Robust-LoRA: Confidence-Aware Fine-Tuning with Knowledge Distillation

1st Hao Wu
Shanghai University,
Shanghai, China
planet@shu.edu.cn

2nd Xiangfeng Luo*
Shanghai University,
Shanghai, China
luoxf@shu.edu.cn

3rd Jianqi Gao
Shanghai University,
Shanghai, China
jianqi_gao@shu.edu.cn

***Abstract*—Low-Rank Adaptation (LoRA) efficiently adapts pre-trained models to downstream tasks by freezing the original parameters and introducing low-rank decomposable trainable matrices. Although this approach significantly reduces computational costs, it suffers from limited performance gains and poor generalization on downstream tasks. To address this challenge, inspired by knowledge distillation that mitigates overfitting risks through soft labels conveying inter-class relationships, we propose Robust-LoRA, a robust low-rank adaptation framework that integrates knowledge distillation with margin-based constraints. Specifically, during fine-tuning, Robust-LoRA leverages knowledge distillation to enable the trainable LoRA components to retain the model's general representations, while simultaneously employing a margin loss function to widen the probability gap between correct and incorrect predictions, thereby stabilizing the decision boundary. Experimental results across multiple pre-trained backbone models and benchmark datasets demonstrate that Robust-LoRA consistently outperforms conventional low-rank adaptation methods, and exhibits superior generalization capabilities in cross-task transfer and few-shot learning scenarios.**

***Index Terms*—Parameter-efficient fine-tuning, low-rank adaptation, knowledge distillation, margin loss, robustness**

I. Introduction

Large Language Models (LLMs) have demonstrated remarkable capabilities across various natural language processing tasks. To effectively adapt these pre-trained models to downstream tasks, Parameter-Efficient Fine-Tuning (PEFT) methods have garnered substantial research attention. Among these approaches, Low-Rank Adaptation (LoRA) [1] has emerged as a particularly popular strategy due to its simplicity and effectiveness: it preserves the original pre-trained parameters while introducing trainable low-rank matrices to capture task-specific adaptations, thereby significantly reducing memory footprint and computational overhead compared to full fine-tuning.

Despite the efficiency advantages of LoRA and its variants [2], [3], the limited parameter capacity and reliance on fixed-rank incremental matrices constrain its representational potential, typically resulting in suboptimal performance gains (e.g., compared to full fine-tuning) and limited generalization capabilities on downstream tasks [4]–[6].

To address these limitations and enhance representational capacity, we propose Robust-LoRA, which tackles the aforementioned issues from two perspectives. First, inspired by knowledge distillation [7]–[9], which transfers knowledge from teacher models to student models to improve generalization performance, we employ a distillation mechanism to transfer robust representations from a teacher model to a low-rank adapted student model, thereby enhancing its representational capacity. Second, building upon this foundation, we design a margin-based loss function that explicitly enlarges the probability margin between correct and incorrect predictions, stabilizing the decision boundary. Consequently, Robust-LoRA exhibits substantial performance improvements and strong generalization capabilities.

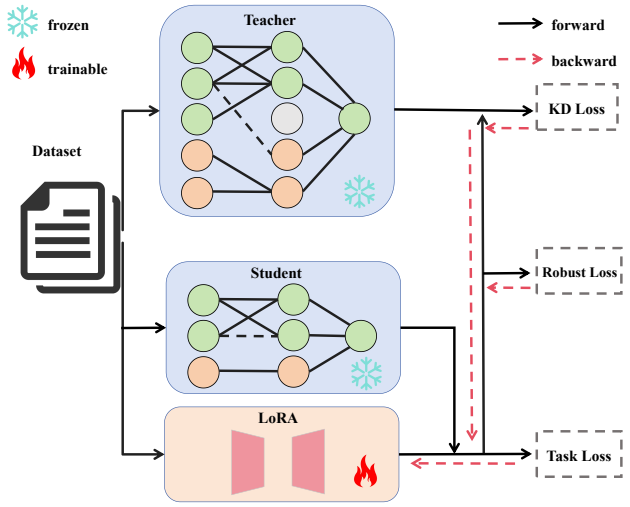


Fig. 1. Architectural overview of the Robust-LoRA framework. The framework comprises a teacher model, a low-rank adapted student model, and a multi-objective loss function that combines task-specific learning, margin-based robust loss, and knowledge distillation. Only the low-rank matrices are updated during adaptation, preserving parameter efficiency.

The main contributions of this paper are as follows:

- We propose Robust-LoRA, a robust and parameter-efficient fine-tuning framework that integrates knowledge distillation with margin-based constraints, enabling LoRA to achieve substantial performance improvements and enhanced generalization capabilities.
- We design a unified distillation protocol that transfers knowledge from teacher models to low-rank student models through temperature-scaled probability matching, thereby enhancing prediction confidence without sacrificing parameter efficiency.

* Xiangfeng Luo is the corresponding author.

- We conduct comprehensive experiments on six natural language understanding benchmark datasets using LLaMA-2-7B and RoBERTa-large as backbone models. Experimental results demonstrate that Robust-LoRA consistently outperforms strong PEFT baseline methods, with particularly pronounced advantages in few-shot and cross-task transfer scenarios.

II. Related Work

A. Parameter-Efficient Fine-Tuning

Parameter-efficient fine-tuning (PEFT) has emerged as a key paradigm for adapting large-scale pre-trained language models to downstream tasks with minimal computational overhead. Among various PEFT strategies, Low-Rank Adaptation (LoRA) [1] stands out due to its simplicity and efficacy. LoRA decomposes weight updates into low-rank matrices, enabling adaptation with fewer than 1% of trainable parameters while preserving performance comparable to full fine-tuning. Subsequent variants, such as QLoRA [10], further reduce memory consumption by integrating 4-bit quantization. In parallel, adapter-based approaches [11] insert lightweight modules between transformer layers, while prefix tuning [12] optimizes continuous task-specific prompts. Despite their efficiency, these methods primarily focus on parameter reduction and lack explicit mechanisms to ensure robustness during adaptation—a critical gap for reliable real-world deployment.

B. Robustness and Margin-Based Learning

Enhancing model robustness has been a longstanding goal in machine learning. Margin-based objectives, which enforce a separation between correct and incorrect predictions, have proven effective in improving generalization and decision boundary stability [13], [14]. Meanwhile, knowledge distillation [15] provides a mechanism to transfer robust representations from teacher models to student models. Recent efforts [16], [17] have further incorporated self-supervised objectives to bolster the reliability of language models. However, these approaches typically assume full-parameter tuning and do not address the unique robustness challenges in parameter-efficient settings. Notably, [18] revealed that PEFT methods exhibit distinct vulnerabilities under distribution shifts and adversarial perturbations. To bridge this gap, we propose a margin-augmented loss tailored for PEFT, integrated into a multi-objective optimization framework that jointly promotes accuracy, stability, and confidence.

C. Knowledge Distillation for Efficient Adaptation

Knowledge distillation (KD) has been extensively studied for model compression and efficient inference [19]. Its integration with PEFT, however, remains relatively underexplored. Recent work has begun to incorporate KD into low-rank adaptation frameworks: for instance, [7], [9], [20] employ distillation to stabilize training and enhance performance in LoRA-based adaptations. Beyond adaptation, KD has also been investigated in broader contexts such as efficient pre-training [21] and model compression [22]. Nevertheless, existing methods often treat distillation as an auxiliary objective isolated from explicit robustness constraints. This decoupling may limit model generalization under challenging conditions, such as few-shot learning or cross-task transfer. To overcome this limitation, our Robust-LoRA framework unifies knowledge distillation with a margin-based robust loss within a cohesive low-rank adaptation loop. This unified approach ensures that the student model retains the teacher's generalization ability while gaining improved prediction confidence and robustness, leading to a more reliable and adaptable PEFT solution.

III. Methodology

In this section, we first briefly revisit the standard LoRA fine-tuning paradigm. We then introduce our proposed enhancements: a margin-based robust loss to increase prediction confidence and a knowledge distillation mechanism to preserve robust representations. The integrated framework, termed Robust-LoRA, is depicted in Fig. 1.

A. Robustness-Constrained LoRA Adaptation

Following the standard LoRA formulation, we consider a pre-trained language model $\mathcal{M}$ with parameters Θ. For a weight matrix $W \in \mathbb{R}^{d\times k}$ in $\mathcal{M}$, the update is parameterized via low-rank decomposition:

$$\Delta W = BA, \quad B \in \mathbb{R}^{d\times r},\ A \in \mathbb{R}^{r\times k},\ r \ll \min(d,k) \quad (1)$$

where only B and A are trainable, keeping Θ frozen. While this design ensures parameter efficiency, the adaptation process may still suffer from ambiguous decision boundaries and overfitting, especially when training data is limited.

To enhance the discriminative ability and prediction confidence of the adapted model, we introduce a margin-based robust loss. For an input x with ground-truth label y, let $p_\theta(y|x)$ denote the predicted probability for class y, and $\max_{y'\neq y} p_\theta(y'|x)$ the highest probability among incorrect classes. The robust loss is defined as:

$$\mathcal{L}_{\text{robust}} = \frac{1}{2}\,\mathbb{E}_{(x,y)\sim\mathcal{D}}\Big[\max\big(0,\ \gamma-\big(p_\theta(y|x)-\max_{y'\neq y} p_\theta(y'|x)\big)\big)\Big]^2 \quad (2)$$

where γ is a tunable margin hyperparameter. This loss encourages the model to maintain a probability gap of at least γ between the correct class and the most competitive incorrect class. By explicitly enlarging this margin, the decision boundary becomes more stable and the model's predictions gain higher confidence, which is particularly beneficial in low-data regimes and cross-task transfers.

B. Knowledge Distillation

To further preserve the robust representations learned during pre-training, we integrate knowledge distillation into the adaptation process. Let the fixed teacher model and the low-rank adapted student model produce logits $\mathbf{z}^{(t)}$ and $\mathbf{z}^{(s)}$, respectively, for a batch of size B with C classes. We employ temperature scaling to soften the output distributions:

TABLE I
PERFORMANCE COMPARISON OF ROBUST-LORA WITH BASELINE METHODS ACROSS SIX GLUE TASKS ON LLAMA2-7B AND ROBERTA-LARGE BACKBONES (%). THE BEST RESULTS ARE IN **BOLD** AND THE SECOND-BEST RESULTS ARE UNDERLINED.

Backbone	Method	#Params	CoLA	SST-2	MRPC	QNLI	RTE	STS-B	Avg.
LLaMA2-7B	MultiLoRA	0.38%	61.3	96.3	86.3	95.5	91.0	91.8	87.0
	MixLoRA	0.17%	**73.2**	96.0	88.2	93.6	87.1	91.3	88.2
	MoRE	0.07%	66.9	96.9	89.2	94.4	**92.1**	92.2	88.6
	LoRA-MT	0.24%	65.9	**97.1**	86.0	**96.0**	91.7	91.9	88.1
	MoELoRA	0.24%	63.7	96.6	91.2	95.7	87.8	92.2	87.9
	LoRA	0.25%	68.5	96.5	90.9	93.5	90.3	90.7	88.4
	QLoRA	0.25%	68.4	96.3	90.7	92.9	90.1	91.6	88.3
	Robust-LoRA	0.25%	71.1	96.9	**91.3**	94.5	91.0	**92.3**	**89.5**
RoBERTa-large	Full FT	100.00%	68.0	96.4	**91.5**	**94.7**	86.6	**92.5**	88.3
	LoRA-XS	0.01%	67.0	95.9	90.7	93.9	88.1	92.0	87.9
	VB-LoRA	0.05%	68.3	96.1	91.4	**94.7**	86.6	91.8	88.2
	Tied-LoRA	0.02%	64.7	94.8	89.7	94.1	81.2	90.8	85.9
	Uni-LoRA	0.01%	68.5	96.3	91.3	94.6	86.6	92.1	88.2
	RED*	0.01%	68.1	96.0	90.3	93.5	86.2	91.3	87.6
	VeRA	0.02%	68.0	96.1	90.9	94.4	85.9	91.7	87.8
	LoReFT*	0.01%	68.0	96.2	90.1	94.1	87.5	91.6	87.9
	RoAd1	0.06%	66.1	96.3	91.0	94.4	**89.2**	91.7	88.1
	WGLoRA	0.17%	68.4	96.2	90.3	94.5	86.3	91.8	87.9
	AdapterFNN*	0.23%	64.4	96.1	90.5	94.3	84.8	90.2	86.7
	Adapter*	0.25%	65.4	95.2	90.5	94.6	85.3	91.5	87.1
	LoRA	0.25%	68.2	96.2	90.2	94.5	85.2	92.1	87.7
	Robust-LoRA	0.25%	**70.6**	**96.5**	91.4	94.6	88.1	92.4	**88.9**

$$\mathbf{p}_\tau^{(t)} = \mathrm{softmax}\left(\frac{\mathbf{z}^{(t)}}{\tau}\right), \qquad \mathbf{p}_\tau^{(s)} = \mathrm{softmax}\left(\frac{\mathbf{z}^{(s)}}{\tau}\right) \quad (3)$$

where τ is the temperature hyperparameter. Increasing τ raises the entropy of the output distribution, spreading probability mass from the dominant class to other classes. This provides richer, softened supervisory signals that help the student model learn the teacher's implicit knowledge about inter-class similarities.

The distillation loss is defined as the temperature-scaled Kullback–Leibler (KL) divergence between the teacher's and student's softened outputs:

$$\mathcal{L}_{\mathrm{KD}} = \tau^2 \cdot D_{\mathrm{KL}}\left(\mathbf{p}_\tau^{(t)} \,\|\, \mathbf{p}_\tau^{(s)}\right) = \frac{\tau^2}{B}\sum_{i=1}^{B}\sum_{c=1}^{C} p_{\tau,i,c}^{(t)} \log\left(\frac{p_{\tau,i,c}^{(t)}}{p_{\tau,i,c}^{(s)}}\right) \quad (4)$$

Through $\mathcal{L}_{\mathrm{KD}}$, the student model inherits not only the teacher's categorical predictions but also its implicit inter-class relational structure, which enhances generalization and mitigates overfitting.

C. Integrated Optimization

The final training objective of Robust-LoRA is a weighted combination of three loss terms:

$$\mathcal{L}_{\text{total}} = \alpha \cdot \mathcal{L}_{\text{task}} + \beta \cdot \mathcal{L}_{\text{robust}} + (1 - \alpha - \beta) \cdot \mathcal{L}_{\mathrm{KD}} \quad (5)$$

where $\mathcal{L}_{\text{task}} = \frac{1}{B}\sum_{i=1}^{B} \mathrm{CE}\big(\mathbf{y}_i, \mathbf{p}_i^{(s)}\big)$ is the standard cross-entropy task loss, and α, β are fixed coefficients that balance the contributions of task accuracy, margin enforcement, and distillation guidance.

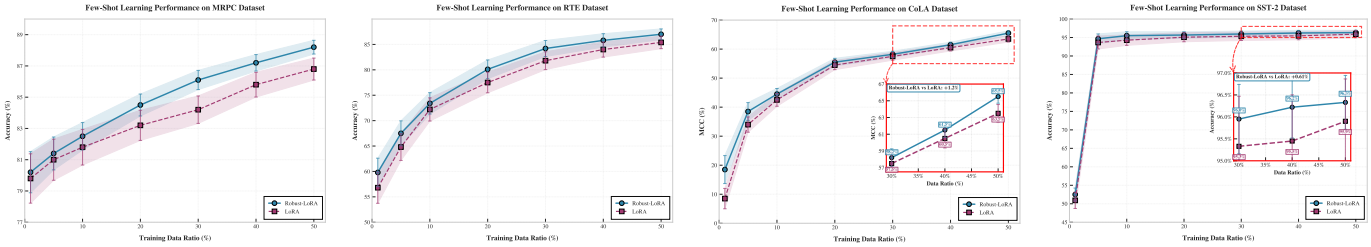
This multi-objective formulation enables simultaneous optimization of task performance via $\mathcal{L}_{\text{task}}$, prediction robustness via $\mathcal{L}_{\text{robust}}$, and representation inheritance via $\mathcal{L}_{\mathrm{KD}}$. All gradients flow exclusively through the low-rank matrices B and A, thereby preserving the parameter efficiency inherent to the LoRA framework. The combined loss function not only stabilizes the training dynamics and accelerates convergence, but also yields a model that exhibits both high accuracy and enhanced reliability across diverse downstream scenarios.

IV. EXPERIMENTS

This section presents a comprehensive empirical evaluation of the proposed Robust-LoRA framework. We begin by detailing the experimental setup, including datasets, baseline methods, and implementation specifics. Subsequently, we report and analyze the main results across multiple benchmark tasks, followed by in-depth investigations into its few-shot learning capability, parameter sensitivity, cross-task transferability, training stability, and representational properties. Ablation studies are also conducted to validate the contribution of each core component.

A. Datasets and Evaluation Metrics

The evaluation is conducted on six natural language understanding tasks from the GLUE benchmark: CoLA (linguistic acceptability), SST-2 (sentiment analysis), MRPC (paraphrase detection), STS-B (semantic textual similarity), QNLI (inference), and RTE (inference). Following common practice, we use the Matthews correlation coefficient for CoLA, the Pearson correlation for STS-B, and accuracy for the other tasks. For STS-B, we discretize the continuous similarity scores into 5 equal-width bins, and treat it as a 5-class classification problem. Results are averaged over three random runs.



(a) MRPC (Accuracy) (b) RTE (Accuracy) (c) CoLA (MCC Score) (d) SST-2 (Accuracy)

Fig. 2. Few-shot learning performance of Robust-LoRA compared to baseline methods (LoRA) across four GLUE tasks.

B. Baselines

We compare Robust-LoRA with a broad range of state-of-the-art PEFT methods. To ensure a direct comparison within the same family of techniques, we include several advanced variants of LoRA. These consist of the standard LoRA method, its quantized counterpart QLoRA [10], and extensions such as MultiLoRA [23], MixLoRA [24], MOELoRA [25], MoRE [18], and LoRA-MT [25], which respectively explore strategies involving multiple parallel adaptations, mixture mechanisms, expert ensembles, and multi-task regularization.

To contextualize the results within the wider PEFT field, we also evaluate against several prominent alternative adaptation techniques on the RoBERTa-large backbone. This set includes Adapter, which inserts trainable bottleneck modules; and several recent efficient variants like Uni-LoRA [26], VB-LoRA [27], Tied-LoRA [28], WGLoRA [29] and LoRA-XS [30].

C. Experimental Setup

We evaluate Robust-LoRA on two widely adopted pre-trained language models: LLaMA 2-7B [31] and RoBERTa-large [32]. We use the same pre-trained model as the teacher in the distillation framework.

To comprehensively assess the impact of low-rank rank selection, we conduct experiments with multiple rank values: for LLaMA 2-7B, we set the rank r to 8 and 16; for RoBERTa-large, we set r to 2 and 4. We use a batch size of 32 and a maximum sequence length of 128 across all experiments. All models are trained for 15 epochs. The model is optimized with the AdamW optimizer, and the learning rate is selected from $\{1 \times 10^{-4}, 2 \times 10^{-4}, 3 \times 10^{-4}\}$ based on validation performance, with linear decay applied during training.

The hyperparameters for the loss components were determined as follows. The margin parameter γ was set to 0.3, and the distillation temperature τ to 2.0 to appropriately smooth the probability distributions. The weighting coefficients α and β were selected via grid search over the set $\{0.1, 0.2, 0.3, 0.4, 0.5\}$ on the validation set to balance the contributions of the task loss, robust loss, and distillation loss. All experiments were conducted on a single NVIDIA A6000 GPU with 48 GB of memory.

D. Main Results

Table I summarizes the performance of Robust-LoRA and competing PEFT methods on six GLUE tasks across LLaMA2-7B and RoBERTa-large backbones. Robust-LoRA achieves the highest average score on both architectures while maintaining competitive parameter efficiency.

On the LLaMA2-7B backbone, Robust-LoRA attains an average score of 89.5%, outperforming all compared PEFT methods. It exceeds standard LoRA [1] by 1.1% and surpasses competitive variants such as MoRE [18] (88.6%) and MixLoRA [24] (88.2%). Notably, Robust-LoRA demonstrates substantial improvements on linguistically challenging tasks: on CoLA, it achieves 71.1% (Matthew's correlation), a gain of 2.7 points over standard LoRA; on STS-B, it reaches 92.3%, outperforming LoRA by 1.6%.

On the RoBERTa-large backbone, Robust-LoRA again achieves the highest average performance (88.9%) among all PEFT methods. Crucially, it outperforms not only standard LoRA (by 1.2 points) but also several ultra-efficient variants that train far fewer parameters, such as LoRA-XS [30] (0.01% of parameters), Tied-LoRA [28] (0.02%), and Uni-LoRA [26] (0.02%). While these methods achieve extreme parameter efficiency, their performance remains substantially lower—by 0.7 to 3.0 points. In contrast, Robust-LoRA attains a more favorable balance between parameter count and predictive performance, using only 0.25% of trainable parameters while delivering robust, state-of-the-art results across all six tasks.

Beyond aggregated scores, Robust-LoRA demonstrates stronger generalization in cross-task and few-shot settings (Sections IV-G and IV-E), as well as improved training stability and representation fidelity (Sections IV-H and IV-I), confirming its well-balanced design for robust and efficient adaptation.

E. Few-Shot Learning Analysis

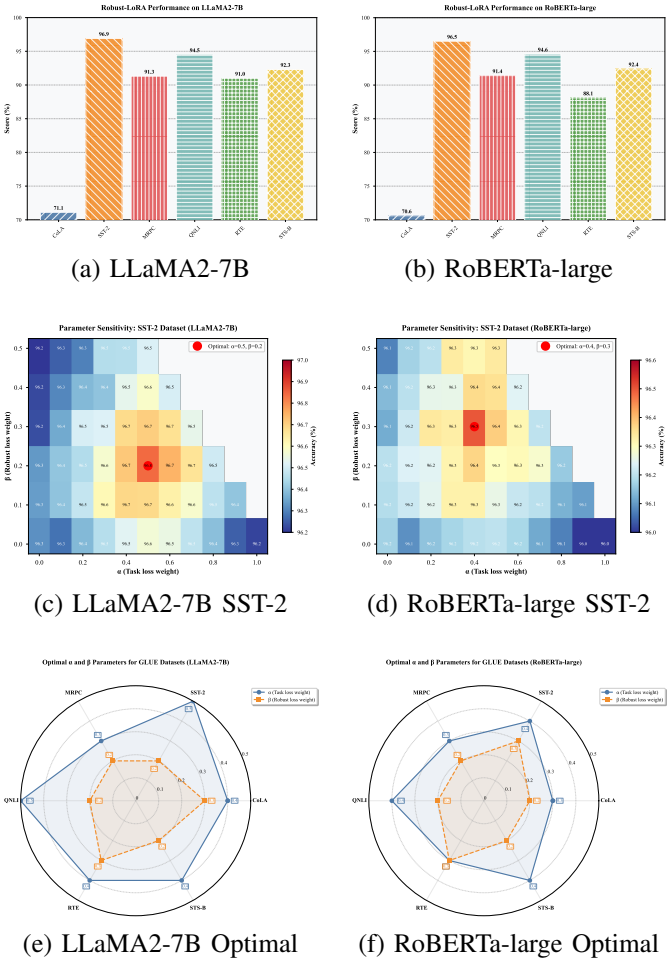
We further evaluate the efficacy of Robust-LoRA under data-scarce conditions through few-shot learning experiments on four representative GLUE tasks: MRPC, RTE, CoLA, and SST-2, using the RoBERTa-large backbone. The training data is subsampled to ratios ranging from 1% to 50% of the original datasets.

As illustrated in Fig. 2, Robust-LoRA consistently outperforms standard LoRA across all data ratios and tasks. The performance advantage is especially pronounced in extremely low-data regimes (1%–10% training data). On MRPC (Fig. 2a), Robust-LoRA sustains higher F1 scores throughout the training progression, demonstrating a more stable performance curve across varying data scales. On RTE (Fig. 2b),

our method yields more accurate and stable predictions, particularly when only a few labeled examples are available. On CoLA (Fig. 2c), Robust-LoRA achieves significantly better Matthews correlation coefficients, indicating its enhanced capacity for linguistic acceptability judgment even with minimal supervision. Similarly, on SST-2 (Fig. 2d), Robust-LoRA maintains higher accuracy across all data subsets, demonstrating robust sentiment classification capability in low-resource settings.

F. Parameter Sensitivity Analysis

We investigate Robust-LoRA's sensitivity to hyperparameters α (task loss weight) and β (robust loss weight) across two model architectures and six GLUE tasks. Robust-LoRA maintains strong performance across all tasks (Fig. 3). Heatmaps (Fig. 3(a,b)) show limited performance degradation across a wide range of parameter combinations. On SST-2, LLaMA2-7B accuracy remains above 96.2% for most α and β choices, peaking at 96.9% with $(\alpha = 0.5, \beta = 0.2)$; RoBERTa-large exhibits similar stability (96.0%–96.5%). Radar charts (Fig. 3(c,d)) reveal task-dependent optimal configurations. For tasks requiring fine-grained semantic discrimination, such as CoLA and RTE, the optimal α lies in a lower range (0.3–0.4), allowing the robust loss to play a more prominent role in stabilizing the decision boundary. In contrast, for tasks with well-separated categories, such as SST-2, a slightly higher α (0.4–0.5) is preferred, placing greater emphasis on task-specific optimization. Across all tasks, β consistently falls within the narrow range of 0.1–0.3. These results demonstrate Robust-LoRA's robustness to hyperparameter choices and its adaptability to diverse tasks, supporting practical deployment with minimal tuning.



(a) LLaMA2-7B (b) RoBERTa-large

(c) LLaMA2-7B SST-2 (d) RoBERTa-large SST-2

(e) LLaMA2-7B Optimal (f) RoBERTa-large Optimal

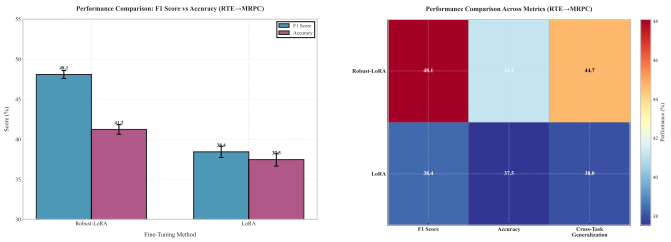
Fig. 3. Parameter sensitivity analysis of Robust-LoRA across LLaMA2-7B and RoBERTa-large. (a,b) Performance on six GLUE tasks. (c,d) Accuracy variation with α and β on SST-2. (e,f) Optimal α and β values across tasks.

G. Cross-Task Transferability Analysis

We evaluate the cross-task transferability of Robust-LoRA by fine-tuning on RTE and directly testing on MRPC using the RoBERTa-large backbone, without task-specific retraining. As shown in Fig. 4(a), Robust-LoRA achieves an F1 score of 48.1% on MRPC, outperforming standard LoRA by 9.7 percentage points (38.4%). Accuracy follows a similar trend: 41.2% versus 37.5%. This consistent improvement across metrics indicates that Robust-LoRA learns more transferable representations. Fig. 4(b) presents a heatmap comparing both methods across F1 score, accuracy, and a cross-task generalization score, defined as the weighted average of F1 and accuracy:

$$\text{Generalization Score} = 0.5 \cdot \text{F1} + 0.5 \cdot \text{Accuracy} \quad (6)$$

Robust-LoRA consistently achieves higher scores across all metrics, with a generalization score of 44.7% compared to 38.0% for LoRA. These results suggest that the advantage of our method stems from enhanced representation quality rather than a single-metric improvement.



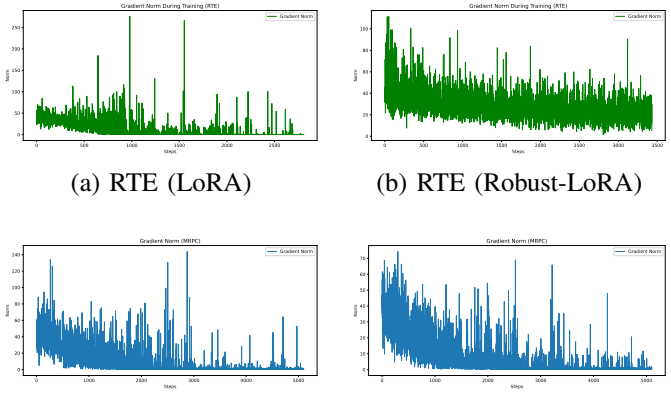
(a) Performance Comparison (b) Comprehensive Evaluation

Fig. 4. Cross-task transfer results for the RTE → MRPC scenario using RoBERTa-large.

H. Training Stability Analysis

To assess the optimization stability of Robust-LoRA, we monitor the gradient norm dynamics during fine-tuning on two GLUE tasks, RTE and MRPC, using the LLaMA2-7B backbone. Gradient norms reflect the magnitude of parameter updates; consistent and bounded gradients typically indicate stable training, whereas large fluctuations may hinder convergence and final performance.

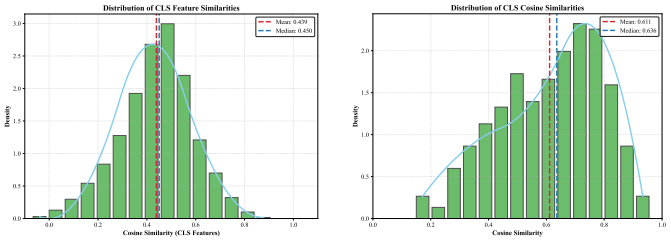
As shown in Fig. 5, Robust-LoRA exhibits significantly smoother gradient trajectories compared to standard LoRA. On RTE (Fig. 5(a,b)), Robust-LoRA reduces gradient fluctuations substantially, with the norm converging more steadily. On MRPC (Fig. 5(c,d)), a similar stabilization effect is observed, where gradient variance is markedly lower and the descent



(a) RTE (LoRA) (b) RTE (Robust-LoRA)

(c) MRPC (LoRA) (d) MRPC (Robust-LoRA)

Fig. 5. Gradient norm dynamics during fine-tuning on RTE and MRPC datasets (LLaMA2-7B). (a,b) RTE dataset showing reduced gradient fluctuations with Robust-LoRA. (c,d) MRPC dataset demonstrating similar stabilization effects. Red lines indicate average gradient norms.



(a) RTE (LoRA) (b) RTE (Robust-LoRA)

Fig. 6. Distributions of cosine similarity between teacher and student CLS embeddings on the RTE dataset. Robust-LoRA shows higher alignment with the teacher model, indicated by a higher mean and median similarity. Red and blue dashed lines mark the mean and median, respectively.

is more monotonic. These quantitative improvements indicate that our method not only achieves smoother convergence but also mitigates the risk of unstable updates that could degrade model performance.

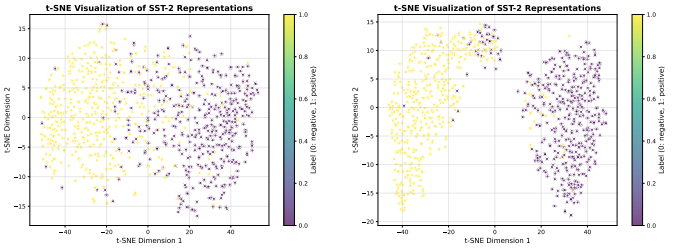
I. Representation Alignment Analysis

We quantify the representational shift induced by parameter-efficient fine-tuning by measuring the cosine similarity between the CLS token embeddings of the pre-trained teacher and the fine-tuned student models on the RTE and MRPC datasets using RoBERTa-large.

As shown in Fig. 6, Robust-LoRA maintains significantly higher cosine similarity with the pre-trained teacher than standard LoRA. On RTE, Robust-LoRA achieves a mean similarity of 0.611 (median: 0.636), compared to 0.439 (median: 0.450) for LoRA. This superior alignment results from the combined effect of the margin-based loss and knowledge distillation. The margin loss anchors the decision boundary close to the teacher's robust configuration, while distillation directly transfers stable representations. Together, they mitigate catastrophic forgetting and preserve semantic knowledge from pre-training.

J. Representation Visualization Analysis

To visually assess the quality of learned representations, we compare t-SNE projections of sentence embeddings obtained



(a) Full Fine-tuning (b) Robust-LoRA

Fig. 7. t-SNE visualizations of [CLS] token embeddings on the SST-2 dataset. (a) Full fine-tuning yields overlapping clusters between positive (yellow) and negative (purple) sentiments. (b) Robust-LoRA produces more compact and well-separated clusters, indicating more discriminative representations.

TABLE II
ABLATION STUDY ON LLAMA2-7B BACKBONE (%).

Configuration	CoLA	SST-2	MRPC	QNLI	RTE	STS-B	Avg.
Robust-LoRA (Full)	**71.1**	**96.9**	**91.3**	**94.5**	**91.0**	**92.3**	**89.5**
w/o Distillation	70.3	96.3	90.6	93.7	90.6	91.4	88.8
w/o Robust Loss	70.6	96.2	90.5	94.0	90.3	91.9	88.9

TABLE III
ABLATION STUDY ON ROBERTA-LARGE BACKBONE (%).

Configuration	CoLA	SST-2	MRPC	QNLI	RTE	STS-B	Avg.
Robust-LoRA (Full)	**70.6**	**96.5**	**91.4**	**94.6**	**88.1**	**92.4**	**88.9**
w/o Distillation	70.0	96.2	90.6	94.0	87.3	91.6	88.3
w/o Robust Loss	69.7	96.0	90.5	94.1	87.1	91.9	88.2

from full fine-tuning and Robust-LoRA on the SST-2 sentiment analysis task using RoBERTa-large.

Fig. 7 shows the t-SNE visualization of [CLS] embeddings. While full fine-tuning (Fig. 7(a)) results in partially overlapping sentiment clusters, Robust-LoRA (Fig. 7(b)) generates tighter and more distinct clusters with minimal overlap. The enhanced cluster separability demonstrates that Robust-LoRA learns more discriminative representations despite updating only a minimal set of parameters. These visual findings align with quantitative results: the clearer separation corresponds to a higher classification accuracy. This confirms that Robust-LoRA not only maintains parameter efficiency but also learns representations that are more effective for downstream tasks.

K. Ablation Studies

We perform ablation experiments to examine the contribution of each component in Robust-LoRA: the knowledge distillation (KD) mechanism and the margin-based robust loss. Results on the LLaMA2-7B and RoBERTa-large backbones are given in Tables II and III.

Removing either component consistently degrades performance, validating their individual necessity. On LLaMA2-7B, discarding KD reduces the average score by 0.7 percentage points, with the largest drop observed on CoLA (0.8%) and STS-B (0.9%). Removing the robust loss leads to a milder average decline of 0.6 percentage points, with the most pronounced degradation on MRPC (0.8%). A similar pattern holds for RoBERTa-large: omitting KD lowers the average by 0.6 percentage points, with notable drops on CoLA (0.6%), SST-2

(0.3%), and RTE (0.8%); removing the robust loss yields an average reduction of 0.7 percentage points, primarily driven by CoLA (0.9%) and MRPC (0.9%). In all cases, the full Robust-LoRA model attains the best performance, confirming that both KD and the robust loss contribute complementarily.

V. Conclusion

We present Robust-LoRA, a parameter-efficient fine-tuning framework that integrates margin-based robustness constraints and knowledge distillation into low-rank adaptation. The method enhances the discriminative ability and preserves robust representations, addressing the limited generalization of standard LoRA. Comprehensive evaluations on GLUE benchmarks with LLaMA2-7B and RoBERTa-large backbones demonstrate that Robust-LoRA consistently outperforms existing PEFT methods, particularly in few-shot and cross-task transfer scenarios. The framework achieves stable optimization, minimizes representational shift, and improves prediction confidence without sacrificing parameter efficiency. Future work may extend the approach to multimodal, continual, and domain-specific learning settings.

References

[1] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al., "Lora: Low-rank adaptation of large language models.," *ICLR*, vol. 1, no. 2, pp. 3, 2022.

[2] Hongyun Zhou, Xiangyu Lu, Wang Xu, Conghui Zhu, Tiejun Zhao, and Muyun Yang, "Lora-drop: Efficient lora parameter pruning based on output evaluation," in *Proceedings of the 31st International Conference on Computational Linguistics*, 2025, pp. 5530–5543.

[3] Dan Biderman, Jose Javier Gonzalez Ortiz, Jacob Portes, Mansheej Paul, Philip Greengard, Connor Jennings, Daniel King, Sam Havens, Vitaliy Chiley, Jonathan Frankle, et al., "Lora learns less and forgets less," *CoRR*, 2024.

[4] Yunsong Deng, Guoxu Zhou, and Qibin Zhao, "Ralo: Rank-aware low-rank adaptation for pre-trained foundation models," *Neural Networks*, p. 108423, 2025.

[5] Reece Shuttleworth, Jacob Andreas, Antonio Torralba, and Pratyusha Sharma, "Lora vs full fine-tuning: An illusion of equivalence," *arXiv preprint arXiv:2410.21228*, 2024.

[6] Krishna Prasad Varadarajan Srinivasan, Prasanth Gumpena, Madhusudhana Yattapu, and Vishal H Brahmbhatt, "Comparative analysis of different efficient fine tuning methods of large language models (llms) in low-resource setting," *arXiv preprint arXiv:2405.13181*, 2024.

[7] Injoon Hwang, Haewon Park, Youngwan Lee, Jooyoung Yang, and Sun-Jae Maeng, "Pc-lora: Low-rank adaptation for progressive model compression with knowledge distillation," *arXiv preprint arXiv:2406.09117*, 2024.

[8] Mingke Yang, Yuqi Chen, Yi Liu, and Ling Shi, "Distillseq: A framework for safety alignment testing in large language models using knowledge distillation," in *Proceedings of the 33rd ACM SIGSOFT International Symposium on Software Testing and Analysis*, 2024, pp. 578–589.

[9] Rambod Azimi, Rishav Rishav, Marek Teichmann, and Samira Ebrahimi Kahou, "Kd-lora: A hybrid approach to efficient fine-tuning with lora and knowledge distillation," *arXiv preprint arXiv:2410.20777*, 2024.

[10] Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer, "Qlora: Efficient finetuning of quantized llms," *Advances in neural information processing systems*, vol. 36, pp. 10088–10115, 2023.

[11] Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly, "Parameter-efficient transfer learning for nlp," in *International conference on machine learning*. PMLR, 2019, pp. 2790–2799.

[12] Xiang Lisa Li and Percy Liang, "Prefix-tuning: Optimizing continuous prompts for generation," *arXiv preprint arXiv:2101.00190*, 2021.

[13] Hui Yuan, Yifan Zeng, Yue Wu, Huazheng Wang, Mengdi Wang, and Liu Leqi, "A common pitfall of margin-based language model alignment: Gradient entanglement," in *13th International Conference on Learning Representations, ICLR 2025*. International Conference on Learning Representations, ICLR, 2025, pp. 97925–97944.

[14] Shuang Ao, Yi Dong, Jinwei Hu, and Sarvapali Ramchurn, "Safe pruning lora: Robust distance-guided pruning for safety alignment in adaptation of llms," *arXiv preprint arXiv:2506.18931*, 2025.

[15] Jingwen Ye, Yining Mao, Jie Song, Xinchao Wang, Cheng Jin, and Mingli Song, "Safe distillation box," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2022, vol. 36, pp. 3117–3124.

[16] Junle Li, Meiqi Tian, and Bingzhuo Zhong, "Automatic generation of safety-compliant linear temporal logic via large language model: A self-supervised framework," *arXiv preprint arXiv:2503.15840*, 2025.

[17] Yichen Gong, Delong Ran, Xinlei He, Tianshuo Cong, Anyu Wang, and Xiaoyun Wang, "Safety misalignment against large language models," in *Proceedings of the 2025 Annual Network and Distributed System Security Symposium (NDSS)*, 2025.

[18] Ajinkya Shekhar More, Rajat Sharma, and Vijay Kumar Patel, "Investigating the robustness of peft methods against adversarial attacks," in *Proceedings of the 2025 IEEE Symposium on Security and Privacy*. IEEE, 2025, pp. 1–18.

[19] Ioannis Sarridis, Christos Koutlis, Giorgos Kordopatis-Zilos, Ioannis Kompatsiaris, and Symeon Papadopoulos, "Indistill: Information flow-preserving knowledge distillation for model compression," in *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2025, pp. 9033–9042.

[20] Zeyu Ju, "A knowledge distillation-enhanced lora fine-tuning approach," in *Proceedings of the 4th International Conference on Artificial Intelligence and Intelligent Information Processing*, 2025, pp. 226–231.

[21] Rishabh Agrawal, Himanshu Kumar, and Shashikant Reddy Lnu, "Efficient llms for edge devices: Pruning, quantization, and distillation techniques," in *2025 International Conference on Machine Learning and Autonomous Systems (ICMLAS)*. IEEE, 2025, pp. 1413–1418.

[22] Runming Yang, Taiqiang Wu, Jiahao Wang, Pengfei Hu, Yik-Chung Wu, Ngai Wong, and Yujiu Yang, "Llm-neo: Parameter efficient knowledge distillation for large language models," *arXiv preprint arXiv:2411.06839*, 2024.

[23] Yiming Wang, Yu Lin, Xiaodong Zeng, and Guannan Zhang, "Multilora: Democratizing lora for better multi-task learning," *arXiv preprint arXiv:2311.11501*, 2023.

[24] Dengchun Li, Yingzi Ma, Naizheng Wang, Zhengmao Ye, Zhiyuan Cheng, Yinghao Tang, Yan Zhang, Lei Duan, Jie Zuo, Cal Yang, et al., "Mixlora: Enhancing large language models fine-tuning with lora-based mixture of experts," *arXiv preprint arXiv:2404.15159*, 2024.

[25] Yaming Yang, Dilxat Muhtar, Yelong Shen, Yuefeng Zhan, Jianfeng Liu, Yujing Wang, Hao Sun, Weiwei Deng, Feng Sun, Qi Zhang, et al., "Mtl-lora: Low-rank adaptation for multi-task learning," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2025, vol. 39, pp. 22010–22018.

[26] Kaiyang Li, Shaobo Han, Qing Su, Wei Li, Zhipeng Cai, and Shihao Ji, "Uni-lora: One vector is all you need," *arXiv preprint arXiv:2506.00799*, 2025.

[27] Yang Li, Shaobo Han, and Shihao Ji, "Vb-lora: Extreme parameter efficient fine-tuning with vector banks," *Advances in Neural Information Processing Systems*, vol. 37, pp. 16724–16751, 2024.

[28] Adithya Renduchintala, Tugrul Konuk, and Oleksii Kuchaiev, "Tied-lora: Enhancing parameter efficiency of lora with weight tying," in *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, 2024, pp. 8694–8705.

[29] Qingyun Lin, Wenlin He, qian Zhang, Zihan Peng, Zhendong Wu, and Lilan Peng, "Wglora: Efficient fine-tuning method integrating weights and gradient low-rank adaptation," in *Proceedings of the 2024 12th International Conference on Communications and Broadband Networking*, 2024, pp. 108–114.

[30] Klaudia Bałazy, Mohammadreza Banaei, Karl Aberer, and Jacek Tabor, "Lora-xs: Low-rank adaptation with extremely small number of parameters," *arXiv preprint arXiv:2405.17604*, 2024.

[31] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al., "Llama 2: Open foundation and fine-tuned chat models," *arXiv preprint arXiv:2307.09288*, 2023.

[32] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov, "Roberta: A robustly optimized bert pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.