NIFT: NOISE-INVARIANT FINE-TUNING FOR PRIVACY-PRESERVING LLMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Privacy-preserving fine-tuning of large language models is appealing for many important applications, but remains challenging in practice. Adding differentially private (DP) noise to input representations can protect private data, but this noise is amplified by deep Transformer layers and sharply degrades performance. Prior work has primarily addressed this problem through noise-correction mechanisms that attempt to recover useful representations after amplification has already occurred, but these approaches still leave a substantial performance gap compared with the corresponding no-noise setting. In this paper, we propose a two-stage training framework that instead prepares the backbone to be less sensitive to the agreed DP noise distribution before downstream task fine-tuning. First, we train a LoRA adapter with a label-free pairwise consistency objective over multiple noisy embedding copies, encouraging the backbone's hidden states to remain stable under DP perturbations. Second, we fine-tune a lightweight client-side ladder side network that fuses local features with selected hidden states returned by the robust server-side backbone. This separation of noise-invariance preparation from task adaptation reduces the burden on the client model and improves robustness under privacy-preserving noise. Our experiments show that NIFT outperforms the state-of-the-art HiddenEcho across multiple datasets and privacy budgets.

1 INTRODUCTION

Privacy-preserving fine-tuning of large language models (LLMs) has become increasingly important for NLP systems that must adapt to user- or organization-specific data. The most straightforward way to achieve this goal is to fine-tune the model entirely on local hardware, so that sensitive user or organizational data never leaves the private environment. However, fully fine-tuning a multi-billion-parameter model locally is often impractical because private hardware may have limited memory, computation, and deployment capacity (Lin et al., 2022; Dettmers et al., 2023).

This resource constraint motivates split learning (Vepakomma et al., 2018; Thapa et al., 2022), which partitions the model at a chosen cut layer and distributes computation between local hardware and a cloud provider while keeping raw private data within the local environment (Ma et al., 2025; Li et al., 2025; Shi et al., 2022). This design reduces direct exposure of user data, but it does not eliminate privacy risk; instead, it shifts the attack surface from raw inputs to intermediate representations. A curious server may still recover sensitive inputs through embedding inversion or gradient leakage attacks (Zhu et al., 2019; Erdogan et al., 2022; Jin et al., 2021). A common defense is to add calibrated noise to the embeddings before they leave the private environment (Dwork & Roth, 2014; Qu et al., 2021; Shen et al., 2023). However, this noise does not remain benign as it propagates through the backbone: each Transformer block transforms and re-normalizes the perturbed embedding, causing deviations in deeper hidden states to grow substantially beyond the initial perturbation. As a result, the downstream classifier receives noisy, distorted representations, leading to sharp accuracy degradation under practical privacy budgets.

Prior work has addressed this accuracy degradation by treating error amplification as a downstream artifact to be repaired. Split-and-Denoise (Mai et al., 2024) introduced a client-side denoiser for noisy embeddings, while HiddenEcho (Li et al., 2026) extended this repair-based strategy to per-layer corrections and currently represents the strongest approach in this direction. Although these methods can partially recover accuracy, their corrective effect is inherently limited: once the server-side

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

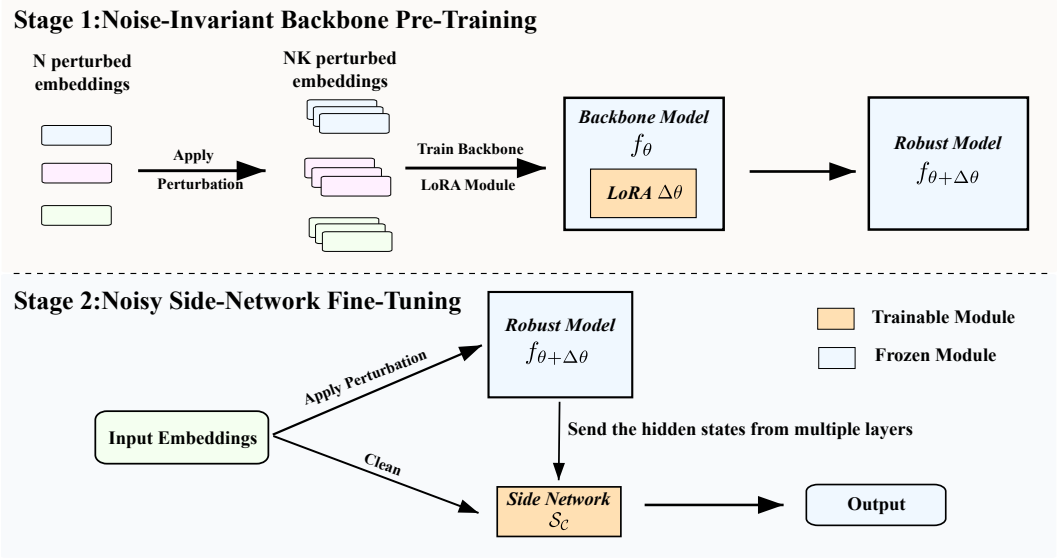

Figure 1: Overview of the proposed two-stage framework. *Stage 1 (top): noise-invariant backbone pre-training.* The client sends N DP-perturbed embeddings to the server; the server then perturbs each received embedding K times, yielding NK noisy copies in total. These copies are passed through the frozen backbone with a low-rank (LoRA) adapter and used to train the robust model $f_{\theta+\Delta\theta}$. *Stage 2 (bottom): noisy side-network fine-tuning.* The client embeds each input, sends a DP-perturbed copy through the frozen robust backbone on the server (which returns selected hidden states to the device), and trains a lightweight client-side side network that fuses these returned hidden states with a clean local embedding. Different clients can each train their own LoRA adapter on top of the same shared backbone.

backbone has amplified a small privacy-preserving perturbation into a much larger representation shift, the client may not be able to reconstruct the lost clean representation, and the resulting model often remains substantially less accurate than the same model evaluated on unperturbed inputs.

To address this limitation, we take the opposite view: **although the backbone remains frozen during each user's task-specific fine-tuning, it can be prepared in advance to make its hidden states less sensitive to the DP noise injected by the client.** We propose NIFT (Noise-Invariant Fine-Tuning), a two-stage framework for privacy-preserving split fine-tuning of LLMs in local–server settings. In the first stage, we fit a low-rank adapter (LoRA) (Hu et al., 2022) to the backbone using a label-free pairwise consistency objective. For each input, the client sends a single DP-perturbed embedding to the server; the server then creates K noisy copies internally, passes them through the backbone, and minimizes the mean pairwise ℓ_2 distance between their hidden states at selected layers. This objective encourages the LoRA parameters to absorb noise-induced variability, so that the resulting backbone-adapter model produces hidden states that are approximately invariant to the agreed-upon noise distribution. Because the consistency objective does not require labels, each client can train its own LoRA adapter on top of the same shared frozen backbone, matched to its local input distribution and privacy budget. In the second stage, we fine-tune the model on the user's downstream task: a lightweight ladder side network (Sung et al., 2022) on the device fuses the returned backbone hidden states with its own local representation to predict the label. Figure 1 summarizes the two stages and the data flow between the client and server.

This two-stage design offers three practical advantages. First, it separates noise-invariance preparation from task-specific fine-tuning, addressing DP-noise amplification before any private labels are used. Second, by suppressing noise-induced variation in the backbone, it lets the ladder side network fuse stable hidden states with local features rather than reconstruct distorted representations, improving accuracy under the same privacy budget. Finally, the resulting stability enables more aggressive activation compression, reducing communication without sacrificing performance.

Contributions. In summary, our contributions are as follows:

- We propose NIFT, a novel two-stage local–server framework for privacy-preserving fine-tuning of LLMs. The first stage prepares a noise-invariant LoRA-augmented backbone, and the second stage performs task-specific split fine-tuning with a lightweight client-side ladder network.
- We develop a server-side noise-consistency objective that improves robustness without increasing client exposure: the client sends only one DP-perturbed embedding, while the server constructs K noisy copies internally to train stable hidden representations.
- We show that noise-invariant representations make privacy-preserving split fine-tuning more effective and communication-efficient, enabling aggressive activation compression while preserving accuracy.
- Experiments on five text classification datasets show that NIFT consistently outperforms HiddenEcho (Li et al., 2026), with especially large gains under stringent privacy budgets. Code will be released upon publication.

2 Related Work

Split learning and side-tuning for LLMs. Split learning was originally proposed as a privacy-preserving distributed training paradigm: the model is partitioned at a cut layer so that raw data remains on the client, while only intermediate activations are exchanged with the server (Vepakomma et al., 2018; Thapa et al., 2022). Recent LLM fine-tuning systems adapt this partitioned execution model to make large models usable when local devices cannot host or update the full backbone: SplitFrozen (Ma et al., 2025) freezes device-side layers and trains a server-side LoRA module, MobiLLM (Li et al., 2025) offloads back-propagation through quantized activation transfer, and Shi et al. (2022) reduces communication via low-rank decomposition of the cut-layer weights. Related systems address similar access and deployment constraints from different angles: Offsite-tuning (Xiao et al., 2023) lets users tune a compact emulator and adapter without accessing the full foundation model, while Petals (Borzunov et al., 2023) distributes Transformer blocks across peers for collaborative inference and parameter-efficient fine-tuning. Although these systems improve model accessibility, compute efficiency, and bandwidth usage, they generally do not address the privacy risks that remain in the transmitted activations and gradients.

Privacy in distributed fine-tuning. Cross-device traffic can leak sensitive information: gradients may reveal training inputs (Zhu et al., 2019; Geiping et al., 2020; Jin et al., 2021), and an honest-but-curious server may reconstruct private data from intermediate activations (Erdogan et al., 2022; Shu et al., 2025). A standard mitigation is to apply local differential privacy to embeddings or gradients (Dwork & Roth, 2014; Qu et al., 2021; Shen et al., 2023; Yu et al., 2022; Li et al., 2022), but the injected perturbation is repeatedly transformed by subsequent Transformer blocks, causing noise amplification and substantial utility loss. The closest prior work addresses this degradation through post-hoc repair: HiddenEcho (Li et al., 2026) returns per-layer hidden states for layer-wise client-side correction, extending the earlier Split-and-Denoise approach (Mai et al., 2024). These methods mitigate the effect of amplified noise after it appears, whereas our goal is to make the server-side representations less sensitive to the injected noise in the first place.

Consistency-based representation learning. The first stage of NIFT is related to consistency-based regularization, which encourages a model to produce similar representations or predictions under input perturbations (Miyato et al., 2019; Xie et al., 2020). Randomized smoothing makes a similar connection between Gaussian noise injection and robustness, including in differentially private settings (Cohen et al., 2019; Lecuyer et al., 2019). We adapt this principle to privacy-preserving split fine-tuning: instead of using perturbations as data augmentation for downstream classification, we prepare the frozen backbone to absorb client-injected DP noise before the resulting representations are passed to the side network.

3 Problem Formulation

Parties and assets. A client device $\mathcal{C}$ owns a private labeled dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$ and seeks to fine-tune a large pre-trained model for its local task. A server $\mathcal{S}$ hosts a backbone model f_θ that

maps an input embedding sequence to a sequence of hidden states in each layer. The client fine-tunes a lightweight local model. For each private data sample (x_i, y_i), the client maps the input tokens x_i to clean embeddings, injects calibrated noise, and transmits only the perturbed embeddings to the server. The server processes the perturbed embeddings and returns a subset of hidden states, which the client uses to make predictions.

Noise mechanism. We adopt the local-DP embedding mechanism used by recent split-LLM systems (Qu et al., 2021; Shen et al., 2023; Mai et al., 2024; Li et al., 2026). Formally, the DP-perturbed embedding is constructed by adding an unbounded noise term ξ_t to the clipped embedding:

$$\tilde{e}_t = \Pi_{\|\cdot\|_2 \le c}(e_t) + \xi_t, \tag{1}$$

where $\Pi_{\|\cdot\|_2 \le c}$ clips each token embedding to the ℓ_2 ball of radius c: embeddings with norm at most c are left unchanged, while embeddings with larger norm are rescaled to have norm c. The noise term ξ_t is drawn independently according to the ℓ_2-Laplace mechanism, with density

$$p(\xi_t) = \frac{\eta^d}{C_d \Gamma(d) c^d} \exp\left(-\frac{\eta \|\xi_t\|_2}{c}\right), \tag{2}$$

where d is the embedding dimension, $C_d = 2\pi^{d/2}/\Gamma(d/2)$ is the surface area of a unit sphere in $\mathbb{R}^d$, and $\Gamma(\cdot)$ denotes the Gamma function.

Threat model. The server is honest-but-curious: it follows the protocol but may inspect and store any message it receives. The attack surface consists of two channels: (i) the embeddings e sent to the server in the forward pass; (ii) the gradients sent back from the client during back propagation. We aim to defend against both attacks.

4 METHODOLOGY

Our solution consists of two stages of training that share the same noise specification $(\mathcal{N}_\eta, c)$.

4.1 STAGE 1: NOISE-INVARIANT BACKBONE PRE-TRAINING

On the server, at first, we train an adaptation $\Delta\theta$ on top of the frozen backbone f_θ so that $f_{\theta+\Delta\theta}$ produces hidden states that are invariant to the noise distribution $\mathcal{N}_\eta$. Specifically, $\Delta\theta = \{B_\ell A_\ell\}$ is a rank-r LoRA adapter applied to the query, key, value, and output projections of every Transformer block, following common practice (Hu et al., 2022). The backbone parameters θ remain fixed, and only $\Delta\theta$ is updated.

Server-side multi-sample perturbations. For each input x in a batch, the client computes the clean embedding $e = Ex$, applies the DP embedding mechanism in Eq. (1), and sends the resulting perturbed embedding $\tilde{e}_0$ to the server. To obtain multiple stochastic views, the server generates K noisy variants $\tilde{e}_1, \ldots, \tilde{e}_K$ by adding independently sampled perturbations to $\tilde{e}_0$ following the same ℓ_2-Laplace noise distribution.

Pairwise consistency objective. Let $h_k^{(\ell)} = f_\ell(\tilde{e}_k; \theta + \Delta\theta)$ denote the hidden state at backbone layer ℓ for the k-th noisy copy. We penalize the average pairwise ℓ_2 distance among the hidden states of the K noisy copies:

$$\mathcal{L}_{\mathrm{P}}^{(\ell)} = \frac{2}{K(K-1)} \sum_{1 \le i < j \le K} \left\| h_i^{(\ell)} - h_j^{(\ell)} \right\|_2. \tag{3}$$

For a set of selected layers $\mathcal{S}$, we define the overall consistency objective as the average loss across these layers:

$$\mathcal{L}_{\mathrm{P}} = \frac{1}{|\mathcal{S}|} \sum_{\ell \in \mathcal{S}} \mathcal{L}_{\mathrm{P}}^{(\ell)}. \tag{4}$$

Minimizing Eq. (3) encourages hidden states of all K noisy copies of the same input to coincide, thereby making the backbone less sensitive to input noise within the batch.

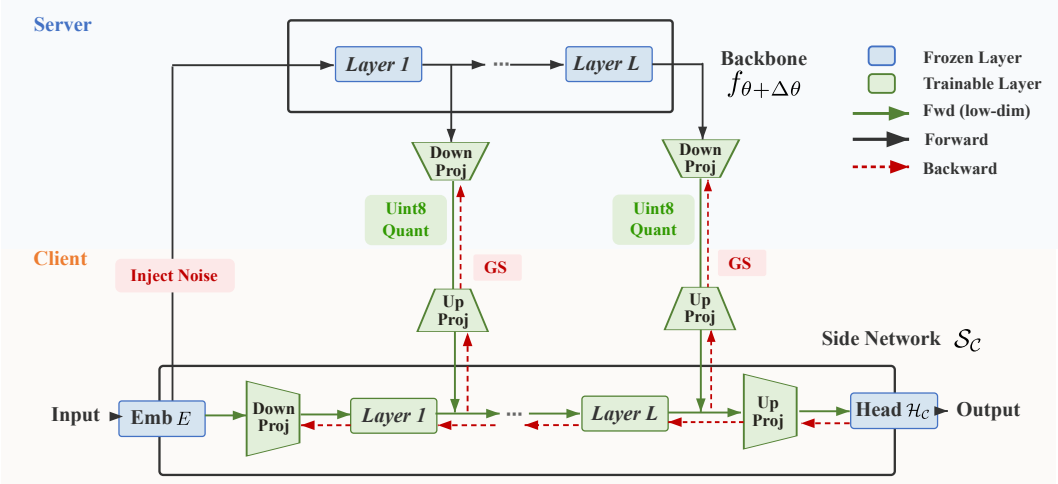

Figure 2: Stage 2 architecture in detail. The server (top) hosts the frozen, robust backbone $f_{\theta+\Delta\theta}$. At each selected layer, a trainable split bottleneck adapter $\mathcal{A}_S$ places its server-side down-projection and client-side up-projection on opposite sides of the cut; the forward wire format is an 8-bit PACT-quantized tensor (`uint8 quant`), and the returned gradients are compressed by Top-k sparsification with error feedback (denoted `GS`). The client side (bottom) holds the frozen embedding E and classification head $\mathcal{H}_C$, together with the trainable side network $\mathcal{S}_C$, which fuses the bottleneck adapter outputs with its own clean embedding. The red box on the left marks the DP perturbation ξ added to the embeddings before they cross the cut.

Anti-collapse regularization. Although Eq. (3) encourages invariance across noisy copies, it admits a degenerate solution in which the backbone maps all inputs to a constant representation. To avoid this failure, we monitor two diversity diagnostics during training: ℓ_2^{sample}, the mean pairwise distance between hidden states $h_k^{(\ell)}$ for different inputs under the same noisy sample, and ℓ_2^{token}, the mean pairwise distance between hidden states at different valid token positions of the same input. Formally, these diagnostics are defined as

$$\ell_2^{\text{sample}} = \frac{2}{|\mathcal{B}|(|\mathcal{B}|-1)} \sum_{i<j} \|h_i^{(\ell)} - h_j^{(\ell)}\|_2, \tag{5}$$

$$\ell_2^{\text{token}} = \frac{2}{T(T-1)} \sum_{t<t'} \|h_{k,t}^{(\ell)} - h_{k,t'}^{(\ell)}\|_2, \tag{6}$$

where $\mathcal{B}$ is the batch of inputs and T is the number of valid tokens. Representation collapse is indicated when either diagnostic approaches zero. To discourage collapse without relying on manual early stopping, we record per-layer step-0 baselines $b_s^{(\ell)}$ and $b_t^{(\ell)}$ from a single no-gradient forward pass before training, normalize the diagnostics as $r_s^{(\ell)} = \ell_2^{\text{sample},(\ell)}/b_s^{(\ell)}$ and $r_t^{(\ell)} = \ell_2^{\text{token},(\ell)}/b_t^{(\ell)}$, and add a squared-hinge penalty that activates only when either ratio falls below a margin $C \in (0, 1]$:

$$\mathcal{L}_{\text{reg}} = \tfrac{1}{|\mathcal{S}|} \sum_{\ell \in \mathcal{S}} \left[\phi(r_s^{(\ell)}) + \phi(r_t^{(\ell)})\right] \tag{7}$$

where $\phi(r) = \max(0, C - r)^2$.

The overall Stage 1 objective is $\mathcal{L}_{\text{P}} + \alpha\, \mathcal{L}_{\text{reg}}$, where $\alpha \geq 0$ controls the strength of the collapse penalty.

4.2 Stage 2: Noisy Side-Network Fine-Tuning

Stage 2 training involves three main components, illustrated in Figure 2:

- $f_{\theta+\Delta\theta}$: the frozen robust backbone on the server;

- $\mathcal{S}_\mathcal{C}$: the trainable client-side side network, implemented as the ladder side network;
- $\mathcal{A}_\mathcal{S}$: the trainable per-layer bottleneck adapter, whose down-projection module is placed on the server side and whose up-projection module is placed on the client side.

Forward pass. The server first evaluates the frozen backbone on the perturbed embedding $\tilde{e}_0$ received from the client:

$$\big(h^{(1)}, h^{(2)}, \ldots, h^{(L)}\big) = f_{\theta+\Delta\theta}(\tilde{e}_0). \tag{8}$$

For each cut layer ℓ in a chosen subset $\mathcal{R} \subseteq \{1, \ldots, L\}$, the server applies the split bottleneck adapter described below and sends a compressed subset of hidden states back to the client. The client's ladder side network (Sung et al., 2022) reduces the embedding to a side-stream tensor and, layer by layer, fuses it with the returned bottleneck features through additive layer normalization followed by a reduced Transformer decoder block. The final pooled representation is fed to the classification head.

Split bottleneck adapter. The cut-layer adapter serves a dual purpose: it compresses the hidden states crossing the device-server boundary to reduce bandwidth, and provides a learnable interface that lets the client adapt to the server's representations. Our adapter design follows the bottleneck side-tuning approach of Quantized Side Tuning (QST) (Zhang et al., 2024), extended with PACT-style activation quantization (Choi et al., 2018) at the cut to further compress the on-the-wire payload from BF16 to uint8. For each cut layer ℓ, the adapter's two projections live on opposite sides of the cut:

$$\text{(server)}\quad u^{(\ell)} = \mathrm{LN}\big(W^{(\ell)}_{\mathrm{down}} h^{(\ell)} + b^{(\ell)}_{\mathrm{down}}\big), \tag{9}$$

$$\text{(wire)}\quad \hat{u}^{(\ell)} = \mathcal{Q}_{\alpha^{(\ell)}}\big(u^{(\ell)}\big), \tag{10}$$

$$\text{(client)}\quad v^{(\ell)} = W^{(\ell)}_{\mathrm{up}}\, \mathcal{Q}^{-1}_{\alpha^{(\ell)}}\big(\hat{u}^{(\ell)}\big), \tag{11}$$

where r is the reduction factor, $u^{(\ell)}, \hat{u}^{(\ell)} \in \mathbb{R}^{T\times b}$ with bottleneck dimension b, $v^{(\ell)} \in \mathbb{R}^{T\times d_s}$ with $d_s = d/r$ the side-network dimension, and $\mathcal{Q}_\alpha(\cdot)$ is a PACT-style activation quantizer with a per-layer learnable clipping bound $\alpha^{(\ell)}$. Concretely, $\mathcal{Q}_\alpha$ clips its input to $[0, \alpha]$, divides by $\alpha/255$, and rounds to the nearest integer. The client then rescales by multiplying with $\alpha/255$ to recover an approximation of the original activation.

Backward pass and gradient sparsification. During back-propagation, the client-side loss sends gradients through the device-side up-projection, after which the corresponding gradient signal must cross the cut to update the server-side down-projection. We reduce this cross-device traffic with Top-k sparsification and error feedback (Lin et al., 2018; Stich et al., 2018), transmitting only a small fraction $\rho \in (0, 1]$ of the elements of the gradient tensor while accumulating the omitted elements for later updates. Beyond bandwidth savings, this compression also limits the gradient information available to a potential adversary. Gradient reconstruction attacks rely on access to the elements of the gradient tensor (Zhu et al., 2019); prior evidence shows that pruning more than 20% of them can already make recovered inputs unrecognizable, and our practical settings keep only $\rho = 0.05$ of them.

Loss function. Let $\hat{y} = \mathcal{H}_\mathcal{C}(\mathcal{S}_\mathcal{C}(\{v^{(\ell)}\}_{\ell\in\mathcal{R}}))$ denote the prediction on the client. Stage 2 minimizes the supervised task loss over the parameters of $\mathcal{S}_\mathcal{C}$ and $\mathcal{A}_\mathcal{S}$:

$$\mathcal{L}_{\mathrm{ft}} = \mathbb{E}_{(x,y)\sim\mathcal{D}}\big[\mathcal{L}_{\mathrm{task}}(\hat{y}, y)\big] \tag{12}$$

Since this work targets downstream classification tasks, we instantiate $\mathcal{L}_{\mathrm{task}}$ as the cross-entropy loss.

4.3 Privacy Interpretation

The end-to-end protection provided by our framework stems from three complementary sources. First, the client-side DP mechanism provides the standard local-DP guarantee for information transmitted from the device during the forward pass. Second, stage 1 does not weaken this guarantee; rather, it reshapes the backbone so that downstream representations become less sensitive to the injected noise. Third, the back-propagated gradients are Top-k sparsified, which not only reduces communication overhead but also places the gradients in a regime where gradient-inversion attacks are empirically unable to recover the input (Zhu et al., 2019).

4.4 COMMUNICATION OVERHEAD

We assume BF16 (2-byte) precision for all transmitted tensors, except for PACT-quantized cut activations, which are represented as uint8 (1-byte). Let N denote the number of training samples, T_{ft} the number of fine-tuning epochs, S the sequence length, L the number of transmitted layers, b the PACT bottleneck dimension used by NIFT, d the reduced hidden dimension used by HiddenEcho, and ρ the Top-k gradient keep ratio. The data transfer costs for hidden states and gradients between server and client, measured in bytes, are

$$\mathcal{C}_{\text{NIFT}} = bSLNT_{\text{ft}}(1 + 2\rho) \tag{13}$$

for NIFT, and

$$\mathcal{C}_{\text{HE}} = 4dSLNT_{\text{ft}} \tag{14}$$

for HiddenEcho. In our experiments, we set $b = 8$ and $\rho = 0.05$ for our method, while HiddenEcho uses $d = 96$. Under these settings, NIFT incurs only approximately 2.3% of HiddenEcho's data communication cost while achieving better accuracy.

5 EXPERIMENTS

5.1 EXPERIMENTAL SETUP

Datasets. We evaluate on five text classification datasets. Three of them are the datasets evaluated by HiddenEcho (Li et al., 2026): MRPC, a binary sentence-pair paraphrase task from GLUE (Wang et al., 2018); BBC News, which assigns news articles to five topics; and Financial Phrasebank, which labels the sentiment of financial news sentences with three classes. We add two harder binary sentence-pair tasks from GLUE: RTE (textual entailment) and WNLI (Winograd-style inference), on which DP noise amplification is most pronounced; their results are reported in Appendix B.

Baselines. Our main comparison is with HiddenEcho (Li et al., 2026), the state of the art among repair-based defenses. On MRPC, BBC News and Financial Phrasebank, we also compare with three perturbation baselines: LDP (Qu et al., 2021), which perturbs the input embeddings with d_χ-privacy noise and sends them to an unmodified server model; GAN-DP (Li et al., 2026), which instead generates the noise with a GAN-trained generator on the client; and Split-and-Denoise (SnD) (Mai et al., 2024), which corrects the noise on the client with a denoising model pre-trained on the server.

Backbones and noise mechanism. Two backbone models were used in our evaluation: Qwen2-1.5B (Yang et al., 2024) and Llama-3.2-1B (Grattafiori et al., 2024). We apply the same DP embedding mechanism as HiddenEcho (Li et al., 2026) (Eq. (1)), at the privacy budgets it uses for each backbone: $\eta \in \{100, 1000, 5000, 6000\}$ for Qwen2-1.5B and $\eta \in \{1000, 4000, 5000\}$ for Llama-3.2-1B; a larger η means less noise. The lowest budget, $\eta = 100$, is the high-noise regime in which backbone amplification is most severe.

Attacks. Following Li et al. (2026), we evaluate privacy under the embedding inversion attack (EIA) within the split-learning framework of Shen et al. (2023), assuming a white-box attacker who observes the perturbed embeddings and knows the parameters of the embedding layer. We report the empirical privacy (EP), which measures how much of the original text the attacker fails to recover; a larger EP means stronger privacy.

Stage 1 and Stage 2 hyperparameters. For Stage 1, we apply a rank-16 LoRA adapter to the query, key, value, and output projections of every Transformer block, draw $K = 16$ noisy copies per input, and fix the collapse weight to $\alpha = 1.0$. The number of Stage 1 training steps is selected per dataset on validation AUC, and zero, which skips Stage 1 altogether, is one of the candidates (Appendix A). For Stage 2, we set the side-network reduction factor to $r = 16$, the gradient keep ratio to $\rho = 0.05$, the batch size to 64, and the learning rate to 4×10^{-4}. More evaluation details and hyperparameter selection procedures are described in Appendix A.

Table 1: AUC and empirical privacy (EP) of different perturbation methods on MRPC, BBC News and Financial Phrasebank with a Qwen2-1.5B backbone.

Dataset		MRPC				BBC News				Financial			
Privacy budget η		100	1000	5000	6000	100	1000	5000	6000	100	1000	5000	6000
GAN-DP	AUC	0.497	0.532	0.597	0.612	0.606	0.620	0.684	0.720	0.501	0.524	0.618	0.629
	EP	*1.000*	*0.999*	*0.999*	*0.998*	*0.995*	*0.991*	*0.971*	*0.962*	*1.000*	*0.999*	*0.997*	*0.992*
LDP	AUC	0.551	0.557	0.553	0.599	0.648	0.646	0.736	0.803	0.596	0.595	0.629	0.617
SnD	AUC	0.513	0.513	0.526	0.533	0.627	0.628	0.629	0.637	0.558	0.565	0.595	0.630
HiddenEcho	AUC	0.660	0.655	0.666	**0.668**	0.732	0.747	0.805	0.951	0.857	0.855	0.860	0.866
Ours	AUC	**0.683**	**0.688**	**0.704**	0.662	**0.993**	**0.992**	**0.995**	**0.991**	**0.890**	**0.901**	**0.892**	**0.896**
	EP	*0.988*	*0.987*	*0.956*	*0.867*	*0.973*	*0.972*	*0.914*	*0.820*	*0.988*	*0.987*	*0.967*	*0.886*

Table 2: Performance comparison based on Llama-3.2-1B.

Dataset		MRPC			BBC News			Financial		
Privacy budget η		1000	4000	5000	1000	4000	5000	1000	4000	5000
HiddenEcho	AUC	0.645	0.653	0.655	0.971	0.972	0.974	0.828	0.824	0.829
Ours	AUC	**0.685**	**0.690**	**0.690**	**0.998**	**0.998**	**0.998**	**0.892**	**0.895**	**0.904**
EP		*0.951*	*0.889*	*0.809*	*0.903*	*0.803*	*0.700*	*0.952*	*0.897*	*0.848*

5.2 Overall Performance

Table 1 compares NIFT with the baseline perturbation methods on MRPC, BBC News and Financial Phrasebank. NIFT outperforms the baselines on almost all the test cases, with most gains in the range of $+0.02$ to $+0.04$ for AUC. The advantage of our method is more prominent under high noise: $+0.26$ AUC on BBC News at $\eta = 100$ and $+0.119$ AUC on WNLI at $\eta = 5000$, compared to the best of previous methods. The single exception is MRPC at $\eta = 6000$, where HiddenEcho leads by 0.006 AUC.

With a Llama-3.2-1B backbone (Table 2), evaluated at the privacy budgets used for this backbone by Li et al. (2026), NIFT outperforms HiddenEcho in all nine settings on MRPC, BBC News and Financial Phrasebank, by $+0.024$ to $+0.075$ AUC. On RTE and WNLI (Table 5 in Appendix B), it wins five of the six settings, by up to $+0.125$ AUC on WNLI; the only exception is RTE at $\eta = 1000$, where HiddenEcho leads by 0.028 AUC. NIFT's advantage therefore carries over to a different backbone family.

LDP, SnD, HiddenEcho and NIFT all perturb the embeddings on the client with the same mechanism (Eq. (1)), so at each η they share a single EP value, reported in the EP rows of Tables 1 and 2 (and of the RTE and WNLI tables in Appendix B), and their AUCs are compared at matched empirical privacy. An EP close to 1 means that the attacker recovers almost none of the original text. EP decreases as η grows, from 0.97–0.99 at $\eta = 100$ to 0.82–0.89 at $\eta = 6000$ in Table 1, because a larger budget adds less noise. GAN-DP applies a different mechanism and attains higher EP (0.96–1.00), but at a large cost in utility: its AUC is near chance on MRPC and Financial Phrasebank at $\eta = 100$ (0.497 and 0.501), and NIFT exceeds it by 0.05 to 0.39 AUC in every setting of Table 1. EP is lower on Llama-3.2-1B at the same budget (e.g., 0.951 vs. 0.987 on MRPC at $\eta = 1000$) because the two embedding matrices differ in scale and are clipped at different radii, so EP should be compared within a backbone rather than across backbones.

Table 3: AUC comparison with and without noise-invariant pre-training (Qwen2-1.5B).

Dataset	MRPC				BBC News				Financial			
Privacy budget η	100	1000	5000	6000	100	1000	5000	6000	100	1000	5000	6000
Stage 2 only	0.657	0.681	0.684	0.658	**0.993**	0.991	0.993	**0.993**	**0.890**	**0.901**	**0.892**	**0.896**
Stage 1 + 2	**0.683**	**0.688**	**0.704**	**0.662**	**0.993**	**0.992**	**0.995**	0.991	0.886	0.885	0.881	0.886

Dataset	RTE				WNLI			
Privacy budget η	100	1000	5000	6000	100	1000	5000	6000
Stage 2 only	0.602	0.576	0.570	0.563	0.693	0.667	0.748	**0.760**
Stage 1 + 2	**0.610**	**0.596**	**0.617**	**0.636**	**0.765**	**0.799**	**0.766**	0.737

5.3 Effect of Noise-Invariance Pre-Training

To isolate the contribution of noise-invariance pre-training, we compare our two-stage training against a Stage 2–only variant, in which the side network is fine-tuned directly with the original backbone.

Table 3 reports this Stage 1 ablation on all five datasets. On MRPC, Stage 1 improves all four privacy budget settings, with gains ranging from $+0.004$ to $+0.026$ AUC. On RTE, Stage 1 also improves all four settings, with the largest gain ($+0.073$ AUC) at $\eta = 6000$. The improvement is most prominent on WNLI, with AUC gains up to $+0.132$. On BBC News, both variants reach AUC ≈ 0.99 across all η, leaving no headroom for Stage 1 to contribute. On Financial Phrasebank, Stage 1 lowers AUC slightly, by 0.004 to 0.016; here the *Stage 1 + 2* row runs Stage 1 for a fixed 200 steps, whereas validation selects zero steps for the final NIFT results in Table 1, which therefore coincide with the Stage-2-only row. The same holds for WNLI at $\eta = 6000$, where the row uses 175 steps.

5.4 Data Communication Overhead

To demonstrate the communication advantage of NIFT over HiddenEcho, we measure the actual client-server traffic for hidden states and gradients. For 10 epochs of fine-tuning on MRPC (3700 samples), NIFT incurs 119 MB of data communication, while HiddenEcho requires 5.08 GB—a $43\times$ reduction. This ratio is dataset-independent. Absolute costs scale linearly with the number of training samples. In practical deployment over a typical mobile uplink ($\sim$10 Mbps), HiddenEcho's 5.08 GB transfer per MRPC fine-tuning run would require approximately 70 minutes of network time; NIFT's 119 MB requires $\sim$95 seconds. This gap becomes more critical as fine-tuning scales to larger datasets or more frequent updates.

We conducted two further ablations to isolate the components responsible for this reduction. Appendix E replaces the split bottleneck adapter with a plain linear projection of the same output dimension that sends bf16 instead of uint8 features, and Appendix F varies the gradient keep ratio between the default $\rho = 0.05$ and a more aggressive $\rho = 0.01$. Neither component trades accuracy for bandwidth: NIFT outperforms the linear-transfer variant in 17 of 20 settings while sending half the forward payload, and lowering ρ to 0.01 changes AUC by at most 0.026 while still outperforming HiddenEcho in every setting of Table 1.

6 Conclusion

We presented NIFT, a two-stage local–server framework for privacy-preserving fine-tuning of LLMs. Its key novelty is the separation of noise-invariance preparation from task adaptation: a label-free pairwise consistency objective stabilizes backbone hidden states under multiple noisy embedding views, while a lightweight client-side ladder side network adapts the robust backbone to downstream tasks. Our experiments show that NIFT improves both accuracy and communication efficiency, yielding a stronger privacy–accuracy–communication trade-off for split LLM fine-tuning.

AI USE STATEMENT

We used large language model (LLM) assistants in two ways: (i) to aid and polish the writing, by suggesting and editing wording, adapting the manuscript to the ICLR template, and laying out tables and figures; and (ii) for retrieval and discovery, by helping to find related work and to locate relevant results in prior papers. The authors reviewed and revised all LLM-assisted text and verified every citation against the original publication. The authors take full responsibility for the content of this paper.

REFERENCES

Alexander Borzunov, Dmitry Baranchuk, Tim Dettmers, Maksim Ryabinin, Younes Belkada, Artem Chumachenko, Pavel Samygin, and Colin Raffel. Petals: Collaborative inference and fine-tuning of large models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, 2023. doi: 10.18653/v1/2023.acl-demo.51.

Jungwook Choi, Zhuo Wang, Swagath Venkataramani, Pierce I-Jen Chuang, Vijayalakshmi Srinivasan, and Kailash Gopalakrishnan. Pact: Parameterized clipping activation for quantized neural networks. *arXiv preprint arXiv:1805.06085*, 2018.

Jeremy Cohen, Elan Rosenfeld, and J. Zico Kolter. Certified adversarial robustness via randomized smoothing. In *International Conference on Machine Learning (ICML)*, pp. 1310–1320, 2019.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. QLoRA: Efficient finetuning of quantized LLMs. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Cynthia Dwork and Aaron Roth. The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science*, 9(3–4):211–407, 2014. doi: 10.1561/0400000042.

Ege Erdogan, Alptekin Kupcu, and A. Ercument Cicek. UnSplit: Data-oblivious model inversion, model stealing, and label inference attacks against split learning. *Proceedings on Privacy Enhancing Technologies*, 2022. doi: 10.56553/popets-2022-0094.

Jonas Geiping, Hartmut Bauermeister, Hannah Droge, and Michael Moeller. Inverting gradients – how easy is it to break privacy in federated learning? In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, et al. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*, 2022.

Xiao Jin, Pin-Yu Chen, Chia-Yi Hsu, Chia-Mu Yu, and Tianyi Chen. CAFE: Catastrophic data leakage in vertical federated learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Mathias Lecuyer, Vaggelis Atlidakis, Roxana Geambasu, Daniel Hsu, and Suman Jana. Certified robustness to adversarial examples with differential privacy. In *IEEE Symposium on Security and Privacy (SP)*, pp. 656–672, 2019. doi: 10.1109/SP.2019.00044.

Liang Li, Xingke Yang, Wen Wu, Hao Wang, Tomoaki Ohtsuki, Xin Fu, Miao Pan, and Xuemin Shen. MobiLLM: Enabling LLM fine-tuning on the mobile device via server assisted side tuning. *arXiv preprint arXiv:2502.20421*, 2025. URL https://arxiv.org/abs/2502.20421.

Wenhao Li, Kunhao Li, and Lei Yang. HiddenEcho: Mitigating noise amplification in differentially private LLMs with hidden-state correction. In *Proceedings of the Fourteenth International Conference on Learning Representations (ICLR)*, 2026. URL https://openreview.net/forum?id=ER9BElK8He.

Xuechen Li, Florian Tramer, Percy Liang, and Tatsunori Hashimoto. Large language models can be strong differentially private learners. In *International Conference on Learning Representations (ICLR)*, 2022.

Ji Lin, Ligeng Zhu, Wei-Ming Chen, Wei-Chen Wang, Chuang Gan, and Song Han. On-device training under 256KB memory. *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Yujun Lin, Song Han, Huizi Mao, Yu Wang, and William J. Dally. Deep gradient compression: Reducing the communication bandwidth for distributed training. In *International Conference on Learning Representations (ICLR)*, 2018.

Jian Ma, Xinchen Lyu, Jun Jiang, Qimei Cui, Haipeng Yao, and Xiaofeng Tao. SplitFrozen: Split learning with device-side model frozen for fine-tuning LLM on heterogeneous resource-constrained devices. *arXiv preprint arXiv:2503.18986*, 2025. URL https://arxiv.org/abs/2503.18986.

Peihua Mai, Ran Yan, Zhe Huang, Youjia Yang, and Yan Pang. Split-and-Denoise: Protect large language model inference with local differential privacy. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, 2024. URL https://arxiv.org/abs/2310.09130.

Takeru Miyato, Shin-ichi Maeda, Masanori Koyama, and Shin Ishii. Virtual adversarial training: A regularization method for supervised and semi-supervised learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(8):1979–1993, 2019. doi: 10.1109/TPAMI.2018.2858821.

Chen Qu, Weize Kong, Liu Yang, Mingyang Zhang, Michael Bendersky, and Marc Najork. Natural language understanding with privacy-preserving BERT. In *Proceedings of the 30th ACM International Conference on Information and Knowledge Management (CIKM)*, pp. 1488–1497, 2021. doi: 10.1145/3459637.3482281.

Xicong Shen, Yang Liu, Huiqi Liu, Jue Hong, Bing Duan, Zirui Huang, Yunlong Mao, Ye Wu, and Di Wu. A split-and-privatize framework for large language model fine-tuning. *arXiv preprint arXiv:2312.15603*, 2023. URL https://arxiv.org/abs/2312.15603.

Shaohuai Shi, Qing Yang, Yang Xiang, Shuhan Qi, and Xuan Wang. An efficient split fine-tuning framework for edge and cloud collaborative learning. *arXiv preprint arXiv:2211.16703*, 2022. URL https://arxiv.org/abs/2211.16703.

Yunmeng Shu, Shaofeng Li, Tian Dong, Yan Meng, and Haojin Zhu. Model inversion in split learning for personalized LLMs: New insights from information bottleneck theory. *arXiv preprint arXiv:2501.05965*, 2025. URL https://arxiv.org/abs/2501.05965.

Sebastian U. Stich, Jean-Baptiste Cordonnier, and Martin Jaggi. Sparsified SGD with memory. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

Yi-Lin Sung, Jaemin Cho, and Mohit Bansal. LST: Ladder side-tuning for parameter and memory efficient transfer learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Chandra Thapa, Pathum Chamikara Mahawaga Arachchige, Seyit Camtepe, and Lichao Sun. SplitFed: When federated learning meets split learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 8485–8493, 2022. doi: 10.1609/aaai.v36i8.20825.

Praneeth Vepakomma, Otkrist Gupta, Tristan Swedish, and Ramesh Raskar. Split learning for health: Distributed deep learning without sharing raw patient data. *arXiv preprint arXiv:1812.00564*, 2018. URL https://arxiv.org/abs/1812.00564.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pp. 353–355, 2018. doi: 10.18653/v1/W18-5446.

Guangxuan Xiao, Ji Lin, and Song Han. Offsite-tuning: Transfer learning without full model. *arXiv preprint arXiv:2302.04870*, 2023. URL https://arxiv.org/abs/2302.04870.

Qizhe Xie, Zihang Dai, Eduard Hovy, Minh-Thang Luong, and Quoc V. Le. Unsupervised data augmentation for consistency training. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, et al. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*, 2024.

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. Differentially private fine-tuning of language models. In *International Conference on Learning Representations (ICLR)*, 2022.

Zhengxin Zhang, Dan Zhao, Xupeng Miao, Gabriele Oliaro, Zhihao Zhang, Qing Li, Yong Jiang, and Zhihao Jia. Quantized side tuning: Fast and memory-efficient tuning of quantized large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1–17, 2024.

Ligeng Zhu, Zhijian Liu, and Song Han. Deep leakage from gradients. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

A Evaluation Setups

The collapse margin C is selected by a sweep over $\{0.55, 0.65, 0.75\}$ on validation AUC: $C = 0.65$ maximizes mean AUC on MRPC and is tied with the best alternative on BBC News and Financial Phrasebank (within ± 0.002 AUC; see Appendix C). We therefore adopt $C = 0.65$ for these three datasets; on Financial Phrasebank it has no effect, since Stage 1 is skipped there (see below). For RTE and WNLI we use $C = 0.55$, the value that emerged from initial exploratory runs on these GLUE tasks (a full C sweep is left to future work). For MRPC, BBC News, and Financial Phrasebank, where the step sweep covers a broader range, we select a single Stage 1 step per dataset that maximizes mean validation AUC across the four privacy budgets, with zero steps (skipping Stage 1) among the candidates; this yields 125 for MRPC, 200 for BBC News, and 0 for Financial Phrasebank, where every non-zero step count lowers validation AUC. For RTE and WNLI, where the step sweep was restricted to a narrower range, we report the best step per $(\text{dataset}, \eta)$ cell from the available sweep, again including zero, which is selected for WNLI at $\eta = 6000$, consistent with standard per-task hyperparameter selection on GLUE (Wang et al., 2018); sensitivity to step choice is reported in Appendix D. We use the Hugging Face Trainer for Stage 2 fine-tuning and report the best validation AUC across epochs, matching HiddenEcho's reporting convention (Li et al., 2026).

B Results on RTE and WNLI

This appendix reports the RTE and WNLI results discussed in Sections 5.2 and 5.3: the comparison with HiddenEcho with a Qwen2-1.5B backbone (Table 4) and a Llama-3.2-1B backbone (Table 5).

Table 4: AUC and empirical privacy (EP) on RTE and WNLI with a Qwen2-1.5B backbone.

Dataset		RTE				WNLI			
Privacy budget η		100	1000	5000	6000	100	1000	5000	6000
HiddenEcho	AUC	0.607	0.562	0.579	0.598	0.675	0.781	0.647	0.663
Ours	AUC	**0.610**	**0.596**	**0.617**	**0.636**	**0.765**	**0.799**	**0.766**	**0.760**
EP		*0.987*	*0.987*	*0.935*	*0.876*	*0.992*	*0.989*	*0.946*	*0.900*

Table 5: AUC and empirical privacy (EP) on RTE and WNLI with a Llama-3.2-1B backbone.

Dataset		RTE			WNLI		
Privacy budget η		1000	4000	5000	1000	4000	5000
HiddenEcho	AUC	**0.638**	0.633	0.583	0.662	0.675	0.649
Ours	AUC	0.610	**0.642**	**0.618**	**0.787**	**0.775**	**0.760**
EP		*0.938*	*0.872*	*0.799*	*0.955*	*0.901*	*0.834*

C Selecting the Collapse Margin C

The collapse margin C controls when the anti-collapse hinge of Eq. (7) fires: smaller C is more permissive, larger C is stricter. We initially derived a theoretical value $C = 0.75$ from the ratio of clean to noisy-step-0 pairwise diversity on MRPC; however, our empirical sweep across $C \in \{0.55, 0.65, 0.75\}$ on the three main datasets finds $C = 0.65$ to be the best or tied-best choice everywhere.

Figure 3 reports mean validation AUC across the four privacy budgets at each Stage 1 step, with separate lines for each C value. On MRPC, $C = 0.65$ reaches the highest mean at step 125, but only 0.002 AUC above the runner-up; its clearer advantage appears in the best AUC at each budget (Table 6), where it leads $C = 0.55$ by $+0.015$ AUC and $C = 0.75$ by $+0.016$. On BBC News, all three C values are within ± 0.002 AUC because the task is already saturated near the AUC ceiling of 1.0. On Financial Phrasebank, the mean AUC across the three C values varies by less than ± 0.002. We therefore adopt $C = 0.65$ uniformly across the three datasets.

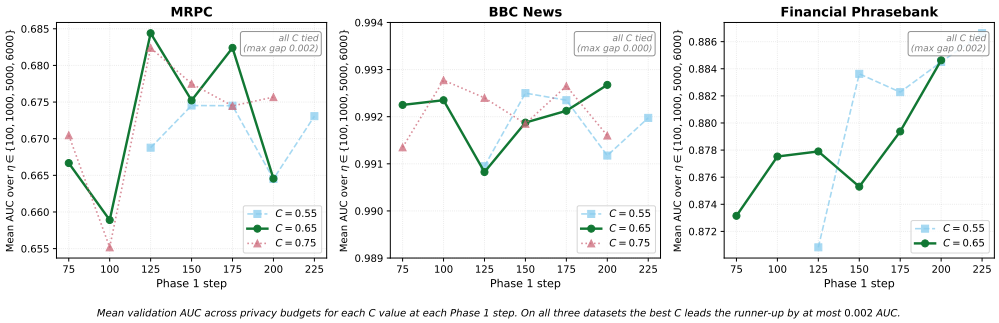

Figure 3: Mean validation AUC across $\eta \in \{100, 1000, 5000, 6000\}$ for each C value at each Stage 1 step. A gold circle would mark a C value that leads the runner-up by at least 0.005 AUC; on all three datasets the lead is at most 0.002, so every panel is marked as tied.

Per-budget results on MRPC. To understand how the collapse margin C in the anti-collapse regularizer of Eq. (7) affects NIFT's performance, we sweep $C \in \{0.55,\ 0.65,\ 0.75\}$ on MRPC across all four privacy budgets. The theoretical reference $C = 0.75$ was derived from clean-baseline diagnostics on MRPC, but our empirical sweep finds $C = 0.65$ to be the strongest choice on average.

Table 6: NIFT validation AUC on MRPC at $C \in \{0.55, 0.65, 0.75\}$. $C = 0.65$ achieves the highest mean AUC across η.

η	100	1000	5000	6000	mean
$C = 0.55$	0.695	0.667	0.674	0.697	0.683
$C = 0.65$	**0.696**	0.688	**0.704**	**0.703**	**0.698**
$C = 0.75$	**0.696**	**0.697**	0.660	0.676	0.682

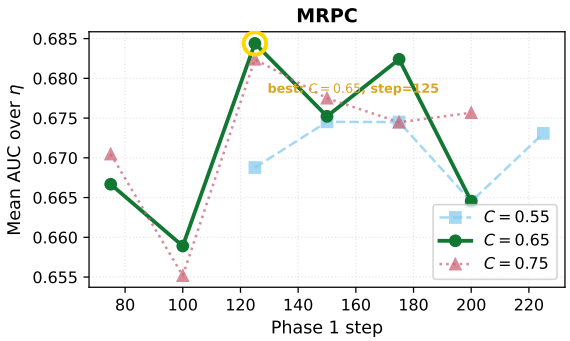

Figure 4: Mean AUC over η versus Stage 1 step on MRPC, for each value of C. The gold circle marks the global optimum at $C = 0.65$, step $= 125$.

Table 6 reports the best AUC at each (η, C) combination, and Figure 4 plots the mean AUC over η against Stage 1 step. $C = 0.65$ achieves the highest mean AUC (0.698), outperforming the theoretical $C = 0.75$ by $+0.016$ AUC and $C = 0.55$ by $+0.015$ AUC. The corresponding sweeps on BBC News and Financial Phrasebank show all three C values within ± 0.002 AUC (Figure 3).

D Sensitivity to Stage 1 Step

Figure 5 reports validation AUC against Stage 1 step for each (dataset, η) cell at the selected $C = 0.65$. Sensitivity varies by dataset. On BBC News, AUC is essentially invariant to step (range < 0.005 across the swept values), consistent with the saturation observed above. On MRPC and Financial

Phrasebank, AUC varies by up to about 0.06 across step values, motivating the per-dataset step selection used in the main results.

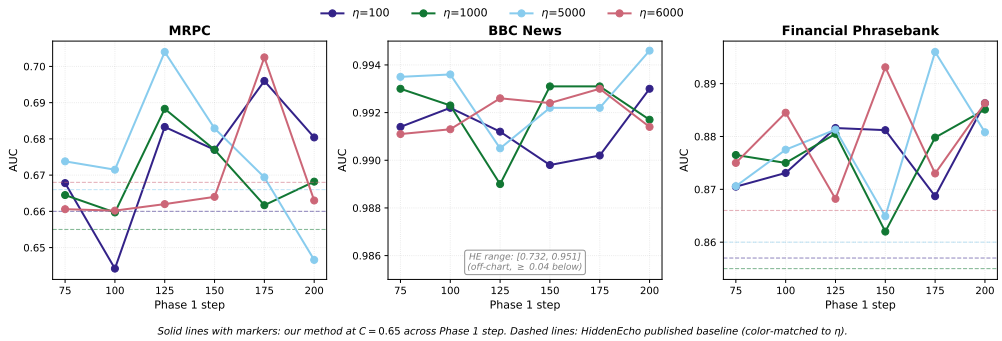

Figure 5: AUC versus Stage 1 step at $C = 0.65$ for each $(\text{dataset}, \eta)$ cell. Solid lines with markers: our method. Dashed lines: HiddenEcho published baseline at matched η. BBC News y-axis is tightened to its near-ceiling range; the HiddenEcho values fall off-chart below.

E Ablation on the Split Bottleneck Adapter

The split bottleneck adapter serves two purposes at the cut: it reduces the dimensionality of the hidden states that cross the device–server boundary, and it quantizes them to uint8 before transmission. To isolate the contribution of the second of these, we compare NIFT against a *linear transfer* variant, which uses a plain linear projection of the same output dimension and sends the projected features in bf16 rather than quantizing them with PACT, doubling the forward payload. All other components, including the Stage 1 backbone and the client-side side network, are unchanged.

Table 7: Effect of the split bottleneck adapter. *Linear transfer* replaces the bottleneck adapter with a plain linear projection at the cut, transmitting bf16 activations without PACT quantization; all other components are identical to NIFT.

Dataset	MRPC				BBC News				Financial			
Privacy budget η	100	1000	5000	6000	100	1000	5000	6000	100	1000	5000	6000
NIFT	**0.683**	**0.688**	**0.704**	0.662	**0.993**	**0.992**	**0.995**	**0.991**	**0.890**	0.901	0.892	**0.896**
Linear transfer	0.662	0.678	0.671	**0.663**	0.637	0.636	0.943	0.943	0.888	**0.903**	**0.893**	0.869

Dataset	RTE				WNLI			
Privacy budget η	100	1000	5000	6000	100	1000	5000	6000
NIFT	**0.610**	**0.596**	**0.617**	**0.636**	**0.765**	**0.799**	**0.766**	**0.760**
Linear transfer	0.579	0.570	0.576	0.573	0.671	0.696	0.629	0.655

F Ablation on the Gradient Sparsification Ratio

On the backward path, only a fraction ρ of the elements of the gradient tensor crosses the cut, with the omitted elements accumulated in an error-feedback buffer. The main experiments use $\rho = 0.05$. Here we compare that setting against $\rho = 0.01$, which transmits five times fewer elements and therefore both reduces backward traffic further and leaves a gradient-inversion adversary with less to work with.

G Local-Only Reference

The local-only reference is the side network trained on clean (unperturbed) embeddings with the backbone path disabled, i.e., only the local stream feeds the classifier and nothing is sent to the server.

Table 8: Effect of the gradient keep ratio ρ. Smaller ρ transmits fewer gradient elements on the backward path; all other components are identical.

Dataset	MRPC					BBC News					Financial				
η	100	1000	5000	6000	Avg.	100	1000	5000	6000	Avg.	100	1000	5000	6000	Avg.
$\rho = 0.05$	0.683	0.688	0.704	0.662	**0.684**	0.993	0.992	0.995	0.991	**0.993**	0.890	0.901	0.892	0.896	**0.895**
$\rho = 0.01$	0.683	0.675	0.678	0.678	0.679	0.992	0.993	0.992	0.992	0.992	0.880	0.890	0.874	0.873	0.879

Because no noise is added and no backbone signal is involved, the value is independent of the privacy budget η. Table 9 reports the AUC for each of the five datasets we evaluate. This baseline is a floor for the split framework: any benefit of the server path must show up as an improvement over it. NIFT exceeds this baseline on 4 of the 5 datasets; on Financial Phrasebank, where Stage 1 is skipped, it stays within 0.008 AUC of the reference and exceeds it at $\eta = 1000$.

Table 9: Local-only reference: side network trained on clean embeddings with the backbone path disabled. Values are η-independent.

Dataset	MRPC	BBC News	Financial	RTE	WNLI
AUC	0.651	0.990	0.898	0.560	0.525