ICLR 2026 Poster

BA-LoRA: Bias-Alleviating Low-Rank Adaptation to Mitigate Catastrophic Inheritance in Large Language Models

Yupeng Chang[1], Yi Chang[1,2,3], Yuan Wu[1,2,*]

[1] School of Artificial Intelligence, Jilin University, [2] Key Laboratory of Symbolic Computation and Knowledge Engineering, Jilin University, [3] International Center of Future Science, Jilin University

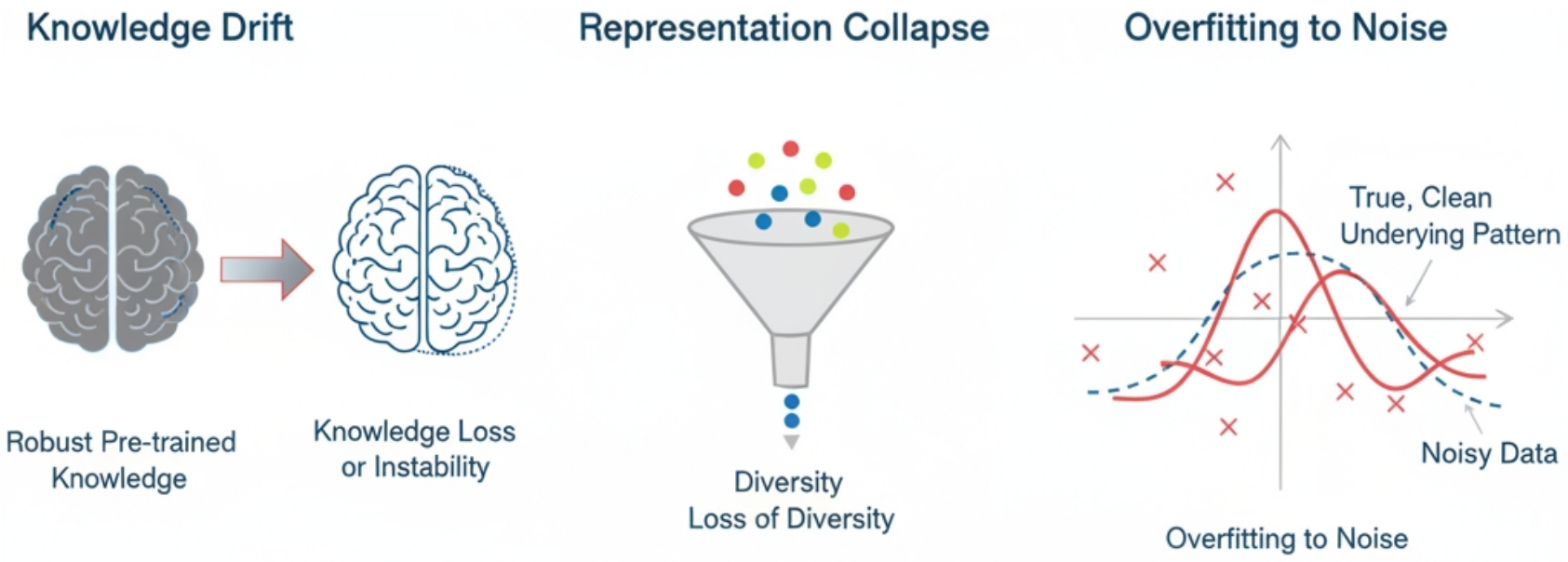
Background: PEFT can amplify inherited bias

Phenomenon: Low-rank bottlenecks may amplify inherited bias and noise during downstream adaptation.

Goal: Mitigate catastrophic inheritance with a lightweight PEFT framework.

Failure Modes of Catastrophic Inheritance

Catastrophic inheritance typically appears in three forms.



Need: a lightweight PEFT solution that addresses all three failure modes.

BA-LoRA for NLU Tasks

1. Consistency Regularization (CR-NLU): Aligns the fine-tuned model with the pre-trained model via KL divergence.

$$\mathcal{L}_{\text{CR\_NLU}} = T^2 \cdot \text{KL}(\text{softmax}(\mathbf{Z}_P/T) \,\|\, \text{softmax}(\mathbf{Z}_F/T))$$

2. Diversity Regularization (DR-NLU): Reduces inter-class output correlation via covariance regularization.

$$\mathcal{L}_{\text{DR\_NLU}} = \frac{1}{D}\sum_{i \neq j}[C(\mathbf{Z}_F)]_{i,j}^2$$

$$C(\mathbf{Z}_F) = \frac{1}{N-1}\mathbf{Z}_{\text{centered}}^T\mathbf{Z}_{\text{centered}}, \quad \text{where } \mathbf{Z}_{\text{centered}} = \mathbf{Z}_F - \bar{\mathbf{Z}}_F$$

3. SVD Regularization (SVDR-NLU): Encourages low-rank and robust output representations via SVD.

$$\mathcal{L}_{\text{SVDR\_NLU}} = -\frac{\sum_{i=1}^{k}\sigma_i}{\sum_{j=1}^{\min(N,D)}\sigma_j}$$

Overall objective for NLU: $\mathcal{L}_{\text{NLU}} = \mathcal{L}_{\text{task\_NLU}} + \lambda_1\mathcal{L}_{\text{CR\_NLU}} + \lambda_2\mathcal{L}_{\text{DR\_NLU}} + \lambda_3\mathcal{L}_{\text{SVDR\_NLU}}$

BA-LoRA for NLG Tasks

1. Consistency Regularization (CR-NLG): Aligns the fine-tuned model with the pre-trained model via KL divergence.

$$\mathcal{L}_{\text{CR\_NLG}} = T^2 \cdot \frac{1}{M} \sum_{i=1}^{M} \text{KL}(\mathcal{P}_P(y_i \mid \mathbf{x}; T) \,\|\, \mathcal{P}_F(y_i \mid \mathbf{x}; T))$$

2. Diversity Regularization (DR-NLG): Promotes diversity within the Top-K candidate tokens.

$$\mathcal{L}_{\text{DR\_NLG}} = \frac{1}{M} \sum_{i=1}^{M} \sum_{v \in \mathcal{V}_{\text{top-}K}^{(i)}} \mathcal{P}'_F(v \mid \mathbf{h}_i) \log \mathcal{P}'_F(v \mid \mathbf{h}_i)$$

3. SVD Regularization (SVDR-NLG): Encourages low-rank and robust output structures via randomized SVD.

$$\mathcal{L}_{\text{SVDR\_NLG}} = -\frac{\sum_{i=1}^{k} \tilde{\sigma}_i}{\|\mathbf{Z}_{\text{valid}}\|_F}$$

Overall objective for NLG: $\mathcal{L}_{\text{NLG}} = \mathcal{L}_{\text{task\_NLG}} + \lambda_1 \mathcal{L}_{\text{CR\_NLG}} + \lambda_2 \mathcal{L}_{\text{DR\_NLG}} + \lambda_3 \mathcal{L}_{\text{SVDR\_NLG}}$

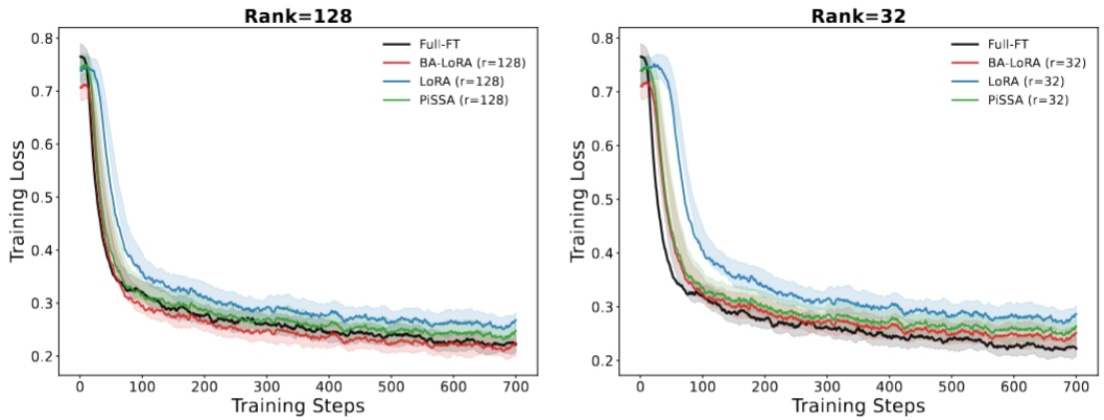
Main Results

Goal: Compare BA-LoRA against strong PEFT baselines on both NLG and NLU benchmarks.
Setup: LLaMA-2-7B for NLG and DeBERTa-v3-base for NLU.
Takeaway: BA-LoRA achieves the best overall performance across both domains. BA-LoRA converges faster and reaches lower training loss.

NLG Benchmarks

Methods	GSM8K	MATH	HumanEval	MBPP	MT-Bench	Avg
Full FT	48.9$_{\pm 0.49}$	7.48$_{\pm 0.22}$	20.52$_{\pm 0.29}$	23.64$_{\pm 0.38}$	4.85$_{\pm 0.09}$	21.08
LoRA	42.68$_{\pm 0.54}$	5.92$_{\pm 0.15}$	16.80$_{\pm 0.38}$	21.51$_{\pm 0.43}$	4.60$_{\pm 0.14}$	18.30
AdaLoRA	41.95$_{\pm 0.90}$	6.24$_{\pm 0.38}$	18.10$_{\pm 0.46}$	20.19$_{\pm 0.71}$	4.79$_{\pm 0.18}$	18.25
DoRA	41.77$_{\pm 0.74}$	6.20$_{\pm 0.48}$	16.86$_{\pm 0.54}$	21.60$_{\pm 0.49}$	4.48$_{\pm 0.14}$	18.18
MiLoRA	43.09$_{\pm 1.16}$	6.31$_{\pm 0.39}$	17.55$_{\pm 0.24}$	20.22$_{\pm 0.37}$	4.50$_{\pm 0.17}$	18.33
LoRA+	47.84$_{\pm 0.39}$	7.21$_{\pm 0.49}$	20.07$_{\pm 0.38}$	23.69$_{\pm 0.29}$	5.11$_{\pm 0.06}$	20.78
LoRA-FA	40.25$_{\pm 0.46}$	5.66$_{\pm 0.47}$	15.91$_{\pm 0.41}$	20.01$_{\pm 0.32}$	4.67$_{\pm 0.12}$	17.30
LoRA-GA	50.47$_{\pm 0.98}$	7.13$_{\pm 0.44}$	19.44$_{\pm 0.45}$	23.05$_{\pm 0.40}$	5.04$_{\pm 0.10}$	21.03
PiSSA	51.48$_{\pm 0.34}$	7.60$_{\pm 0.18}$	19.48$_{\pm 0.45}$	23.84$_{\pm 0.46}$	4.92$_{\pm 0.07}$	21.46
CorDA	53.90$_{\pm 0.56}$	8.52$_{\pm 0.27}$	21.03$_{\pm 0.37}$	24.15$_{\pm 0.44}$	5.15$_{\pm 0.09}$	22.55
CorDA++	55.03$_{\pm 0.52}$	8.95$_{\pm 0.37}$	21.76$_{\pm 0.39}$	24.74$_{\pm 0.47}$	**5.64**$_{\pm 0.12}$	23.22
BA-LoRA	**55.86**$_{\pm 0.35}$	**9.47**$_{\pm 0.52}$	**23.58**$_{\pm 0.25}$	**36.86**$_{\pm 0.31}$	5.11$_{\pm 0.05}$	**26.18**

Optimization Dynamics



NLU Benchmarks

Methods	#Params	MNLI	SST-2	MRPC	CoLA	QNLI	QQP	RTE	STS-B	Avg
Full FT	184M	90.34$_{\pm 0.18}$	**96.33**$_{\pm 0.11}$	89.95$_{\pm 1.07}$	71.43$_{\pm 0.72}$	94.24$_{\pm 0.10}$	92.11$_{\pm 0.28}$	83.75$_{\pm 1.81}$	91.04$_{\pm 0.48}$	88.65
BitFit	0.1M	89.54$_{\pm 0.29}$	94.68$_{\pm 0.11}$	87.95$_{\pm 1.33}$	67.31$_{\pm 0.49}$	92.45$_{\pm 0.17}$	88.72$_{\pm 0.45}$	79.12$_{\pm 0.39}$	91.63$_{\pm 0.37}$	86.43
HAdapter	1.22M	90.23$_{\pm 0.07}$	95.38$_{\pm 0.06}$	89.97$_{\pm 0.27}$	68.73$_{\pm 0.27}$	94.31$_{\pm 0.29}$	91.99$_{\pm 0.28}$	84.76$_{\pm 0.39}$	91.58$_{\pm 0.13}$	88.37
PAdapter	1.18M	90.42$_{\pm 0.36}$	95.49$_{\pm 0.10}$	89.71$_{\pm 0.35}$	69.04$_{\pm 0.10}$	94.38$_{\pm 0.26}$	92.15$_{\pm 0.43}$	85.53$_{\pm 0.18}$	91.69$_{\pm 0.13}$	88.55
LoRA	1.33M	90.71$_{\pm 0.16}$	94.79$_{\pm 0.16}$	89.85$_{\pm 0.21}$	70.05$_{\pm 0.34}$	93.94$_{\pm 0.09}$	92.07$_{\pm 0.48}$	85.43$_{\pm 0.09}$	91.67$_{\pm 0.29}$	88.56
DoRA	1.27M	90.48$_{\pm 0.10}$	95.85$_{\pm 0.08}$	91.04$_{\pm 0.15}$	71.03$_{\pm 0.18}$	94.21$_{\pm 0.37}$	92.34$_{\pm 0.16}$	86.19$_{\pm 0.25}$	91.92$_{\pm 0.38}$	89.13
AdaLoRA	1.27M	90.87$_{\pm 0.08}$	96.18$_{\pm 0.43}$	90.81$_{\pm 0.40}$	71.64$_{\pm 0.12}$	94.68$_{\pm 0.46}$	92.37$_{\pm 0.35}$	87.78$_{\pm 0.36}$	91.97$_{\pm 0.43}$	89.54
PiSSA	1.33M	90.47$_{\pm 0.44}$	95.81$_{\pm 0.45}$	91.48$_{\pm 0.49}$	72.27$_{\pm 0.29}$	94.41$_{\pm 0.41}$	92.21$_{\pm 0.26}$	87.14$_{\pm 0.08}$	91.93$_{\pm 0.25}$	89.47
BA-LoRA	1.33M	**91.26**$_{\pm 0.49}$	96.25$_{\pm 0.09}$	**92.11**$_{\pm 0.55}$	**75.46**$_{\pm 0.62}$	**95.35**$_{\pm 0.14}$	**93.63**$_{\pm 0.52}$	**88.58**$_{\pm 0.73}$	**92.71**$_{\pm 0.38}$	**90.67**

Why Does BA-LoRA Help?

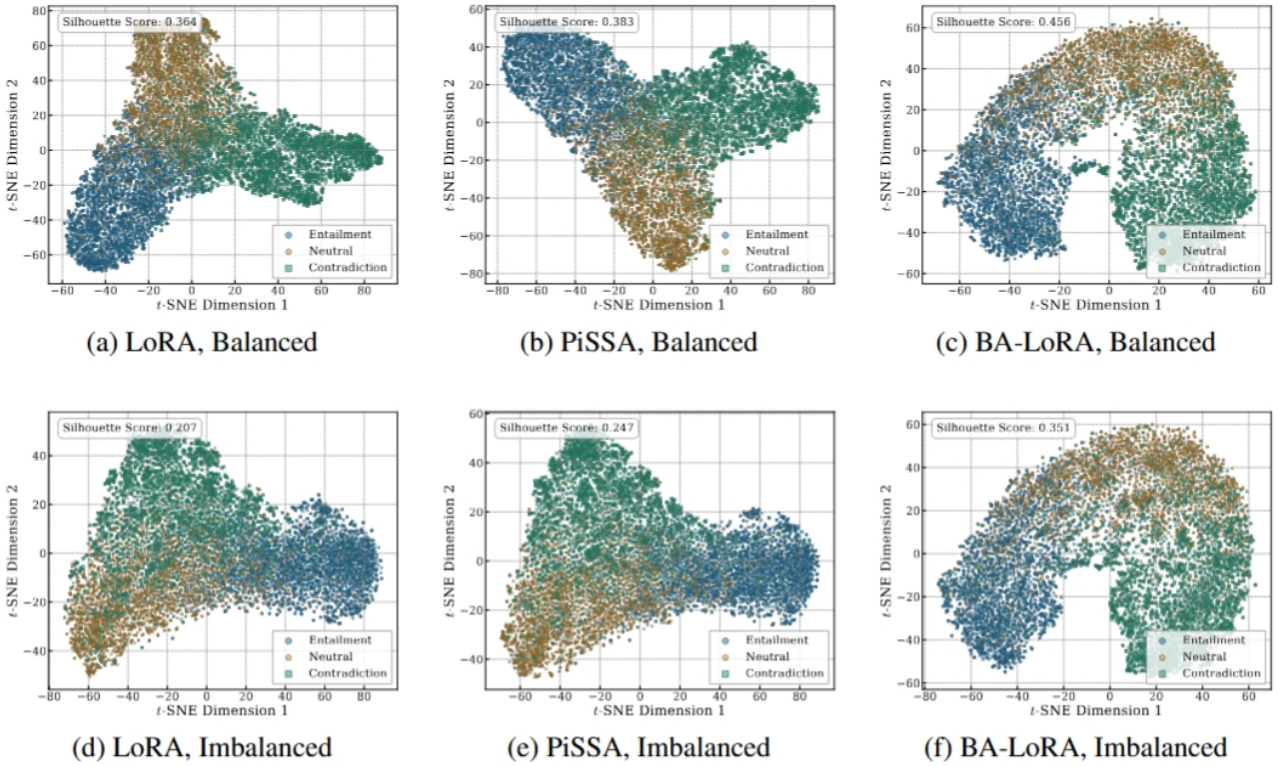
Noisier pre-training:
BA-LoRA shows larger gains when inherited noise is stronger.

Model	Methods	MNLI	SST-2	CoLA	QNLI	MRPC	Avg
RoBERTa-base	LoRA	$85.63_{\pm 0.01}$	$94.03_{\pm 0.02}$	$62.40_{\pm 0.71}$	$91.37_{\pm 0.97}$	$87.98_{\pm 0.23}$	84.28
	PiSSA	$85.72_{\pm 0.40}$	$93.64_{\pm 0.13}$	$67.28_{\pm 0.59}$	$91.40_{\pm 0.54}$	$88.11_{\pm 0.24}$	85.23
	BA-LoRA	**86.59**$_{\pm 0.58}$	**94.83**$_{\pm 0.45}$	**67.91**$_{\pm 0.21}$	**92.28**$_{\pm 0.37}$	**90.07**$_{\pm 0.32}$	**86.34**
T5-base	LoRA	$85.30_{\pm 0.04}$	$94.04_{\pm 0.11}$	$69.35_{\pm 0.05}$	$92.96_{\pm 0.09}$	$68.38_{\pm 0.01}$	82.01
	PiSSA	$85.75_{\pm 0.07}$	$94.07_{\pm 0.06}$	$74.27_{\pm 0.39}$	$93.15_{\pm 0.14}$	$76.31_{\pm 0.51}$	84.71
	BA-LoRA	**86.91**$_{\pm 0.48}$	**95.20**$_{\pm 0.29}$	**80.19**$_{\pm 1.03}$	**94.12**$_{\pm 0.32}$	**83.43**$_{\pm 0.71}$	**87.97**

Gain over PiSSA: +1.11

Gain over PiSSA: +3.26

Imbalanced data:
BA-LoRA preserves clearer feature separation under severe imbalance.



(a) LoRA, Balanced (b) PiSSA, Balanced (c) BA-LoRA, Balanced

(d) LoRA, Imbalanced (e) PiSSA, Imbalanced (f) BA-LoRA, Imbalanced

Ablation and Efficiency

Ablation Study: All three regularizers contribute positively, and the full BA-LoRA model achieves the best overall performance.

Configuration	GSM8K	MATH	Average of GLUE
Baseline (PiSSA)	51.48 ± 0.34	7.60 ± 0.18	89.47
+ $\mathcal{L}_{\text{CR}}$	54.25 ± 0.59	9.15 ± 0.25	90.18
+ $\mathcal{L}_{\text{DR}}$	53.60 ± 0.46	8.95 ± 0.18	89.85
+ $\mathcal{L}_{\text{SVDR}}$	52.95 ± 0.55	8.70 ± 0.22	89.71
BA-LoRA (Full)	$\mathbf{55.86 \pm 0.35}$	$\mathbf{9.47 \pm 0.52}$	**90.67**

Efficiency Analysis: BA-LoRA delivers strong gains with only modest additional memory and training time, achieving a favorable performance-cost trade-off.

Method	Memory Cost	Training Time	GSM8K
Full FT	>96 GB	>24h	48.9 ± 0.49
LoRA	66.32 GB	4h 31min	42.68 ± 0.54
PiSSA	66.59 GB	4h 17min	51.48 ± 0.34
BA-LoRA	77.34 GB	4h 48min	$\mathbf{55.86 \pm 0.35}$

ICLR

Thank you!