{THINSVD}(R_l)$ ;
4  $\tau_l \leftarrow \text{NOISEFLOOR}(\sigma, P_M^\perp)$ ; // 0.95-quantile MC estimate
5 for  $j = 1, \dots, \text{rows}(A_l^P)$  do
6    $\tilde{s}_j \leftarrow \sqrt{\max(\Sigma_{jj}^2 - \gamma \tau_l^2, 0)}$ ;
7 end
8  $k \leftarrow \min(m_t^{\max}, |\{j : \tilde{s}_j > 0\}|)$ ;
9 if  $k > 0$  then
10    $A_l^M \leftarrow [A_l^M; V_{1:k}]$ ;
11    $B_l^M \leftarrow [B_l^M; U_{1:k} \text{diag}(\tilde{s}_{1:k})]$ ;
12 end
13  $P_{M,\text{new}}^\perp \leftarrow I - (A_l^M)^\top A_l^M$ ;
14  $r_P^{\text{next}} \leftarrow r - \text{rows}(A_l^M)$ ;
15 sample  $A_l^P \in \mathbb{R}^{r_P^{\text{next}} \times d_{\text{in}}}$  from  $\text{Im}(P_{M,\text{new}}^\perp)$ ;
16  $B_l^P \leftarrow \mathbf{0}$ ;

```

and $r_P = r$. After segment $t < M$, the maximum number of directions that can be written is

$$m_t^{\max} = \left\lceil \frac{r - r_P^{\min} - r_M^{(t)}}{M - t} \right\rceil, \quad (5)$$

where $r_M^{(t)} = \text{rows}(A_l^M)$ before the boundary. The rule distributes the remaining non-plastic capacity across future boundaries rather than allocating all available rank to early memory. Once r and $r_P^{\min}$ are fixed, the schedule is deterministic.

To avoid storing DP noise as memory, the server estimates a layer-wise spectral noise floor by simulating an equivalent aggregated DP noise matrix N_l and projecting it through the current memory complement: $\tau_l = Q_q(\|N_l P_M^\perp\|_2)$ with $q = 0.95$. N_l is constructed solely from public quantities such as σ , clipping, steps, and aggregation weights. Note that τ_l is a relative scale, not a formal DP utility bound; the formal privacy guarantee still derives from client-side DP-SGD. With s_j denoting the singular values of R_l , the shrinkage rule is

$$\tilde{s}_j = \sqrt{\max(s_j^2 - \gamma \tau_l^2, 0)}, \quad (6)$$

and only directions with $\tilde{s}_j > 0$ are eligible for memory insertion. The shrinkage coefficient γ is fixed before the main runs; Appendix L reports the diagnostic sweep used to choose it.

3.6 Privacy and Cost across FedAvg, Residual Projection, Thin SVD, and Memory Write

PRISM provides *record-level LDP* for local examples within each client. Neighboring datasets differ in one example within one fixed client; client-level DP is not claimed. A separate RDP accountant is maintained per client over the full stream, and the maximum achieved ε among all clients at $\delta = 10^{-5}$ is reported as the formal privacy guarantee. The mean per-client ε is reported only as a diagnostic metric for the sampled participation schedule. The privacy budget is not reset between segments. The accountant binary search set the upper bound of σ to 64; when this limit is binding, the achieved ε is reported rather than the preset target value.

The number of trainable and uploaded parameters will scale with the plastic rank instead of the total rank. For segment t , the per-round uplink consists of $\sum_l d_{\text{out},l} r_P^{(t)}$ adapter parameters. The downlink contains the fixed-rank adapter state needed for inference and the next local update, thus scaling with r rather than with the full backbone. The boundary SVD is computationally lightweight since the rank of R_l is bounded by $r_P^{(t)}$; it is computed through QR reductions and an $r_P^{(t)} \times r_P^{(t)}$ SVD. Hence, the additional cost compared with DP-FedSVD lies in memory bookkeeping and noise-floor Monte Carlo computation, while the dominant local training cost decreases as r_P shrinks over the stream.

4 Experimental Setup

Research questions. The experiments explore four research questions: whether PRISM improves final accuracy compared with existing private FL references under the same record-level LDP budget (RQ1), whether it reduces forgetting and single-class collapse under intense noise (RQ2), whether the gains stem from rank scheduling and shrinkage-gated auxiliary memory rather than extra capacity or privacy budget (RQ3), and whether such trends hold across stream order and a second domain stream (RQ4).

Datasets. The experiments use human-labeled English text classification benchmarks and do not collect or redistribute new text. *MNLI genre* is a five-segment domain stream for the same NLI task, built from MNLI (Williams et al., 2018): *fiction* $\rightarrow$ *government* $\rightarrow$ *slate* $\rightarrow$ *telephone* $\rightarrow$ *travel*, plus the reverse order. Each genre is split 90/10 from

the MNLI train split and capped at 2,000 train and 1,000 evaluation examples. The public initialization is RoBERTa-large (Liu et al., 2019) warmed up on 50,000 SNLI (Bowman et al., 2015) examples. *Sentiment3* consists of SST-2 (Socher et al., 2013; Wang et al., 2018) $\rightarrow$ IMDb (Maas et al., 2011) $\rightarrow$ Yelp Polarity (Zhang et al., 2015), with 2,000/1,000 class-balanced train/evaluation examples per task and a Rotten-Tomatoes (Pang and Lee, 2005) public warmup. Dataset construction, licenses, and preprocessing details are in Appendix C.

Baselines. All private methods use the same public initialization, private domain stream, client schedule, calibrated noise multiplier, optimizer, and LoRA target modules. The comparison includes Public-Init (no private FL), DP-FedAvg (train both LoRA factors), DP-FFA (freeze A_l , train B_l), DP-FedSVD (Lee et al., 2025; freeze A_l locally and SVD-refactorize after aggregation), an O-LoRA-style orthogonal continual-adapter baseline, and PRISM. Since gains under strict DP can be explained by lower active rank, Section 6 includes static low-rank DP-FedSVD and rank-scheduled FedSVD as primary controls rather than treating them solely as auxiliary diagnostics. Appendix E lists the shared settings and method-specific differences.

Metrics. Let $\text{Acc}_{i,j}$ be accuracy on segment j after training through segment i . The main metric is final average accuracy $\overline{\text{Acc}} = \frac{1}{M} \sum_j \text{Acc}_{M,j}$. The evaluation also reports average forgetting, $\frac{1}{M-1} \sum_{j < M} (\max_{i < M} \text{Acc}_{i,j} - \text{Acc}_{M,j})$, and inspects task-retention matrices and maximum-class prediction share to separate true retention from collapsed predictors.

Implementation and privacy. Unless otherwise stated, all private runs use RoBERTa-large with LoRA rank of 8, alpha of 8, dropout rate of 0.05, and adapters on query/value projections. The simulations use six clients, Dirichlet label skew $\alpha = 0.5$, 50% participation, 50 rounds per segment, five local steps, SGD with learning rate 5×10^{-3} , maximum sequence length of 128, and logical batch size of 64. DP-SGD uses $\delta = 10^{-5}$ and clipping norm $C = 1.0$ on MNLI and $C = 2.0$ on Sentiment3, with one calibrated noise multiplier per privacy/seed cell. For PRISM, the minimum plastic rank is $r_P^{\min} = 2$, thereby yielding the deterministic schedule in Section 3.5: $(r_M, r_P) = (0, 8) \rightarrow (2, 6) \rightarrow (4, 4) \rightarrow (5, 3) \rightarrow (6, 2)$ for MNLI

Target ϵ	σ	Achieved max ϵ	Mean client ϵ	σ cap
<i>MNLI</i>				
6	6.27	6.000	2.32	no
3	11.99	3.000	1.18	no
2	17.55	2.000	0.79	no
<i>Sentiment3</i>				
3	25.03	3.000	1.39	no
2	36.02	2.000	0.93	no
1	64.00	1.065	0.50	yes

Table 1: Privacy accounting over three seeds, $\delta = 10^{-5}$. The MNLI strict-budget diagnostic is reported in Appendix L; the Sentiment3 stress row is σ -capped with achieved $\epsilon_{\max} \approx 1.07$.

and $(0, 8) \rightarrow (3, 5) \rightarrow (6, 2)$ for Sentiment3. The shrinkage coefficient is fixed by privacy regime from a complete single-seed diagnostic sweep (Appendix L): $\gamma_{\epsilon=6} = 0.25$ and $\gamma_{\epsilon \leq 3} = 1.0$. The privacy unit is one local record inside a fixed client; the per-client RDP accountant composes across the whole stream. Additional implementation details are provided in Appendix C.

Statistical protocol. Main results are reported as means with standard deviations. For MNLI, results are averaged over three seeds and two task orders, yielding six seed-order runs per privacy budget. Sentiment3 averages three seed runs in the forward order. For a fixed seed, all methods share the same data split, client partition, client participation schedule, public initialization, and noise multiplier. We evaluate the final model after the full stream, with no early stopping or test-set-based checkpoint selection.

Privacy accounting summary. Table 1 reports the calibrated noise multiplier, the max-client ϵ used for the formal guarantee, and the mean per-client ϵ used only as a participation diagnostic. Among the main-table cells, the σ -cap binds only at the Sentiment3 stress point, which is reported as σ -capped with achieved $\epsilon_{\max} \approx 1.07$. Appendix B elaborates the record-level LDP theorem, the post-processing argument, and the σ -cap details.

5 Main Results

The results are consistent with the research questions in Section 4: utility under the same LDP budget, stability against forgetting and collapse, and operating regimes in which PRISM provides limited additional utility. Mean differences and standard deviations are reported descriptively; no

Method	MNLI (Avg. Acc., %)			Sentiment3 (Avg. Acc., %)		
	$\epsilon = 6$	$\epsilon = 3$	$\epsilon = 2$	$\epsilon = 3$	$\epsilon = 2$	Stress*
Public-Init	79.4	79.4	79.4	88.9	88.9	88.9
DP-FedAvg	78.5	43.0	33.4	49.8	49.9	49.9
DP-FFA	80.8	73.5	51.1	90.3	82.0	50.4
DP-FedSVD	81.0	73.9	52.3	90.0	85.0	50.8
O-LoRA	80.6	77.8	56.2	90.9	62.2	50.1
PRISM (ours)	81.2	80.3	79.8	92.2	92.3	90.8

Table 2: **Final average accuracy under strict LDP.** Means are across seeds and, for MNLI, task orders; standard deviations and forgetting are in Appendix H. DP-FedSVD is the full-rank static reference; rank-scheduled and static low-rank controls are analyzed in Tables 6 and 7. *Sentiment3 σ -capped stress with achieved $\epsilon_{\max} \approx 1.07$.

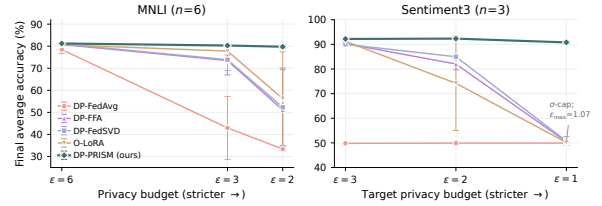


Figure 2: **Privacy-utility curves.** O-LoRA is included as an orthogonal continual-adaptor control. The Sentiment3 stress point is σ -capped with achieved $\epsilon_{\max} \approx 1.07$.

hypothesis tests are conducted, and no statistical-significance claims are made.

Final accuracy under the formal privacy budget.

Table 2 first compares PRISM with full-rank private FL references under the same record-level LDP accountant. On MNLI, PRISM outperforms full-rank DP-FedSVD by 0.2% at $\epsilon = 6$, indicating limited value from boundary consolidation when the privacy budget is less restrictive. The gap grows to 6.4% at $\epsilon = 3$ and 27.4% at $\epsilon = 2$, where full-rank DP-FedSVD averages 52.3% with high variance while PRISM remains at 79.8%. This comparison diagnoses the instability of full-rank static adaptation; it should not be interpreted as evidence that memory alone accounts for the gain. Section 6 isolates the stronger low-rank and rank-scheduled alternatives. On Sentiment3, PRISM improves over full-rank DP-FedSVD by 2.2% at $\epsilon = 3$, 7.4% at $\epsilon = 2$, and 40.0% in the σ -capped stress setting. The O-LoRA-style control is competitive under less restrictive budgets, but becomes unstable under stricter privacy: at $\epsilon = 2$, it drops to $56.2 \pm 21.3\%$ on MNLI and $62.2 \pm 21.9\%$ on Sentiment3 due to collapse in individual seed-order runs.

Public initialization and private learning. The MNLI public checkpoint is intentionally strong. At

Method	$\varepsilon = 3$	$\varepsilon = 2$
DP-FedAvg	0.79 $\pm$ 0.19	0.93 $\pm$ 0.11
DP-FFA	0.43 $\pm$ 0.05	0.74 $\pm$ 0.27
DP-FedSVD	0.41 $\pm$ 0.04	0.64 $\pm$ 0.21
PRISM (ours)	0.38 $\pm$ 0.03	0.38 $\pm$ 0.03
Balanced (1/3)	0.33	0.33

Table 3: **Maximum-class prediction share on MNLI.** Final maximum-class prediction share; values closer to the balanced reference indicate less concentration.

$\varepsilon = 2$, PRISM surpasses Public-Init by only 0.4%. Thus, the MNLI advantage lies in stability under private continual training rather than in a substantial gain over the public model. Appendix D reports a three-seed sensitivity study. Private adaptation brings a substantial gain for the undertrained SNLI-10k-e1 checkpoint, no reliable gain for the lowest-accuracy SNLI-5k-e1 checkpoint, and achieves modest gains as the public warmup approaches saturation.

Retention and collapse explain RQ2. Aggregate accuracy alone cannot distinguish retention of previous domains from adaptation to the latest domain. Figure 3 shows the MNLI task-retention matrix with $\varepsilon = 2$ for the forward order. Cell (i, j) is accuracy on task j after training reaches task i ; cells with $j > i$ are blank because the corresponding tasks are unseen at that stage. DP-FedAvg approaches chance-level performance by the second task. DP-FedSVD shows a directional overwrite effect, where subsequent tasks overwrite prior feature directions and rows below the diagonal degrade. The O-LoRA-style orthogonal-adapter control suffers from seed instability under strict LDP. In contrast, PRISM maintains steady performance across tasks and stages, verifying that its superiority is not limited to last-task adaptation.

Model failures under strict DP also manifest as class-biased collapse rather than uniform accuracy degradation. Table 3 shows the final maximum-class prediction share on MNLI; lower values, closer to the balanced reference 1/3, indicate less class concentration. At $\varepsilon = 2$, the values of DP-FedAvg and DP-FedSVD reach 0.93 and 0.64, respectively, whereas PRISM stays at 0.38 at both budgets, suggesting that rank control combined with shrinkage mitigates noise-driven class dominance.

Sentiment3 forgetting corroborates retention stability. Table 4 reports average forgetting on

Method	Sentiment3 forgetting (%)		
	$\varepsilon = 3$	$\varepsilon = 2$	Stress*
DP-FedAvg	17.72 $\pm$ 14.19	5.79 $\pm$ 6.63	0.02 $\pm$ 0.03 [†]
DP-FFA	2.20 $\pm$ 0.91	9.76 $\pm$ 2.49	2.88 $\pm$ 3.98 [†]
DP-FedSVD	2.51 $\pm$ 0.70	5.75 $\pm$ 1.83	1.31 $\pm$ 0.68 [†]
O-LoRA	− 0.78 $\pm$ 2.72	10.38 $\pm$ 16.19	0.06 $\pm$ 0.97 [†]
PRISM (ours)	0.31 $\pm$ 0.14	− 0.12 $\pm$ 0.56	0.86 $\pm$ 0.95

Table 4: Sentiment3 average forgetting. Lower is better. * σ -capped stress with achieved $\varepsilon_{\max} \approx 1.07$; [†]near-zero forgetting at the cap is caused by collapsed predictors rather than stable retention. Bold marks the best non-collapsed entry within a privacy setting.

Method	Uplink params	Downlink params	Round (s)
DP-FedAvg	2,313,035	5,409,138	5.79
DP-FedSVD	1,156,518	5,409,138	3.29
PRISM (ours)	664,998	5,409,138	3.35

Table 5: **Communication and runtime.** MNLI $\varepsilon = 2$ main runs, averaged across seeds and orders.

the binary stream. At $\varepsilon = 2$, DP-FedSVD exhibits 5.75% average forgetting, whereas PRISM achieves −0.12%. O-LoRA is competitive at $\varepsilon = 3$ but collapses in two seeds at $\varepsilon = 2$. Note that at the σ -cap, the low forgetting of the DP baselines reflects chance-level final accuracy in Table 2 rather than stable retention. Appendix K reports the per-seed range and forgetting trace.

Communication efficiency rather than per-round latency. Table 5 reports the cost of the MNLI $\varepsilon = 2$ main runs. PRISM trains only the current plastic factor B^P , reducing uplink parameters by 42.5% relative to full-rank DP-FedSVD. Running time remains comparable (3.35s vs. 3.29s). Thus, the efficiency advantage lies in communication volume rather than latency; static low-rank FedSVD can reduce uplink but yields lower accuracy (Table 7).

6 Ablations

The ablations address RQ3 by testing whether PRISM’s gains arise from the proposed mechanism rather than extra capacity or a less restrictive privacy budget. DP-FedSVD is augmented with +RS for rank scheduling, +RS+Mem for ungated memory, +RS+Shrink for shrinkage, and +RS+Mem+Shrink for the full mechanism. These internal controls use the same privacy accountant and public initialization. Additionally, a static low-rank DP-FedSVD sweep further tests whether training fewer parameters is sufficient.

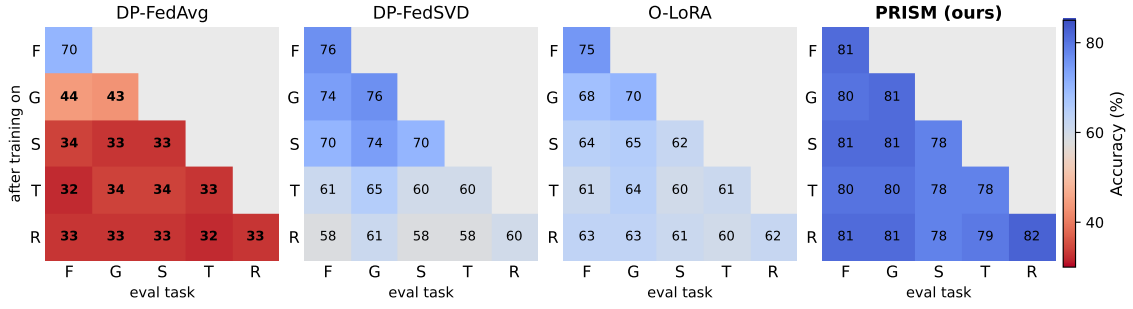


Figure 3: **MNLI retention at $\varepsilon = 2$.** Forward order, mean across seeds. Each panel reports accuracy after training reaches a given genre and evaluating on each observed genre. Tasks F/G/S/T/R denote fiction/government/slate/telephone/travel.

Variant	$\varepsilon=6$	$\varepsilon=3$	$\varepsilon=2$
DP-FedSVD	81.04 $\pm$ 0.55	73.88 $\pm$ 4.90	52.32 $\pm$ 17.78
+RS	80.91 $\pm$ 0.67	79.38 $\pm$ 0.78	77.55 $\pm$ 1.33
+RS+Mem	81.32 $\pm$ 0.65	78.41 $\pm$ 1.03	70.82 $\pm$ 7.59
+RS+Shrink	80.82 $\pm$ 0.57	79.97 $\pm$ 0.88	78.55 $\pm$ 0.53
+RS+Mem+Shrink	81.23 $\pm$ 0.65	80.31 $\pm$ 0.40	79.75 $\pm$ 0.76

Table 6: **MNLI module ablations.** Final average accuracy over forward and reverse orders with seeds 42/43/44 ($n=6$ for each cell).

Method	Uplink	Avg. Acc. (%)	Fgt (%)	BWT (%)	Δ_{pub} (%)
DP-FedSVD ($r=8$)	1.16M	54.59 $\pm$ 13.35	14.38 $\pm$ 8.29	-14.25 $\pm$ 8.34	-24.56 $\pm$ 13.05
DP-FedSVD ($r=2$)	0.29M	78.32 $\pm$ 1.02	1.04 $\pm$ 1.01	-0.80 $\pm$ 1.23	-0.84 $\pm$ 1.22
PRISM ($r=8$)	0.66M	79.68 $\pm$ 0.85	0.24 $\pm$ 0.18	-0.05 $\pm$ 0.26	0.53 $\pm$ 0.54

Table 7: **Strong static low-rank control.** MNLI $\varepsilon=2$, forward order, seeds 42–46 ($n=5$). Uplink is average client-to-server parameters per round.

Table 6 shows that rank scheduling is the dominant intervention under strict DP: at $\varepsilon=2$, +RS improves DP-FedSVD by 25.23% and reduces forgetting from 15.08% to 0.75%. Shrinkage remains effective under strong noise, improving over +RS by 1.00% at $\varepsilon=2$ and by 0.59% at $\varepsilon=3$, while keeping forgetting below 1%. Memory without shrinkage is unstable under strict DP: +RS+Mem drops to 70.82% at $\varepsilon=2$, whereas the full method recovers to 79.75%. These results support a constrained mechanism interpretation: rank scheduling dominates performance recovery, shrinkage functions as a noise filter, and memory contributes after shrinkage gates memory writing.

Static rank-reduction control. Table 7 shows the strongest competing explanation. Static DP-FedSVD with $r=2$ reaches $78.3 \pm 1.0\%$ at MNLI $\varepsilon=2$, only 1.36% below PRISM. The result supports a conservative interpretation: rank reduction accounts for most gains, and shrinkage-gated mem-

ory brings a modest retention improvement at fixed total rank.

7 Discussion

The results identify a specific operating regime. Under loose privacy, static DP-FedSVD is competitive and boundary processing brings only marginal improvement. Under strict privacy with a useful public initialization, rank scheduling reduces the directions exposed to DP-SGD, and shrinkage-gated memory prevents low-SNR boundary features from entering persistent state. Findings under extreme test scenarios cannot be generalized to practical deployment, which only reveals the collapse mode addressed by our method.

Rank scheduling accounts for most recovery. Ungated memory can preserve noisy singular directions, while shrinkage makes memory a retention store for filtered directions. This explains why static low-rank DP-FedSVD serves as a strong baseline but cannot fully match PRISM’s forgetting and backward-transfer capability.

PRISM uses no private validation data, logits, or growing task modules. The method is therefore designed for deployed classifiers with fixed label interfaces and known domain transitions. Its efficiency concerns communication rather than per-round latency.

8 Conclusion

PRISM addresses private continual LoRA where shifting client domains share a fixed label interface and one record-level LDP budget. It targets DP-noise amplification and overwriting of earlier directions. Scheduling plastic rank and writing shrinkage-filtered high-SNR directions improves strict-budget stability. Rank scheduling provides most gains, and memory adds modest retention.

Limitations

Scope and assumptions. The evidence is limited to English public benchmarks and two fixed-output text-classification domain streams: MNLI genres and Sentiment3. This work does not evaluate low-resource languages, multilingual transfer, generation, long-context tasks, retrieval-augmented systems, or class-incremental settings with disjoint label spaces. The capped per-segment private subsets and benchmark annotations cannot confirm consistent performance for larger private corpora, noisier labels, or deployment data with different domain shifts. PRISM also assumes known domain boundaries, a fixed classifier/label interface, a frozen classification head, and a useful public initialization. The rank schedule and shrinkage-gated memory update cannot solve label-space expansion or learn a classifier from scratch under DP.

Privacy and evaluation boundaries. The formal guarantee is record-level LDP for examples within a fixed client. Client-level DP is not provided for the addition or removal of an entire client. Server-side SVD, memory writing, rank scheduling, and shrinkage are post-processing of locally privatized updates; they do not strengthen the privacy guarantee beyond local DP-SGD. The noise-floor estimate uses only public quantities, but it is a heuristic utility filter rather than a formal DP utility bound. Means and standard deviations are measured over a limited number of seeds, and no significance tests are conducted.

Resources and risk. The experiments use RoBERTa-large and multi-run federated simulations, so they require nontrivial GPU resources. Software versions, privacy accounting, seeds, and communication costs are reported, and the implementation will be released after the review process. The method is intended to reduce exposure of private training records, but privacy-preserving adaptation can still be misused to deploy models on sensitive or biased data. Safety-sensitive deployment requires additional audits for bias, downstream misuse, and stakeholder impact.

Acknowledgments

This work was supported by the Key Research and Development Program of Xizang (XZ202601ZY0042), Beijing Advanced Innovation Center for Future Blockchain and Privacy Comput-

ing (GJJ-26) and the BUPT Innovation and Entrepreneurship Support Program (2025-YC-T033).

References

- Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. 2016. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, pages 308–318.
- Samuel R Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, pages 632–642.
- Zhiqi Bu, Yu-Xiang Wang, Sheng Zha, and George Karypis. 2022. Differentially private bias-term fine-tuning of foundation models. *arXiv preprint arXiv:2210.00036*.
- Arslan Chaudhry, Marc’Aurelio Ranzato, Marcus Rohrbach, and Mohamed Elhoseiny. 2018. Efficient lifelong learning with a-gem. *arXiv preprint arXiv:1812.00420*.
- Jiahua Dong, Lixu Wang, Zhen Fang, Gan Sun, Shichao Xu, Xiao Wang, and Qi Zhu. 2022. Federated class-incremental learning. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10154–10163. IEEE.
- Mehrdad Farajtabar, Navid Azizan, Alex Mott, and Ang Li. 2020. Orthogonal gradient descent for continual learning. In *International conference on artificial intelligence and statistics*, pages 3762–3773. PMLR.
- Charlie Hou, Akshat Shrivastava, Hongyuan Zhan, Ryan Conway, Trang Le, Adithya Sagar, Giulia Fanti, and Daniel Lazar. 2024. [PrE-text: Training language models on private federated data in the age of LLMs](#). In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 19043–19061. PMLR.
- Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.
- Peter Kairouz, H. Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Kallista Bonawitz, Zachary Charles, Graham Cormode, Rachel Cummings, Rafael G. L. D’Oliveira, Hubert Eichner, Salim El Rouayheb, David Evans, Josh Gardner, Zachary Garrett, Adrià Gascón, Badih Ghazi, Phillip B. Gibbons, and 40 others. 2021. [Advances and open problems in federated learning](#). *Found. Trends Mach. Learn.*, 14(1–2):1–210.

- 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.
- Seanie Lee, Sangwoo Park, Dong Bok Lee, Dominik Wagner, Haebin Seong, Tobias Bocklet, Juho Lee, and Sung Ju Hwang. 2025. [Fedsvd: Adaptive orthogonalization for private federated learning with lora](#). In *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pages 119733–119757. Curran Associates, Inc.
- Chenghao Liu, Xiaoyang Qu, Jianzong Wang, and Jing Xiao. 2023. Fedet: a communication-efficient federated class-incremental learning framework based on enhanced transformer. *arXiv preprint arXiv:2306.15347*.
- Xiao-Yang Liu, Rongyi Zhu, Daochen Zha, Jiechao Gao, Shan Zhong, Matt White, and Meikang Qiu. 2025. Differentially private low-rank adaptation of large language model using federated learning. *ACM Transactions on Management Information Systems*, 16(2):1–24.
- Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.
- David Lopez-Paz and Marc’Aurelio Ranzato. 2017. Gradient episodic memory for continual learning. *Advances in neural information processing systems*, 30.
- Andrew Maas, Raymond E Daly, Peter T Pham, Dan Huang, Andrew Y Ng, and Christopher Potts. 2011. Learning word vectors for sentiment analysis. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies*, pages 142–150.
- Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Agüera y Arcas. 2017a. [Communication-Efficient Learning of Deep Networks from Decentralized Data](#). In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54 of *Proceedings of Machine Learning Research*, pages 1273–1282. PMLR.
- H Brendan McMahan, Daniel Ramage, Kunal Talwar, and Li Zhang. 2017b. Learning differentially private recurrent language models. *arXiv preprint arXiv:1710.06963*.
- Ilya Mironov. 2017. Rényi differential privacy. In *2017 IEEE 30th computer security foundations symposium (CSF)*, pages 263–275. IEEE.
- Bo Pang and Lillian Lee. 2005. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *Proceedings of the 43rd annual meeting of the association for computational linguistics (ACL’05)*, pages 115–124.
- Gobinda Saha, Isha Garg, and Kaushik Roy. 2021. Gradient projection memory for continual learning. *arXiv preprint arXiv:2103.09762*.
- Raghav Singhal, Kaustubh Ponkshe, and Praneeth Vepakomma. 2025. [FedEx-LoRA: Exact aggregation for federated and efficient fine-tuning of large language models](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1316–1336, Vienna, Austria. Association for Computational Linguistics.
- James Seale Smith, Leonid Karlinsky, Vyshnavi Gutta, Paola Cascante-Bonilla, Donghyun Kim, Assaf Arbel, Rameswar Panda, Rogerio Feris, and Zolt Kira. 2023. Coda-prompt: Continual decomposed attention-based prompting for rehearsal-free continual learning. In *2023 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pages 11909–11919. IEEE.
- Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 1631–1642.
- Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. 2018. Glue: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP workshop BlackboxNLP: Analyzing and interpreting neural networks for NLP*, pages 353–355.
- Xiao Wang, Tianze Chen, Qiming Ge, Han Xia, Rong Bao, Rui Zheng, Qi Zhang, Tao Gui, and Xuanjing Huang. 2023. [Orthogonal subspace learning for language model continual learning](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10658–10671, Singapore. Association for Computational Linguistics.
- Zifeng Wang, Zizhao Zhang, Chen-Yu Lee, Han Zhang, Ruoxi Sun, Xiaoqi Ren, Guolong Su, Vincent Perot, Jennifer Dy, and Tomas Pfister. 2022. Learning to prompt for continual learning. In *2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pages 139–149. IEEE.
- Ziyao Wang, Zheyu Shen, Yexiao He, Guoheng Sun, Hongyi Wang, Lingjuan Lyu, and Ang Li. 2024. Flora: Federated fine-tuning large language models with heterogeneous low-rank adaptations. *Advances in Neural Information Processing Systems*, 37:22513–22533.

- Adina Williams, Nikita Nangia, and Samuel R Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long papers)*, pages 1112–1122.
- Honghui Xu, Shiva Shrestha, Wei Chen, Zhiyuan Li, and Zhipeng Cai. 2025. Dp-fedlora: Privacy-enhanced federated fine-tuning for on-device large language models. In *2025 IEEE International Conference on Data Mining (ICDM)*, pages 813–822. IEEE.
- Jie Xu, Karthikeyan Saravanan, Rogier van Dalen, Haaris Mehmood, David Tuckey, and Mete Ozay. 2026. Dp-dylora: Fine-tuning transformer-based models on-device under differentially private federated learning using dynamic low-rank adaptation. In *ICC 2026-IEEE International Conference on Communications*, pages 1–7. IEEE.
- Jaehong Yoon, Wonyong Jeong, Giwoong Lee, Eunho Yang, and Sung Ju Hwang. 2021. Federated continual learning with weighted inter-client transfer. In *International conference on machine learning*, pages 12073–12086. PMLR.
- Da Yu, Saurabh Naik, Arturs Backurs, Sivakanth Gopi, Huseyin A Inan, Gautam Kamath, Janardhan Kulkarni, Yin Tat Lee, Andre Manoel, Lukas Wutschitz, Sergey Yekhanin, and Huishuai Zhang. 2022. [Differentially Private Fine-Tuning of Language Models](#). In *International Conference on Learning Representations*.
- Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. *Advances in neural information processing systems*, 28.

Appendix Overview

Appendix A adds mechanism-level interpretation and practical guidance beyond the main discussion. Appendix B states the record-level LDP guarantee for PRISM, including client-side RDP composition, the server-side post-processing argument, and the interpretation of the σ cap. Appendix C records dataset construction, preprocessing, metric definitions, software versions, seeds, and checkpoint-selection protocol. Appendix D reports public-warmup sensitivity, and Appendix E documents baseline fairness controls. Appendix F expands the comparison with private parameter-efficient adaptation, federated LoRA aggregation, and continual subspace methods. Appendix G reports the SST-2/Rotten Tomatoes overlap check and decontaminated Sentiment3 results. The remaining appendices provide expanded results, MNLI order robustness, class-collapse diagnostics, Sentiment3 seed ranges, and stress and shrinkage diagnostics.

A Additional Discussion

Mechanism interpretation. Strict LDP changes the dominant failure mode of federated LoRA adaptation. The problem is not limited to stochastic gradient noise at client training time. Noisy high-dimensional adapter directions may also become persistent after server-side re-parameterization. PRISM separates this issue into two operations. Rank scheduling reduces the number of trainable directions exposed to client-side DP-SGD at later segments, and shrinkage prevents low-SNR boundary directions from being written into memory as task-relevant signal. Memory then serves as an auxiliary retention store for filtered directions. This interpretation is consistent with the main results: the largest gains occur under strict privacy, whereas DP-FedSVD remains stable under less restrictive budgets.

Relation to prior design principles. The results are consistent with private parameter-efficient adaptation and continual subspace learning, while delimiting their applicability. Private parameter-efficient adaptation and DP-FedSVD show that reducing the locally trained parameter dimension is important under DP. Continual subspace methods show that orthogonal directions can improve retention. The O-LoRA-style control demonstrates that orthogonality alone is insufficient under strict LDP: it is competitive under less restrictive budgets but

becomes high-variance at $\varepsilon = 2$ and collapses in the Sentiment3 stress setting. PRISM combines these principles in a fixed-rank federated adapter by training a scheduled plastic subspace, filtering boundary directions against a DP noise floor, and storing only the retained directions.

Applicability. PRISM targets strict private continual streams rather than all static private LoRA regimes. When the privacy budget is less restrictive, static DP-FedSVD is already competitive and the boundary logic has limited additional value. When the budget is extremely strict, the accountant can enter the σ -cap regime; those results are reported with achieved rather than nominal ε . The method also depends on a usable public initialization. If the public model is near chance-level on the private label interface, the private stage becomes classifier learning under DP, which is outside the intended LoRA adaptation setting.

Future directions. Future work can replace the empirical shrinkage coefficient with a calibration rule tied to clipping, participation, non-IIDness, and local steps. It can also study adaptive memory allocation across layers and stream segments instead of the current deterministic schedule, and extend the privacy model to client-level DP when the presence of an entire client must be protected.

B Privacy Guarantee of PRISM

This appendix states the formal privacy guarantee for PRISM under per-client record-level LDP. The public SNLI and Rotten-Tomatoes warmups are excluded from the private budget because they use public data. The private stage isolates the privacy-consuming operation, client-side DP-SGD, from server-side post-processing, which operates solely on privatized client updates.

B.1 Notation

Consider K clients. Client k owns a private dataset D_k . Record-level neighboring datasets $D_k \sim D'_k$ are considered, where D_k and D'_k differ in one training example. The global private dataset is $D = (D_1, \dots, D_K)$.

The private continual stream contains T tasks. Task t has R_t communication rounds. Let $z_{k,t,r} \in \{0, 1\}$ indicate whether client k participates in round r of task t . The client participation schedule is sampled independently of private records and is fixed before training.

When selected, client k runs $S_{k,t,r}$ DP-SGD steps. At local step s , examples are sampled at rate $q_{k,t,r,s}$. Under the fixed-batch approximation used by the accountant,

$$q_{k,t,r,s} \approx \frac{b}{|D_{k,t}|},$$

where b is the logical batch size and $D_{k,t}$ is the local dataset of client k for task t . Each per-example gradient is clipped to ℓ_2 norm C , and Gaussian noise with multiplier σ is added. Let

$$\rho_{\text{SGM}}(\alpha; q, \sigma)$$

denote the RDP cost at order $\alpha > 1$ of one subsampled Gaussian DP-SGD step with sampling rate q and noise multiplier σ . In implementation, this value is obtained from the Opacus RDP accountant.

For client k , the accumulated RDP cost over the entire continual stream is

$$\rho_k(\alpha) = \sum_{t=1}^T \sum_{r=1}^{R_t} z_{k,t,r} \sum_{s=1}^{S_{k,t,r}} \rho_{\text{SGM}}(\alpha; q_{k,t,r,s}, \sigma).$$

The corresponding (ε, δ) value is

$$\varepsilon_k(\delta) = \inf_{\alpha > 1} \left[\rho_k(\alpha) + \frac{\log(1/\delta)}{\alpha - 1} \right].$$

The implementation calibrates one global noise multiplier σ so that

$$\varepsilon_{\max}(\delta) = \max_{k \in [K]} \varepsilon_k(\delta) \leq \varepsilon_{\text{target}}$$

when the accountant search is not capped. This max-client value is the privacy guarantee used for the main claims.

Theorem B.1 (Record-level LDP of PRISM). *Assume that (i) the public initialization checkpoint is trained only on public data; (ii) the client participation schedule is independent of private data; (iii) each selected client runs DP-SGD with per-example clipping norm C , noise multiplier σ , and the sampling rates used by the RDP accountant; (iv) aggregation weights are public, fixed by the protocol, or depend only on public protocol quantities; and (v) server-side aggregation, SVD reparameterization, memory writing, rank scheduling, and noise-floor shrinkage use only privatized client updates, public hyperparameters, and public randomness. Then the released transcript of PRISM, including all intermediate global models, memory*

states, plastic states, and the final model, satisfies per-client record-level LDP. More precisely, for any client k , neighboring datasets $D_k \sim D'_k$, and measurable output event $\mathcal{O}$,

$$\begin{aligned} & \Pr[\mathcal{M}(D_1, \dots, D_k, \dots, D_K) \in \mathcal{O}] \\ & \leq e^{\varepsilon_k} \Pr[\mathcal{M}(D_1, \dots, D'_k, \dots, D_K) \in \mathcal{O}] + \delta, \end{aligned} \quad (7)$$

where

$$\varepsilon_k = \inf_{\alpha > 1} \left[\rho_k(\alpha) + \frac{\log(1/\delta)}{\alpha - 1} \right].$$

Consequently, if σ is calibrated so that $\max_k \varepsilon_k \leq \varepsilon_{\text{target}}$, then PRISM satisfies $(\varepsilon_{\text{target}}, \delta)$ record-level LDP for all clients.

Proof. Fix a client k . Consider two global private datasets that differ in exactly one record of client k ; datasets of all other clients are identical. When client k is not selected in a communication round, the mechanism does not access D_k ; that round therefore incurs no privacy loss for client k . When client k is selected, its local update is produced by DP-SGD. Each DP-SGD step is a subsampled Gaussian mechanism with RDP cost $\rho_{\text{SGM}}(\alpha; q_{k,t,r,s}, \sigma)$.

By adaptive sequential composition of RDP, all DP-SGD steps performed by client k over all selected rounds and all tasks satisfy $(\alpha, \rho_k(\alpha))$ -RDP, where $\rho_k(\alpha)$ is the stream-wide sum above. Adaptive composition applies because the initial model broadcast to a client in a later round may depend on previous privatized updates, but the DP guarantee of the next local randomizer remains valid when conditioned on prior DP outputs. Converting RDP to approximate DP gives $\varepsilon_k(\delta)$.

The server receives only privatized client messages. Weighted aggregation, SVD reparameterization, memory writing, dynamic rank scheduling, noise-floor shrinkage, and model release are deterministic or randomized functions of those privatized messages, public hyperparameters, and public randomness. By the post-processing property of differential privacy, these operations do not increase privacy loss. Thus the entire transcript and final released model satisfy (ε_k, δ) record-level LDP for client k . Taking the maximum over clients gives the uniform guarantee $(\varepsilon_{\max}, \delta)$ with $\varepsilon_{\max} = \max_k \varepsilon_k$. $\square$

B.2 Why Server-Side Updates Are Post-Processing

At each stream boundary, PRISM applies the following operations for each LoRA layer:

$$\Delta W_l = B_l^M A_l^M + B_l^P A_l^P, \quad (8)$$

$$P_M^\perp = I - (A_l^M)^\top A_l^M, \quad (9)$$

$$R_l = B_l^P A_l^P P_M^\perp, \quad (10)$$

$$R_l = U \Sigma V^\top. \quad (11)$$

It then updates memory and plastic subspaces, applies rank scheduling, and optionally applies noise-floor shrinkage

$$\tilde{s}_j = \sqrt{\max(s_j^2 - \gamma \tau_l^2, 0)}.$$

These steps do not access raw private records. They use only privatized client updates produced by DP-SGD, public hyperparameters such as r , $r_{\min}$, γ , and the noise-floor quantile, and public randomness or synthetic noise simulation used to estimate τ_l . Therefore, SVD, memory writing, rank scheduling, and shrinkage may affect utility, but they do not change the formal privacy budget.

B.3 Practical Accountant and the σ -Cap

The experiments generate the full client participation schedule before private training. The schedule is independent of private data. For every client, the complete privacy history is computed over the full stream; privacy is not reset at segment boundaries. RDP is then converted to (ε, δ) -DP, and the following quantities are reported:

$$\varepsilon_{\max} = \max_k \varepsilon_k, \quad \varepsilon_{\text{mean}} = \frac{1}{K} \sum_{k=1}^K \varepsilon_k.$$

Only $\varepsilon_{\max}$ is the privacy guarantee. The mean is reported as a diagnostic showing how the sampled participation schedule distributes privacy loss across clients.

The accountant calibrates σ through a bisection search over the noise multiplier needed to achieve the target $\varepsilon_{\max}$. To prevent runaway noise levels, the search is capped at $\sigma = 64$. For the Sentiment3 stress row, the bisection saturates at this cap and the achieved max-client ε is approximately 1.07 rather than the nominal unit target. This condition is therefore referred to as σ -capped, achieved $\varepsilon_{\max} \approx 1.07$, and the achieved ε is reported explicitly.

B.4 Scope of the Guarantee

The theorem establishes per-client record-level LDP. It protects the contribution of any single training record inside a client. It does not provide client-level DP for the addition or removal of an entire client. The guarantee covers all privatized local updates released by clients, all aggregated server updates, all SVD-reparameterized LoRA factors, all memory and plastic subspace states, and all final released model parameters.

The guarantee does not cover non-DP evaluation on private validation or test records. In the experiments, evaluation uses public benchmark splits. If evaluation were performed on private client-held validation records and released publicly, that evaluation would require separate DP accounting.

C Reproducibility Details: Data, Metrics, Software, and Seeds

Data provenance and splits. This appendix records the dataset and implementation details underlying Section 4. All datasets are public English text-classification benchmarks with human-provided labels. They are loaded through the Hugging Face datasets interface and cached locally. No new text is collected, no raw examples are released, and no additional personal-information or offensive-content filtering is applied beyond standard benchmark preprocessing. Table 8 reports the source, split construction, label space, and sample caps used in the pipeline.

SNLI is distributed under CC BY-SA 4.0. MNLI inherits mixed licensing from its source materials; the dataset card lists several license tags and describes most of the corpus as OANC-derived, with fiction examples under permissive or Creative Commons licenses. GLUE/SST-2 and IMDB are listed with license field other on their dataset cards. For Yelp Polarity and Rotten Tomatoes, the provider terms exposed through the dataset loaders are followed.

Public initialization. Private FL starts from a public task-aligned classifier head rather than a random head. For NLI, RoBERTa-large is used with public SNLI warmup, training the classifier head and public LoRA B factors while keeping the backbone frozen. The public LoRA update is merged into the backbone before private FL. The warmup uses 50,000 train examples, 2,000 eval examples, 3 epochs, learning rate 5×10^{-4} , weight decay 0.01, LoRA rank 8, alpha 8, and dropout 0.05. For

Role		Dataset	Source name	Labels	Split construction and examples used
Public NLI init.		SNLI	snli	3-way NLI	Public train/validation; 50,000 train and 2,000 evaluation examples.
Private NLI stream		MNLI genres	multi_nli	3-way NLI	Per-genre 90/10 split from the train split; capped at 2,000 train and 1,000 evaluation examples per genre.
Public sentiment init.		Rotten Tomatoes	rotten_tomatoes	Binary	Public train/validation; full train split and 1,000 evaluation examples.
Private stream	sentiment	SST-2	glue/sst2	Binary	GLUE train/validation; 2,000 train and 1,000 evaluation examples.
Private stream	sentiment	IMDb	imdb	Binary	Train/test split; 2,000 train and 1,000 evaluation examples.
Private stream	sentiment	Yelp Polarity	yelp_polarity	Binary	Train/test split; 2,000 train and 1,000 evaluation examples.

Table 8: **Dataset sources, label spaces, split construction, and sample caps used for all experiments.** All tasks are English classification benchmarks loaded through Hugging Face datasets. When the original split is larger than the cap, the subset is sampled deterministically from the run seed.

Sentiment3, the Rotten-Tomatoes warmup follows the same pattern with binary labels and a 1,000-example public evaluation subset. Public warmup uses no private examples and is not counted in the DP budget.

Metric definitions and diagnostics. In addition to final average accuracy and average forgetting in Section 4, the logs include backward transfer $\frac{1}{M-1} \sum_{j < M} (Acc_{M,j} - Acc_{j,j})$, last-task accuracy $Acc_{M,M}$, and gain over Public-Init $\frac{1}{M} \sum_j (Acc_{M,j} - Acc_{0,j})$. Accuracy is exact-match classification accuracy under the dataset label mapping. The logs also include prediction distributions at every segment boundary, communication cost in uploaded and downloaded parameter counts, run-time per round, rank evolution, singular values, estimated noise floor, SNR, kept rank, discarded energy, and reconstruction error. The tasks are standard classification tasks; the experiments do not use prompt templates, decoding strategies, retrieval, reranking, or LLM-as-judge evaluation.

Implementation details. Private training uses maximum sequence length 128, SGD, no weight decay, 50 rounds per task, 5 local steps, logical batch 64, and main seeds 42, 43, and 44. DP-FedAvg uses physical batch 16 with 4-step accumulation to match the logical batch. PRISM uses $r_P^{\min} = 2$, no memory write after the final task, noise-floor quantile 0.95, 32 Monte Carlo samples for noise-floor estimation and the γ defaults in Section 3.5. Text is tokenized with the RoBERTa to-

kenizer, truncated or padded to the maximum sequence length, and evaluated with the dataset label mappings described in the main text. No early stopping or test-set-based checkpoint selection is used. All reported results are obtained from the final model after the full task stream.

The implementation uses PyTorch 2.8.0 with CUDA 12.8, Transformers 5.7.0, Datasets 4.8.5, and Opacus 1.5.4. Runs were scheduled on NVIDIA RTX 4080-class GPUs. Every run stores its full configuration, client schedule, partition sizes, privacy accounting report, segment-level accuracy matrix, communication logs, prediction distributions, and method-specific SVD/rank logs.

Statistical protocol. The main MNLI tables report mean and standard deviation over three random seeds for both the forward and reverse task orders, giving six seed-order runs per privacy budget. Sentiment3 reports mean and standard deviation over three seeds in the forward order. Methods are paired by seed whenever possible, because a fixed seed determines the data split, client partition, client schedule, public initialization, and calibrated noise multiplier. Targeted diagnostics such as the γ sweep are reported separately from the main multi-seed comparisons. The γ sweep uses seed 45 and fixes a regime-level default: $\gamma = 0.25$ for the less restrictive $\varepsilon = 6$ setting, where all values are nearly tied, and $\gamma = 1.0$ for $\varepsilon \leq 3$, where stronger shrinkage is clearly favored.

D Public Warmup Sensitivity

The main MNLI experiments use a strong SNLI public warmup. This choice matches the target deployment assumption: the model already has a publicly trained task interface, and private FL should adapt it to sequential private domains without degrading earlier performance. To separate retention of public initialization from learning from the private stream, Table 9 reports a three-seed diagnostic that varies the amount of public SNLI warmup before running PRISM at $\varepsilon = 2$. The diagnostic shows that private adaptation is sensitive to public-checkpoint quality. It yields a substantial gain for the SNLI-10k-e1 checkpoint, no reliable gain for the weakest SNLI-5k-e1 checkpoint, and modest positive gains once the public checkpoint is near saturation. The main SNLI-50k-e3 setting is therefore a conservative retention-oriented baseline rather than a configuration chosen to maximize private-stage gains.

Public warmup	Public-Init	PRISM	Gain
SNLI-5k-e1	41.25 $\pm$ 0.40	41.19 $\pm$ 1.61	-0.07 $\pm$ 1.49
SNLI-10k-e1	55.97 $\pm$ 0.16	65.99 $\pm$ 0.45	+9.85 $\pm$ 0.86
SNLI-20k-e1	76.79 $\pm$ 0.18	77.21 $\pm$ 0.50	+0.42 $\pm$ 0.64
SNLI-30k-e1	75.67 $\pm$ 0.25	76.75 $\pm$ 1.36	+1.07 $\pm$ 1.10
SNLI-40k-e1	77.83 $\pm$ 0.32	78.37 $\pm$ 1.03	+0.54 $\pm$ 0.96
SNLI-50k-e3	79.34 $\pm$ 0.46	80.08 $\pm$ 0.65	+0.74 $\pm$ 0.29

Table 9: **MNLI public-warmup sensitivity.** Three-seed diagnostic at $\varepsilon = 2$. The table separates private-stream learning from retention of a strong public initialization; gain is computed per seed as PRISM minus Public-Init.

E Baseline Configuration and Fairness Controls

All compared private FL methods share the same RoBERTa-large backbone, frozen private-stage backbone, frozen pretrained classifier head, public checkpoint, private task stream, client partition, client participation schedule, number of clients, participation rate, rounds per task, local steps, logical batch size, optimizer, learning rate, clipping norm, privacy accountant, calibrated noise multiplier, train/evaluation caps, and final-model evaluation protocol within each dataset, budget, seed, and stream order. Thus differences among methods reflect the private adaptation rule rather than a different privacy budget, data split, public initialization, or client sampling pattern.

The hyperparameter protocol is conservative. The main private optimizer is SGD with learning

rate 5×10^{-3} , fixed before the main comparison. All LoRA methods use total rank 8 unless a static rank sweep is explicitly reported. Rank-scheduled FedSVD and PRISM use the same deterministic plastic-rank schedule and the same minimum plastic rank. FedSVD-style methods use the same server-side SVD period. PRISM uses a fixed SNR threshold and a fixed synthetic noise-floor estimator for a given privacy regime. The O-LoRA-style baseline is not claimed to be an exact reproduction of the original non-private, non-federated algorithm; it is an orthogonality-only control implemented under the same frozen backbone, public initialization, LoRA rank, local DP-SGD accountant, optimizer, and client schedules as the other methods.

F Extended Related Work

Private and federated parameter-efficient adaptation. DP-SGD (Abadi et al., 2016) clips per-example gradients and adds Gaussian noise, and Rényi DP accounting (Mironov, 2017) tracks the composed privacy loss. In federated learning, FedAvg (McMahan et al., 2017a) and private FL variants (McMahan et al., 2017b; Kairouz et al., 2021) provide the standard optimization template, but full updates are vulnerable to the high-dimensional noise burden. Parameter-efficient private fine-tuning reduces that burden: Yu et al. (2022) adapt parameter-efficient methods to DP language-model fine-tuning, Bu et al. (2022) train only bias terms, and federated DP-LoRA methods perturb low-rank updates rather than full models (Liu et al., 2025; Xu et al., 2026, 2025). These studies establish the importance of reducing trainable dimensionality under DP. They do not address sequential domain streams in which a fixed low-rank capacity is reused under a single privacy accountant.

Closest federated LoRA methods. FLoRA (Wang et al., 2024) and FedEx-LoRA (Singhal et al., 2025) show that direct averaging of LoRA factors is often an inadequate federated aggregation rule. FLoRA handles heterogeneous ranks by stacking adapters, while FedEx-LoRA makes aggregation exact by storing a residual update in the frozen backbone. These methods target non-private or aggregation-accuracy settings rather than LDP continual learning. FedSVD (Lee et al., 2025) is closer because it is explicitly private: clients train only B , and the server uses SVD to update an orthonormal

Method	Trainable private parameters	Capacity/control role	Continual mechanism
Public-Init	None	Public checkpoint only	None
DP-FedAvg	LoRA A and B	Total rank 8; largest DP-noise exposure	None
DP-FFA	LoRA B only	Total rank 8 with fixed A	None
DP-FedSVD	LoRA B only	Total rank 8; static FedSVD reference	None
DP-FedSVD ($r = 2/4/6/8$)	LoRA B only	Static rank sweep for active-rank sensitivity	None
DP-FedSVD+RS	Plastic B^P only	Same dynamic plastic-rank schedule as PRISM	Rank throttling only
PRISM variants	Plastic B^P only	Total rank 8 with scheduled plastic rank	Optional memory and shrinkage gates
O-LoRA-style	LoRA factors with orthogonality penalty	Total rank 8 orthogonal-subspace control	Orthogonality regularization

Table 10: **Method-specific differences after shared controls are fixed.** The static rank sweep and rank-scheduled FedSVD are included to test whether PRISM gains are explained by lower active trainable rank.

A , reducing LoRA noise amplification under DP-SGD. PRISM uses FedSVD as the static re-parameterization baseline but adds plastic-rank scheduling and a shrinkage threshold before any memory write in continual streams. PrE-Text (Hou et al., 2024) is another recent private federated language-model route, but it replaces on-device training with DP synthetic data. The experiments retain the client-side DP-SGD training path.

Continual subspaces. Regularization and replay methods such as EWC (Kirkpatrick et al., 2017) and GEM/A-GEM (Lopez-Paz and Ranzato, 2017; Chaudhry et al., 2018) mitigate forgetting but either constrain gradients indirectly or require stored examples. Subspace methods such as OGD (Farajtabar et al., 2020) and GPM (Saha et al., 2021) explicitly protect old-task directions. Recent parameter-efficient continual methods reduce this cost by using pretrained models: L2P (Wang et al., 2022) and CODA-Prompt (Smith et al., 2023) learn prompt memories, while O-LoRA (Wang et al., 2023) separates language tasks with orthogonal LoRA subspaces. PRISM is closest in spirit to O-LoRA, but differs in three operational details: subspaces are written from DP-aggregated updates, the total rank does not grow with the number of tasks, and low-SNR directions are discarded rather than preserved as new task-specific capacity.

Federated continual learning. FedWeIT (Yoon et al., 2021), federated class-incremental learning (Dong et al., 2022), and FedET (Liu et al., 2023) model forgetting across clients and tasks through decomposed parameters, distillation, or compact task modules. Their problem definitions are pri-

marily class-incremental and do not provide the record-level LDP accounting used here. Replay or distillation can also require raw examples or private logits. PRISM therefore addresses a narrower but privacy-compatible setting: fixed-task domain-incremental streams with a public initialization and only post-processed private aggregates at domain boundaries.

G SST-2/Rotten Tomatoes Overlap Check

Normalized-text matching finds that 44.6% of the Rotten Tomatoes warmup examples occur in the SST-2 training split and 8.0% occur in the validation split, for 52.5% in total (4,481/8,530). The corresponding overlap is zero for IMDb and at most three examples for Yelp. The same warmup checkpoint is shared by all compared methods.

We remove the overlapping examples, retrain the public initialization, and rerun Sentiment3 at $\varepsilon = 2$ over three seeds under the same protocol, with σ ranging from 34.1 to 39.2. Table 11 reports the resulting final average accuracy and forgetting.

Method	Final Avg. Acc. (%)	Fgt. (%)
Public-Init	92.3 $\pm$ 0.3	–
DP-FedSVD (full rank)	78.2 $\pm$ 6.0	12.1 $\pm$ 2.7
O-LoRA-style	59.3 $\pm$ 16.5 [†]	–
DP-FedSVD+RS	90.6 $\pm$ 1.0	1.3 $\pm$ 0.6
PRISM	91.8 $\pm$ 0.3	0.4 $\pm$ 0.3

Table 11: Sentiment3 results at $\varepsilon = 2$ using the decontaminated Rotten Tomatoes warmup. Values are mean $\pm$ standard deviation over three seeds. [†]Two of three O-LoRA-style runs collapse; forgetting is therefore not reported.

H Expanded Accuracy, Forgetting, and Public-Init Gains

Table 12 expands Table 2 to include standard deviations for final average accuracy and forgetting (Fgt), and the per-cell gain over Public-Init, Δ_{pub} . Two patterns are salient. First, the variance of PRISM is consistently the smallest across all cells: at MNLI $\varepsilon = 2$ its standard deviation is 0.76%, an order of magnitude smaller than the 17.78% standard deviation of DP-FedSVD; on Sentiment3 the same ordering holds at every budget. The O-LoRA-style control is competitive under less restrictive budgets but adds a high-variance failure mode under stricter privacy, especially on MNLI $\varepsilon = 2$ and Sentiment3 $\varepsilon = 2$. Second, the forgetting of PRISM is near zero or slightly negative on Sentiment3, where negative forgetting indicates positive backward transfer on average; accuracy on earlier tasks improves rather than degrades after boundary processing.

I MNLI Order Robustness

The compact MNLI numbers in Table 2 merge the two stream orders. Since stream order is a deterministic experimental condition, not a sampling source, it is reported separately here. The qualitative ordering of methods is preserved in both orders, but the high-variance baselines (DP-FFA, DP-FedSVD, and O-LoRA under strict privacy) are also unstable in both orders, with reverse generally amplifying their variance, consistent with the reverse stream front-loading the harder-to-learn genres. PRISM is order-stable: the forward/reverse gap stays below 1% at every budget.

J MNLI Class-Collapse Diagnostic Across Orders

The retention matrix in Section 5 shows that failures under strict DP are non-uniform: some tasks degrade more than others. The class-collapse diagnostic provides a complementary view by asking whether the failure is uniform across classes or biased toward one. The final maximum-class prediction share is computed as the fraction of evaluation predictions that fall into the single most-frequent predicted class, averaged over forward and reverse runs and across genres.

Figure 4 shows the result. At $\varepsilon = 3$, DP-FedAvg already exhibits a clear single-class bias (0.79); DP-FFA and DP-FedSVD are mildly biased (~ 0.41) and PRISM stays near 0.38, slightly above the balanced value of $1/3$. At $\varepsilon = 2$ the concentration

Method	Final Avg. Acc. (%)	Fgt (%)	Δ_{pub} (%)
<i>MNLI, $\varepsilon = 6$</i>			
Public-Init	79.35 ± 0.68	0.00 ± 0.00	0.00
DP-FedAvg	78.47 ± 1.75	1.95 ± 1.46	-0.87
DP-FFA	80.77 ± 0.49	0.53 ± 0.32	+1.42
DP-FedSVD	81.04 ± 0.55	0.23 ± 0.41	+1.69
O-LoRA	80.62 ± 0.70	0.10 ± 0.46	+1.27
PRISM (ours)	81.23 ± 0.65	0.30 ± 0.37	+1.89
<i>MNLI, $\varepsilon = 3$</i>			
DP-FedAvg	42.96 ± 14.29	18.42 ± 11.5	-36.39
DP-FFA	73.50 ± 6.53	4.58 ± 3.32	-5.85
DP-FedSVD	73.88 ± 4.90	4.18 ± 2.74	-5.47
O-LoRA	77.78 ± 1.78	0.61 ± 1.04	-1.57
PRISM (ours)	80.31 ± 0.40	0.56 ± 0.44	+0.97
<i>MNLI, $\varepsilon = 2$</i>			
DP-FedAvg	33.37 ± 0.85	9.12 ± 8.41	-45.97
DP-FFA	51.07 ± 18.29	17.91 ± 9.30	-28.28
DP-FedSVD	52.32 ± 17.78	15.08 ± 10.7	-27.03
O-LoRA	56.21 ± 21.26	9.87 ± 9.68	-23.14
PRISM (ours)	79.75 ± 0.76	0.44 ± 0.51	+0.41
<i>Sentiment3, $\varepsilon = 3$</i>			
DP-FedAvg	49.81 ± 0.20	17.72 ± 14.19	-39.11
DP-FFA	90.31 ± 1.06	2.20 ± 0.91	+1.40
DP-FedSVD	90.02 ± 0.71	2.51 ± 0.70	+1.11
O-LoRA	90.94 ± 1.05	-0.78 ± 2.72	+2.03
PRISM (ours)	92.19 ± 0.10	0.31 ± 0.14	+3.28
<i>Sentiment3, $\varepsilon = 2$</i>			
DP-FedAvg	49.90 ± 0.35	5.79 ± 6.63	-39.02
DP-FFA	82.01 ± 2.33	9.76 ± 2.49	-6.91
DP-FedSVD	84.97 ± 0.28	5.75 ± 1.83	-3.94
O-LoRA	62.22 ± 21.85	10.38 ± 16.19	-26.70
PRISM (ours)	92.33 ± 0.43	-0.12 ± 0.56	+3.41
<i>Sentiment3, σ-capped stress, achieved $\varepsilon_{\text{max}} \approx 1.07$</i>			
DP-FedAvg	49.90 ± 0.35	0.02 ± 0.03	-39.02
DP-FFA	50.45 ± 0.79	2.88 ± 3.98	-38.47
DP-FedSVD	50.75 ± 1.83	1.31 ± 0.68	-38.16
O-LoRA	50.10 ± 0.35	0.06 ± 0.97	-38.81
PRISM (ours)	90.79 ± 0.60	0.86 ± 0.95	+1.87

Table 12: **Expanded main results for MNLI genre and Sentiment3.** Final average accuracy is reported in percent; Fgt is average forgetting in percent; Δ_{pub} is gain over the Public-Init model. Entries for final average accuracy and Fgt are mean $\pm$ standard deviation over seeds and, for MNLI, both task orders. The Sentiment3 stress block is σ -capped with achieved $\varepsilon_{\text{max}} \approx 1.07$.

Method	Forward Avg. Acc.	Reverse Avg. Acc.
$\varepsilon = 6$		
DP-FedAvg	78.76 $\pm$ 2.63	78.19 $\pm$ 0.69
DP-FFA	80.97 $\pm$ 0.45	80.57 $\pm$ 0.52
DP-FedSVD	81.35 $\pm$ 0.67	80.73 $\pm$ 0.14
O-LoRA	80.85 $\pm$ 0.67	80.39 $\pm$ 0.79
PRISM (ours)	81.43 $\pm$ 0.59	81.03 $\pm$ 0.76
$\varepsilon = 3$		
DP-FedAvg	44.29 $\pm$ 16.92	41.63 $\pm$ 14.79
DP-FFA	75.88 $\pm$ 2.70	71.12 $\pm$ 9.08
DP-FedSVD	75.50 $\pm$ 2.74	72.26 $\pm$ 6.68
O-LoRA	78.40 $\pm$ 2.15	77.15 $\pm$ 1.47
PRISM (ours)	80.53 $\pm$ 0.35	80.10 $\pm$ 0.39
$\varepsilon = 2$		
DP-FedAvg	32.80 $\pm$ 0.61	33.95 $\pm$ 0.68
DP-FFA	56.02 $\pm$ 17.11	46.11 $\pm$ 21.68
DP-FedSVD	58.84 $\pm$ 15.09	45.80 $\pm$ 20.87
O-LoRA	61.74 $\pm$ 23.39	50.68 $\pm$ 22.16
PRISM (ours)	80.08 $\pm$ 0.65	79.43 $\pm$ 0.83

Table 13: **MNLI genre order robustness.** Final average accuracy (%) for forward and reverse MNLI task orders, reported as mean $\pm$ standard deviation over three seeds per order. Each method uses the same public initialization, client schedule, and privacy budget within a seed/order cell.

becomes more pronounced: the max-class share of DP-FedAvg rises to 0.93, the share of DP-FedSVD rises to 0.64, and the share of DP-FFA rises to 0.74, while PRISM stays at 0.38. Two implications follow. First, the aggregate accuracy drop of the DP baselines reflects a pathology, single-class collapse, rather than only a uniform precision loss. Second, the methods that suppress this pathology are the same methods that restore retention in Section 5: rank control plus shrinkage prevents biased noisy directions from dominating adapter predictions.

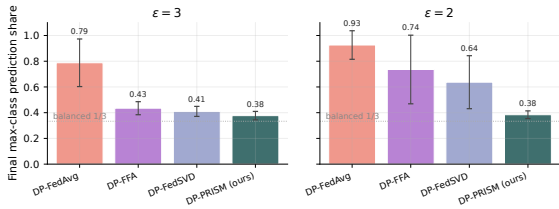


Figure 4: **MNLI class-collapse diagnostic across forward and reverse orders.** The metric is final maximum-class prediction share: the fraction of evaluation predictions assigned to the most frequent predicted label. MNLI is three-way, so the balanced reference is $1/3$; larger values indicate class-biased collapse.

Budget	Method	Seeds (3) min-max	Mean
3	PRISM	92.07–92.26	92.19
3	O-LoRA	89.84–91.93	90.94
3	DP-FedSVD	89.46–90.81	90.02
3	DP-FFA	89.08–90.94	90.31
3	DP-FedAvg	49.69–50.04	49.81
2	PRISM	91.83–92.58	92.33
2	O-LoRA	49.50–87.45	62.22
2	DP-FedSVD	84.76–85.29	84.97
2	DP-FFA	79.97–84.54	82.01
2	DP-FedAvg	49.69–50.31	49.90
Stress [†]	PRISM	90.37–91.47	90.79
Stress [†]	O-LoRA	49.69–50.31	50.10
Stress [†]	DP-FedSVD	49.69–52.87	50.75
Stress [†]	DP-FFA	49.73–51.30	50.45
Stress [†]	DP-FedAvg	49.69–50.31	49.90

Table 14: **Sentiment3 final average accuracy ranges across seeds.** Values are min-max and mean over three seeds for the SST-2 $\rightarrow$ IMDb $\rightarrow$ Yelp Polarity stream. [†]The stress condition is σ -capped with achieved $\varepsilon_{\max} \approx 1.07$ (Table 1).

K Sentiment3 Forgetting Trace and Per-Seed Accuracy Range

Sentiment3 is a binary stream, so single-class collapse manifests as final average accuracy $\approx 50\%$; forgetting metrics therefore require careful interpretation under collapse. Table 14 reports the per-seed range of final average accuracy: it shows that the within-cell variance for PRISM is low ($\leq 1.1\%$ at every budget), while every other DP method has either tight-but-collapsed variance (50 ± 1) or moderate variance around a non-collapsed mean (e.g. DP-FFA at $\varepsilon = 2$). O-LoRA provides a useful contrast: it is stable at $\varepsilon = 3$ but has a wide range at $\varepsilon = 2$ because only one seed avoids collapse.

The forgetting curve in Figure 5 shows why the low DP-baseline forgetting at the σ -cap does not indicate stability. The DP baselines collapse to single-class prediction *at the very first task* under the cap; after collapse occurs at the first task, subsequent evaluations remain near chance-level, so the forgetting metric approaches zero by construction. Only PRISM maintains both accuracy and controlled forgetting at the cap.

L Hyperparameter and Stress Diagnostics: γ , MNLI Strict-Budget Stress, and Boundary SVD Statistics

This appendix reports three auxiliary diagnostics referenced from Sections 3, 4, and 6: the γ sweep, the MNLI strict-budget stress test, and boundary

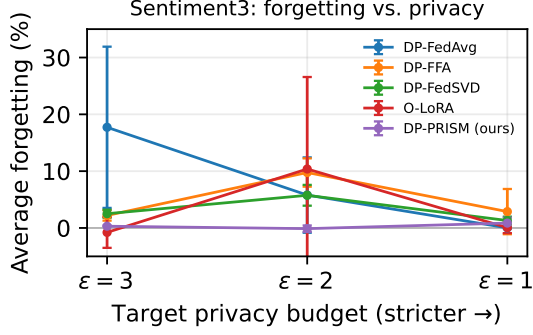


Figure 5: **Sentiment3 average forgetting across privacy budgets.** Lower is better, but near-zero forgetting for collapsed DP baselines at the σ -cap does not indicate stability; their final accuracy is near chance-level in Table 14. O-LoRA is competitive at $\varepsilon = 3$ but unstable at stricter budgets.

ε	$\gamma = 0$	$\gamma = 0.25$	$\gamma = 0.5$	$\gamma = 1.0$	Best
2	73.58	75.42	78.56	79.74	1.0
3	79.20	78.38	78.56	79.98	1.0
6	80.52	80.52	80.46	80.46	0 / 0.25

Table 15: **Single-seed shrinkage sweep on MNLI seed 45.** Entries are final average accuracy. The less restrictive budget setting ($\varepsilon = 6$) is insensitive to γ , while stricter privacy budgets favor $\gamma = 1.0$.

SVD statistics.

Shrinkage coefficient γ . Table 15 sweeps the shrinkage coefficient on seed 45. At $\varepsilon = 6$, all settings are effectively tied (80.46–80.52), so $\gamma = 0.25$ is used as a mild reference value. At $\varepsilon \in \{3, 2\}$, $\gamma = 1.0$ gives the best final accuracy. We therefore use $\gamma_{\varepsilon=6} = 0.25$ and $\gamma_{\varepsilon \leq 3} = 1.0$. This is a single-seed calibration of a coarse regime-level rule, not a per-run hyperparameter search.

MNLI strict-budget stress. Figure 6 reports the forward-order stress test under a near-unit privacy target. All DP baselines collapse to three-way chance accuracy, while PRISM reaches 77.4 ± 2.5 . This extreme-budget diagnostic identifies a shared baseline failure mode rather than establishing general near-unit-budget reliability. The achieved max-client ε remains near the target and reaches approximately 1.06 in the capped seed.

Boundary SVD statistics. Figure 7 reports the boundary quantities. The noise-floor scale τ_l rises as ε tightens, and reconstruction error follows the same trend while remaining below τ_l . The discarded update energy rises from 11% at $\varepsilon = 6$ to

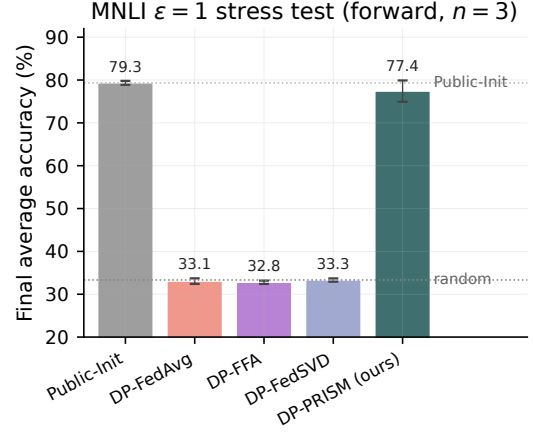


Figure 6: **MNLI strict-budget stress.** Final average accuracy over three seeds under a near-unit privacy target.

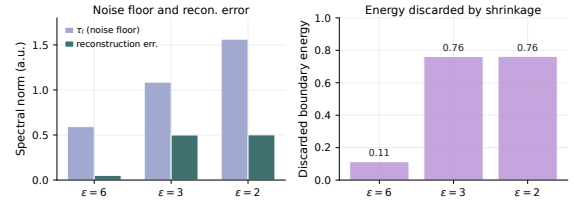


Figure 7: **Boundary SVD statistics on MNLI.** Noise floor, reconstruction error, and discarded boundary-update energy.

approximately 76% at $\varepsilon \in \{3, 2\}$. This is consistent with low-SNR removal, although discarded energy cannot be uniquely attributed to noise without ground-truth signal directions.