ACL 2026

BEFT: Bias-Efficient Fine-Tuning of Language Models in Low-Data Regimes

Baichuan Huang[1], Ananth Balashankar[2], Amir Aminifar[1]

[1]Lund University, Sweden
[2]Google DeepMind, USA

This research has been partially supported by the Wallenberg AI, Autonomous Systems and Software Program (WASP), the Swedish Research Council (VR), the ELLIIT Strategic Research Environment, European Union, and an unrestricted gift from Google.

Introduction and Background

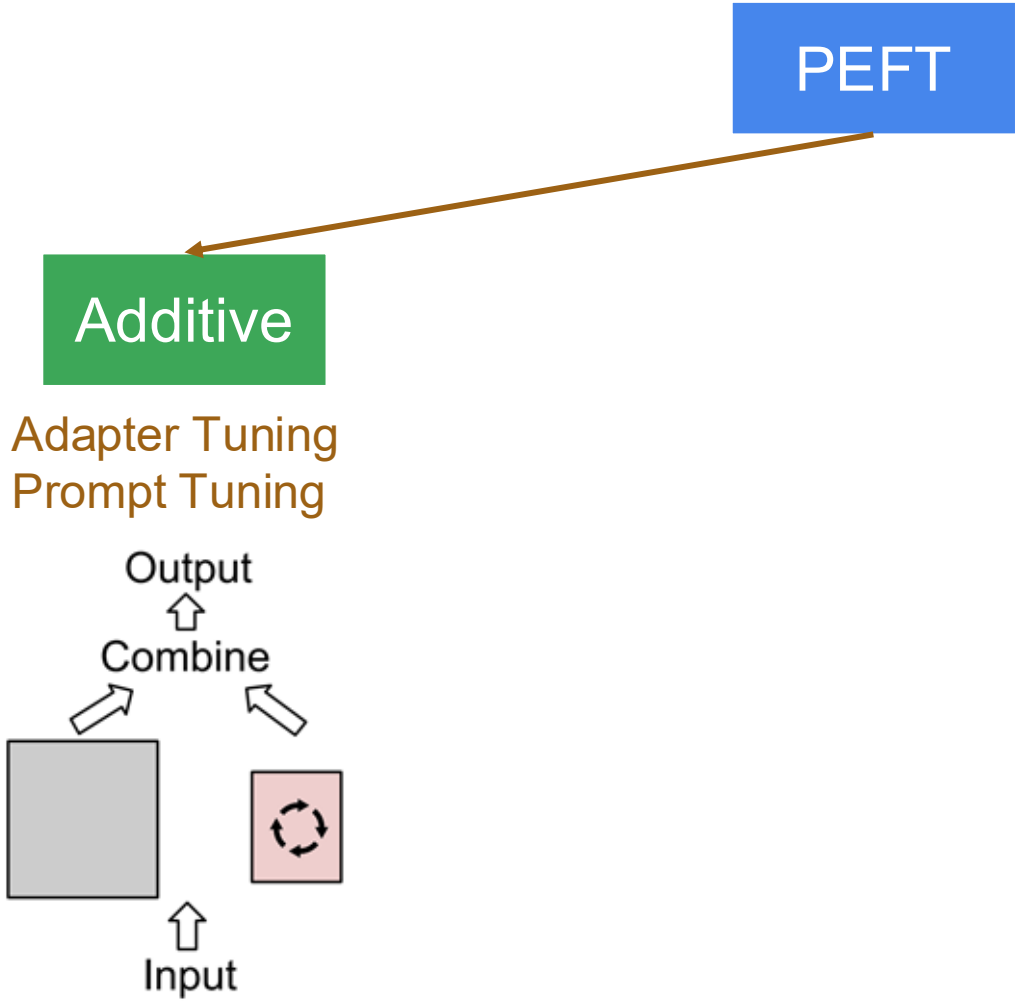
Fine-tuning LLMs

Parameter-Efficient Fine-Tuning

PEFT

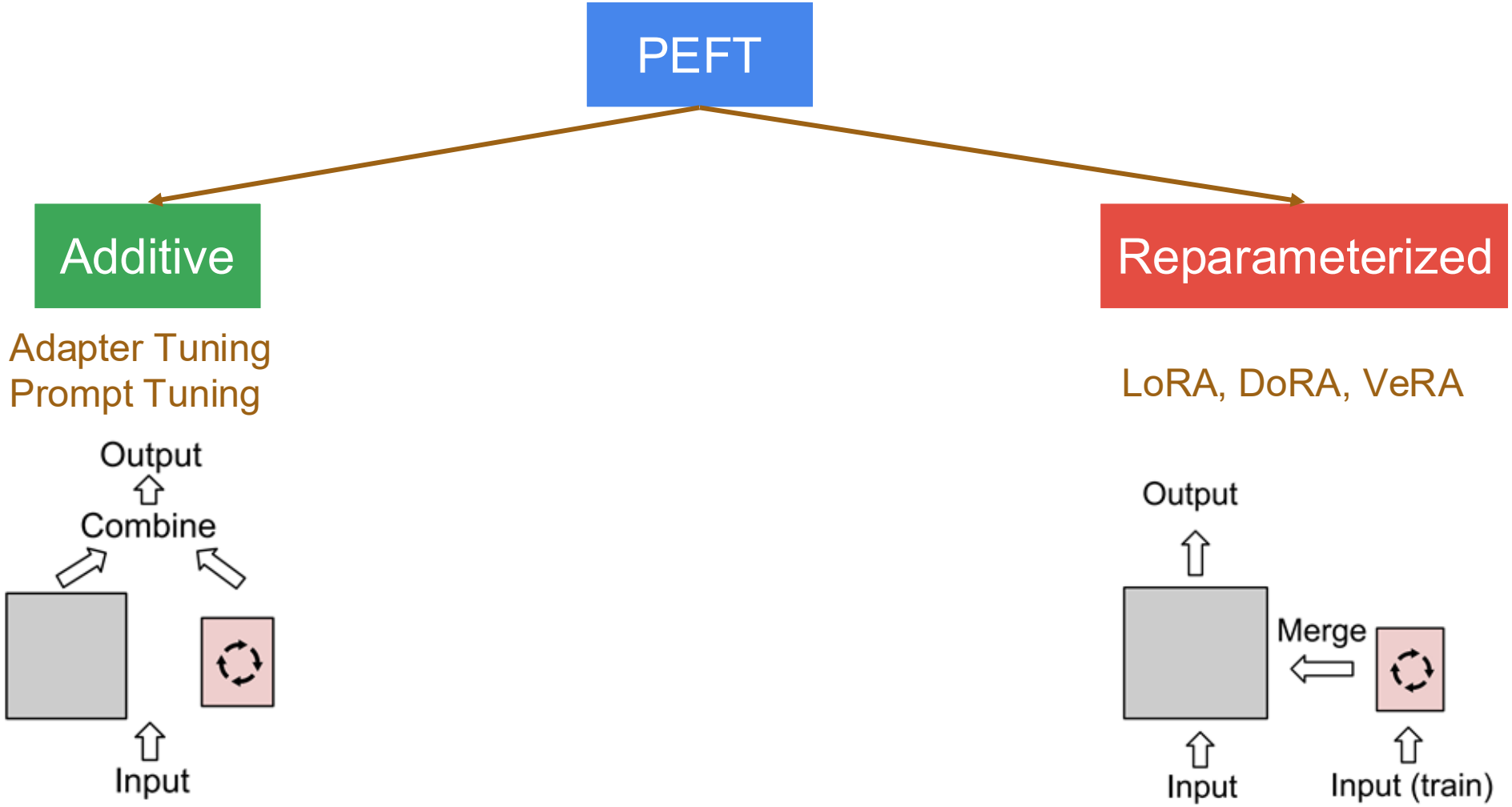
Han, Zeyu, et al. "Parameter-Efficient Fine-Tuning for Large Models: A Comprehensive Survey." Transactions on Machine Learning Research.
Zaken, Elad Ben, et al. Bitfit: Simple parameter-efficient fine-tuning for transformer-based masked language-models. ACL, 2022

Parameter-Efficient Fine-Tuning



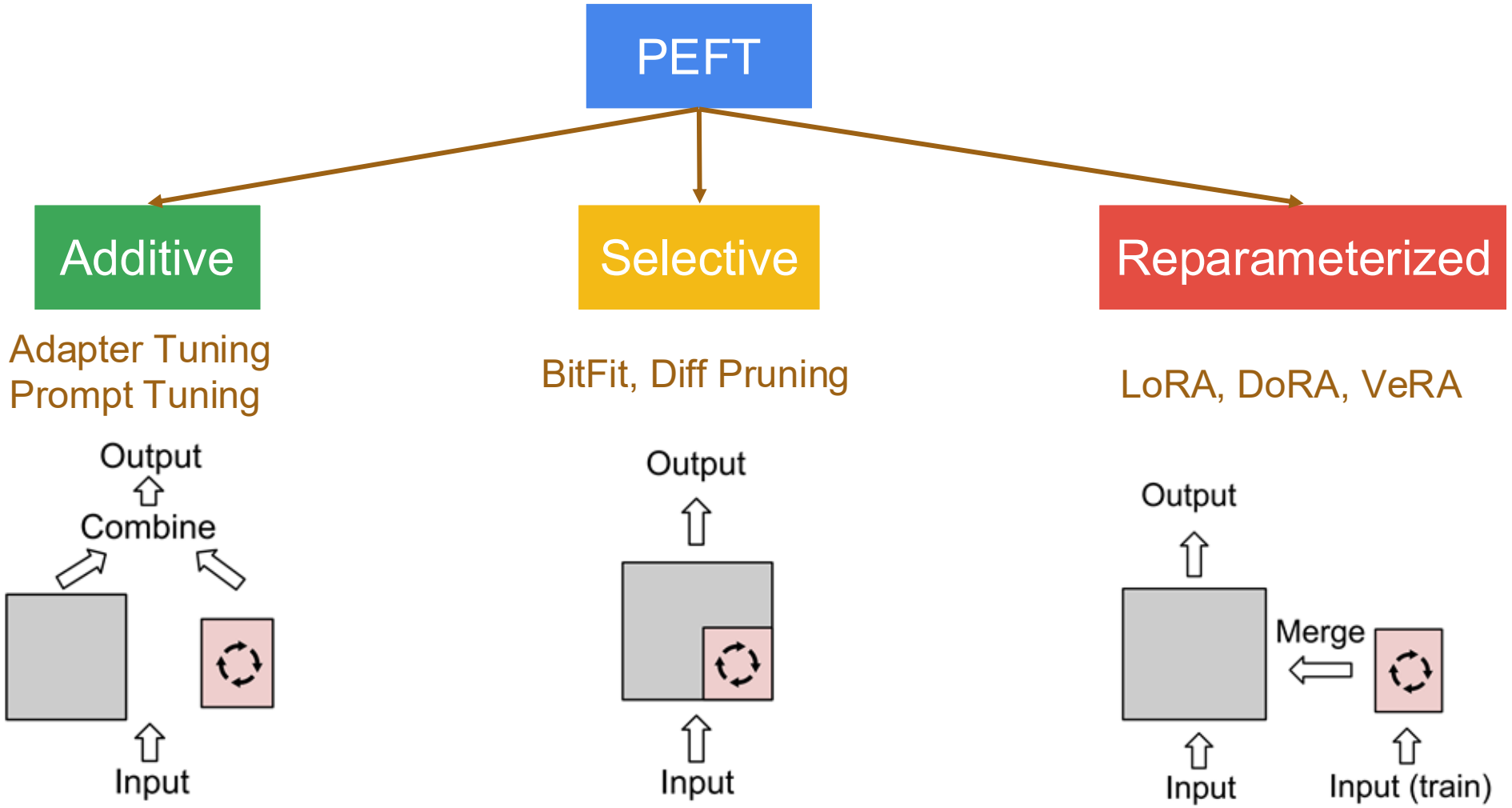
Han, Zeyu, et al. "Parameter-Efficient Fine-Tuning for Large Models: A Comprehensive Survey." Transactions on Machine Learning Research.
Zaken, Elad Ben, et al. Bitfit: Simple parameter-efficient fine-tuning for transformer-based masked language-models. ACL, 2022

Parameter-Efficient Fine-Tuning



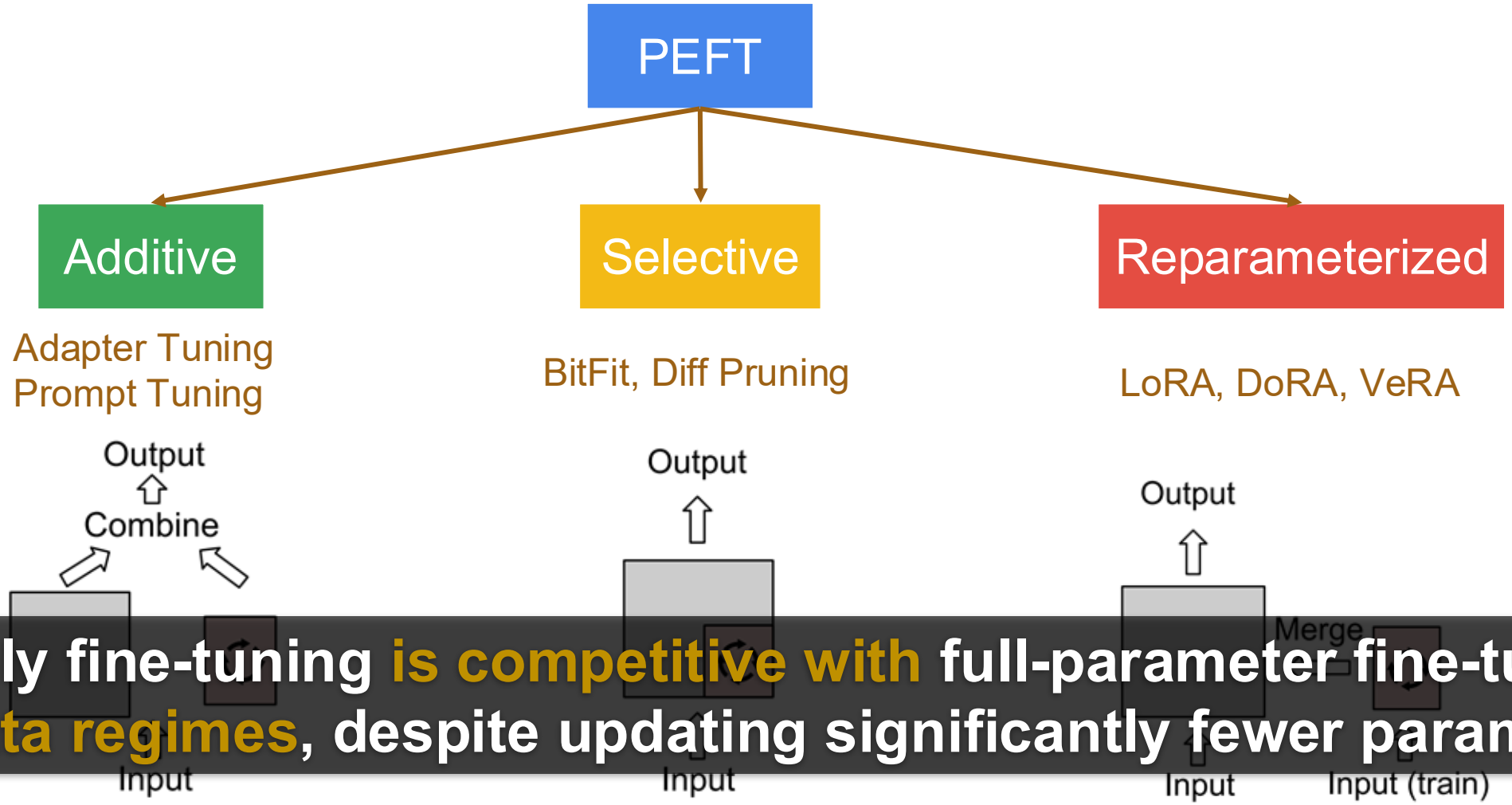
Han, Zeyu, et al. "Parameter-Efficient Fine-Tuning for Large Models: A Comprehensive Survey." Transactions on Machine Learning Research.
Zaken, Elad Ben, et al. Bitfit: Simple parameter-efficient fine-tuning for transformer-based masked language-models. ACL, 2022

Parameter-Efficient Fine-Tuning



Han, Zeyu, et al. "Parameter-Efficient Fine-Tuning for Large Models: A Comprehensive Survey." Transactions on Machine Learning Research.
Zaken, Elad Ben, et al. Bitfit: Simple parameter-efficient fine-tuning for transformer-based masked language-models. ACL, 2022

Parameter-Efficient Fine-Tuning



Bias-only fine-tuning is competitive with full-parameter fine-tuning in low-data regimes, despite updating significantly fewer parameters.

Han, Zeyu, et al. "Parameter-Efficient Fine-Tuning for Large Models: A Comprehensive Survey." Transactions on Machine Learning Research.
Zaken, Elad Ben, et al. Bitfit: Simple parameter-efficient fine-tuning for transformer-based masked language-models. ACL, 2022

Expressive Power of Bias Terms

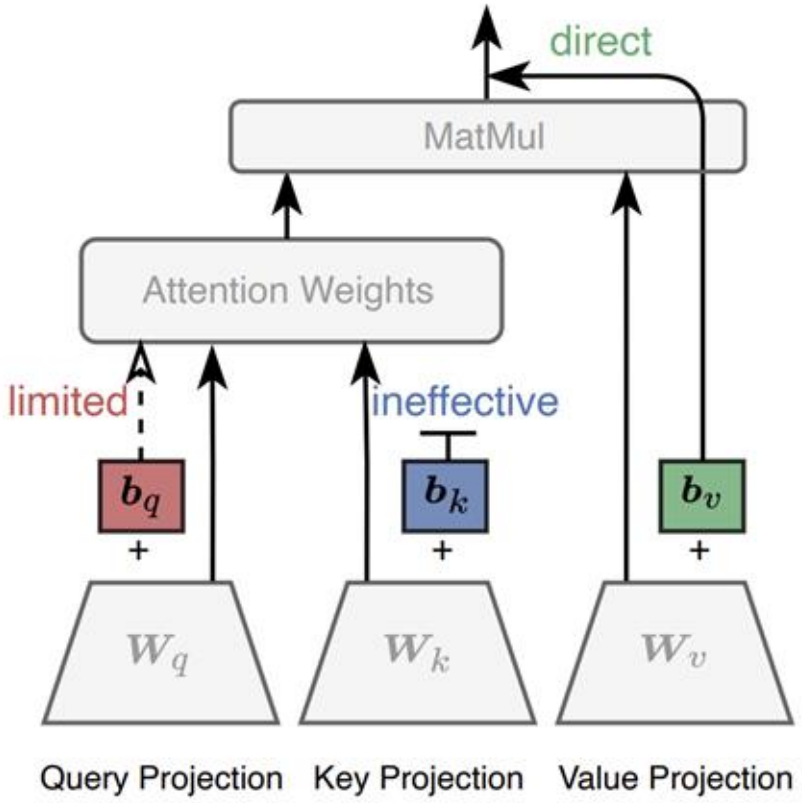
$$\text{Attention}(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) = \text{softmax}(\frac{\boldsymbol{Q}\boldsymbol{K}^T}{\sqrt{d_k}})\boldsymbol{V}$$

Expressive Power of Bias Terms

$$\text{Attention}(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) = \text{softmax}(\frac{\boldsymbol{Q}\boldsymbol{K}^T}{\sqrt{d_k}})\boldsymbol{V}$$

refer to our paper for theoretical analysis

b_k: **Softmax is shift-invariant**



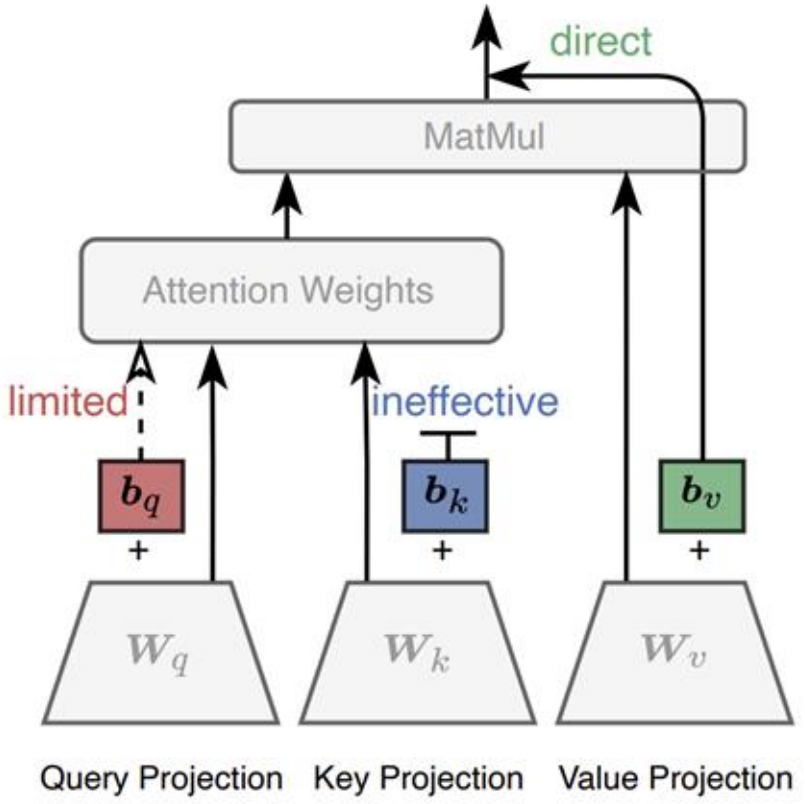
Expressive Power of Bias Terms

$$\text{Attention}(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) = \text{softmax}(\frac{\boldsymbol{Q}\boldsymbol{K}^T}{\sqrt{d_k}})\boldsymbol{V}$$

refer to our paper for theoretical analysis

b_k: **Softmax is shift-invariant**

b_q: **Softmax is insensitive where the Jacobian approaching zero**



Expressive Power of Bias Terms

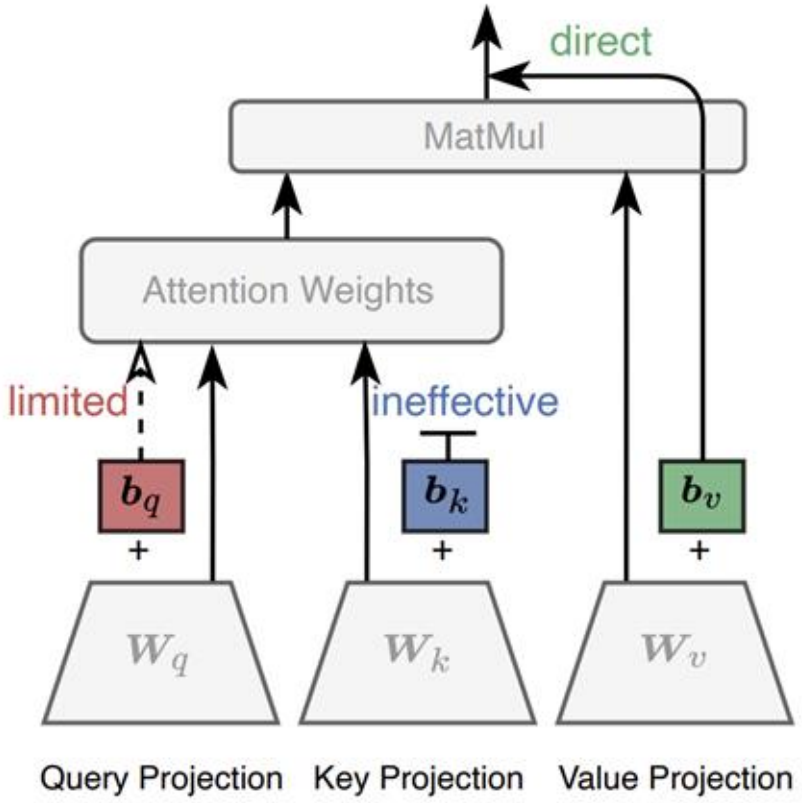
$$\text{Attention}(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) = \text{softmax}(\frac{\boldsymbol{Q}\boldsymbol{K}^T}{\sqrt{d_k}})\boldsymbol{V}$$

refer to our paper for theoretical analysis

b_k: **Softmax is shift-invariant**

b_q: **Softmax is insensitive where the Jacobian approaching zero**

b_v: **Without restriction by the softmax**



Bias-Only Fine-Tuning

Zaken, Elad Ben, et al. Bitfit: Simple parameter-efficient fine-tuning for transformer-based masked language-models. ACL, 2022

Bias-Only Fine-Tuning

Fine-tuning Different Bias

Downstream Performance

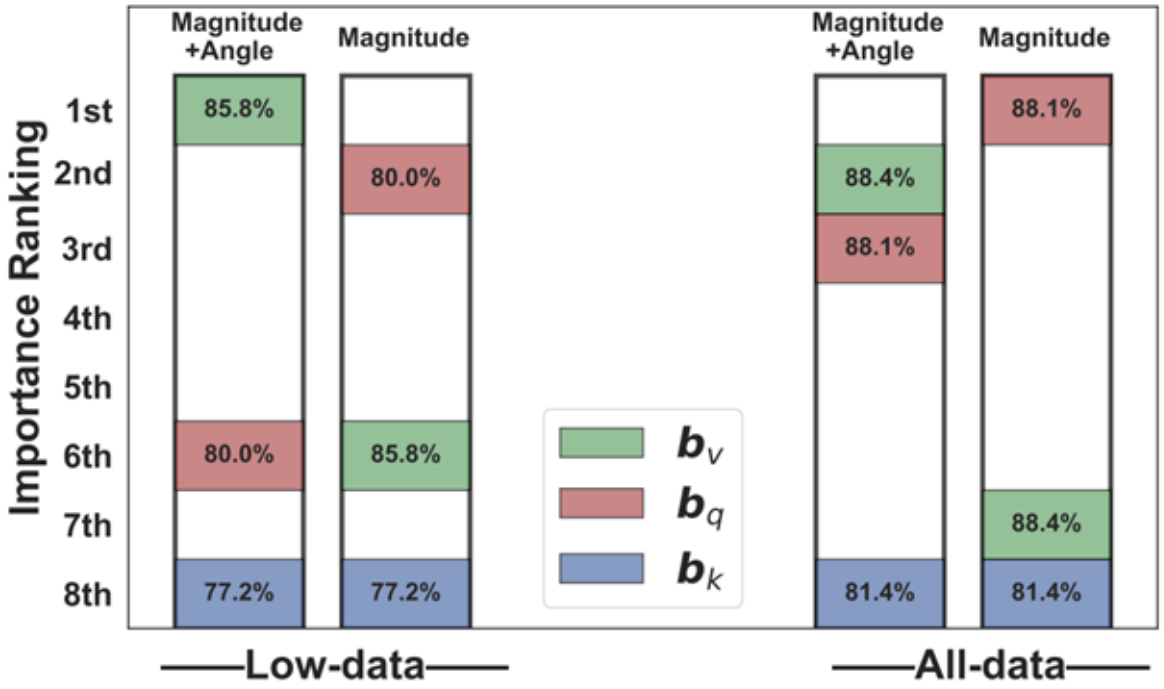
Zaken, Elad Ben, et al. Bitfit: Simple parameter-efficient fine-tuning for transformer-based masked language-models. ACL, 2022

Bias-Only Fine-Tuning

Fine-tuning Different Bias

↓ ?

Downstream Performance



(BERT$_{BASE}$ on SST-2)

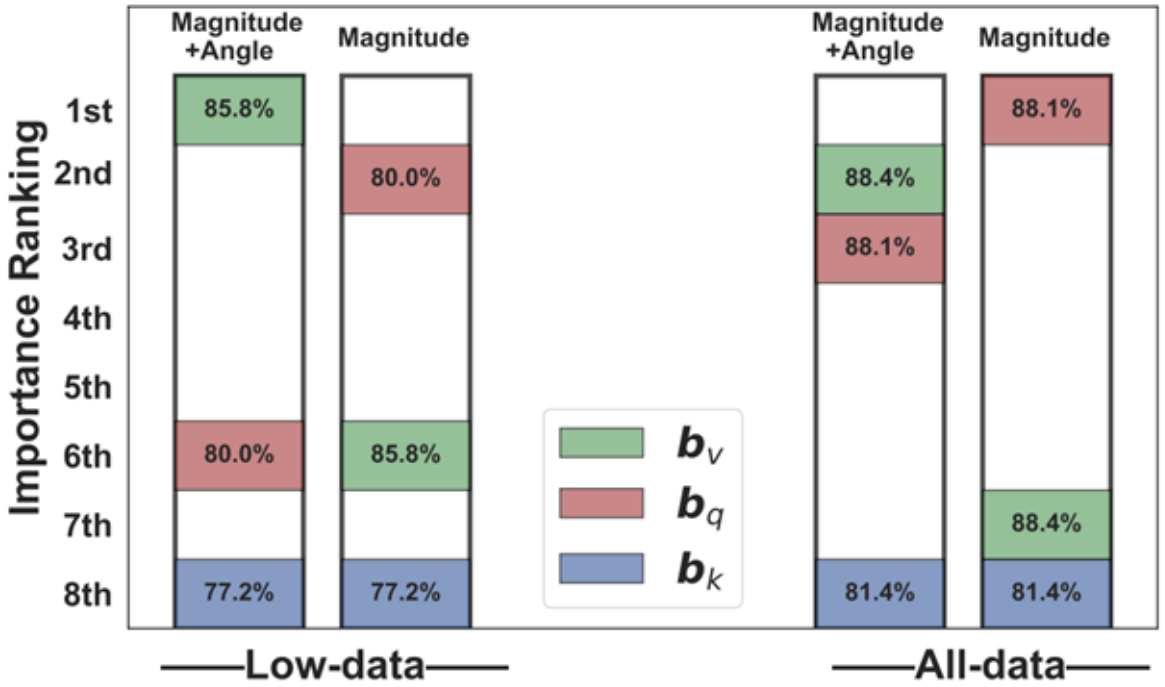
Zaken, Elad Ben, et al. Bitfit: Simple parameter-efficient fine-tuning for transformer-based masked language-models. ACL, 2022

Bias-Only Fine-Tuning

Fine-tuning Different Bias

↓ ?

Downstream Performance



(BERT$_{BASE}$ on SST-2)

Magnitude:
The change in b_q is more than b_v, **but** the accuracy of b_q is lower than b_v.

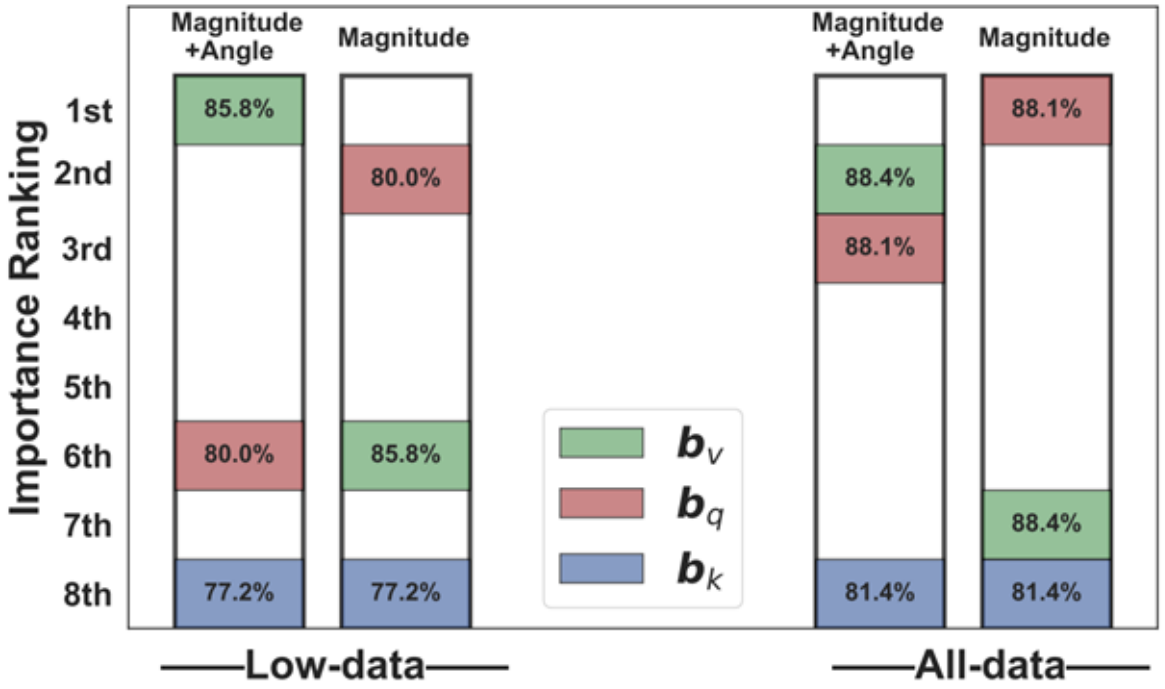
Zaken, Elad Ben, et al. Bitfit: Simple parameter-efficient fine-tuning for transformer-based masked language-models. ACL, 2022

Bias-Only Fine-Tuning

Fine-tuning Different Bias

↓ ?

Downstream Performance



(BERT$_{BASE}$ on SST-2)

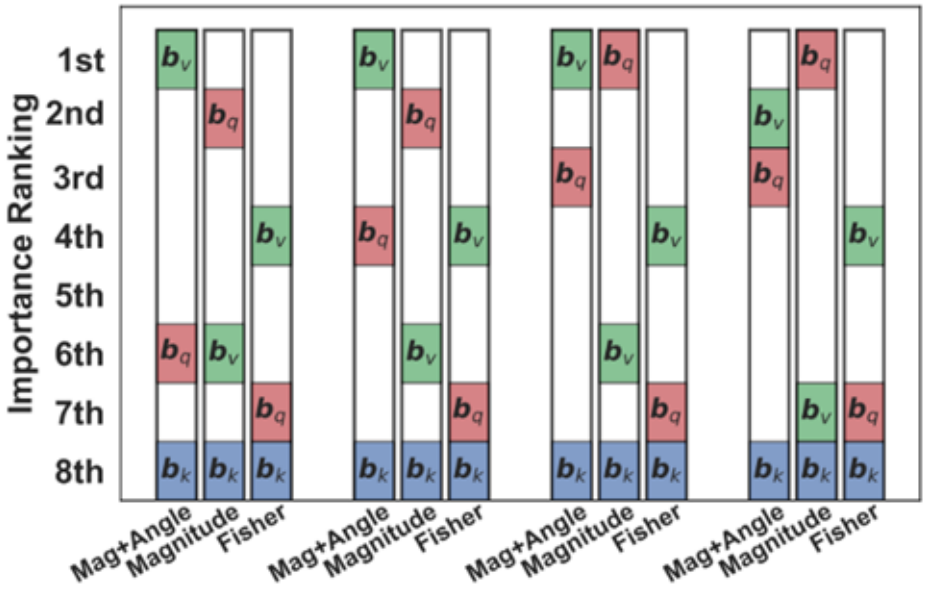
Magnitude:
The change in b_q is more than b_v, **but** the accuracy of b_q is lower than b_v.

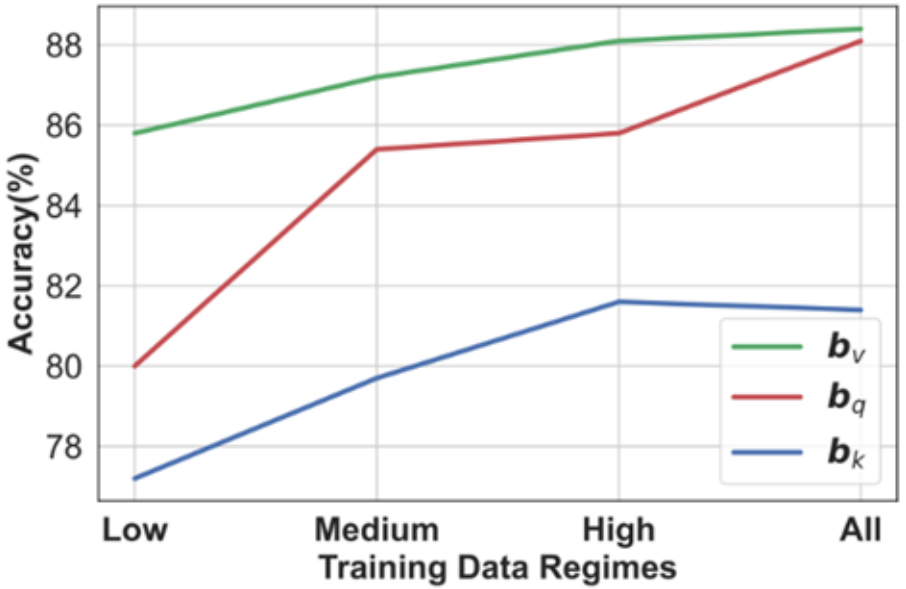
Magnitude+Angle:
The change in b_v is more than b_q, and the accuracy of b_v is higher than b_q.

Zaken, Elad Ben, et al. Bitfit: Simple parameter-efficient fine-tuning for transformer-based masked language-models. ACL, 2022

Empirical Evaluation

Importance Ranking & Downstream Performance

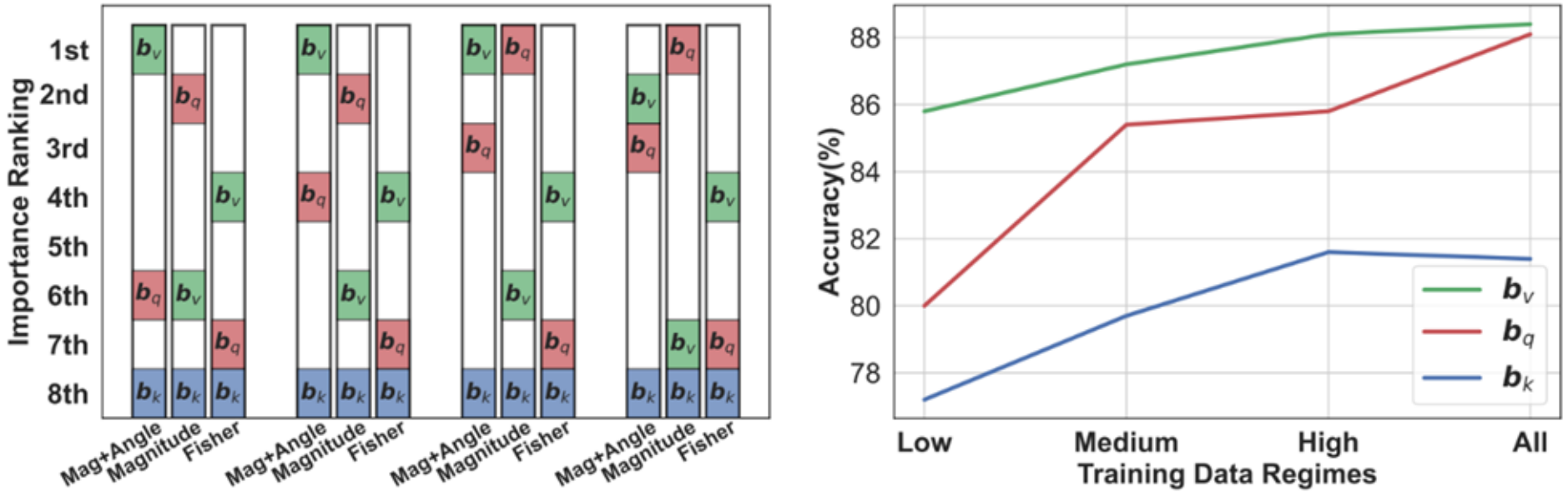




(BERT$_{BASE}$ on SST-2)

refer to our paper for more datasets

Importance Ranking & Downstream Performance



(BERT$_{BASE}$ on SST-2)

refer to our paper for more datasets

Mag+Angle shows a precise and dynamic link between bias-term rankings and downstream performance across diverse data regimes.

Key Finding

Bias	RTE	QQP	QNLI	$MNLI_m$	$MNLI_{mm}$	CoLA	MRPC	STSB
b_v	**59.5%**	**68.8%**	**73.8%**	**43.8%**	**45.7%**	**43.2%**	**84.0%**	**75.6%**
b_q	46.9%	65.1%	67.7%	40.1%	40.5%	30.8%	81.1%	67.8%
b_k	46.5%	60.5%	66.9%	39.5%	40.1%	6.5%	79.2%	65.4%

($BERT_{BASE}$ in Low Data Regimes)

Key Finding

Bias	RTE	QQP	QNLI	$MNLI_m$	$MNLI_{mm}$	CoLA	MRPC	STSB
b_v	**59.5%**	**68.8%**	**73.8%**	**43.8%**	**45.7%**	**43.2%**	**84.0%**	**75.6%**
b_q	46.9%	65.1%	67.7%	40.1%	40.5%	30.8%	81.1%	67.8%
b_k	46.5%	60.5%	66.9%	39.5%	40.1%	6.5%	79.2%	65.4%

($BERT_{BASE}$ in Low Data Regimes)

Directly fine-tuning b_v in low-data regimes!

BEFT Can be Combined with Other PEFT Methods

Bias Term	BEFT	+LoRA	+VeRA	+DoRA
b_v	**85.8%**	**85.0%**	**82.2%**	**86.0%**
b_q	80.0%	82.6%	81.8%	83.5%
b_k	77.2%	76.9%	81.7%	76.9%

($BERT_{BASE}$ on SST-2)

BEFT Can be Combined with Other PEFT Methods

Bias Term	BEFT	+LoRA	+VeRA	+DoRA
b_v	**85.8%**	**85.0%**	**82.2%**	**86.0%**
b_q	80.0%	82.6%	81.8%	83.5%
b_k	77.2%	76.9%	81.7%	76.9%

($BERT_{BASE}$ on SST-2)

Directly fine-tuning b_v+PEFT in low-data regimes!

SOTA LLMs Tend to be Bias-Free

Extension to Bias-Free LLMs

```
from torch import nn
class AddbiasLinear(nn.Linear):
    def __init__(self, in_features, out_features):
        nn.Linear.__init__(self, in_features, out_features)
        # Add trainable bias term
        self.add_bias = nn.Parameter(self.weight.new_zeros((out_features)))
        # Freezing the pre-trained weight matrix
        self.weight.requires_grad = False

    def forward(self, x):
        result = nn.functional.linear(x, self.weight) + self.add_bias
        return result
```

Extension to Bias-Free LLMs

Bias Term	Adding b_v	Adding b_q	Adding b_k
LLaMA2-7B	**94.9%**	90.0%	68.0%
DeepSeek-Coder-Base-1.3B	**76.9%**	67.4%	60.3%
GPT-J-6B	**92.8%**	88.3%	63.8%

(SST-2)

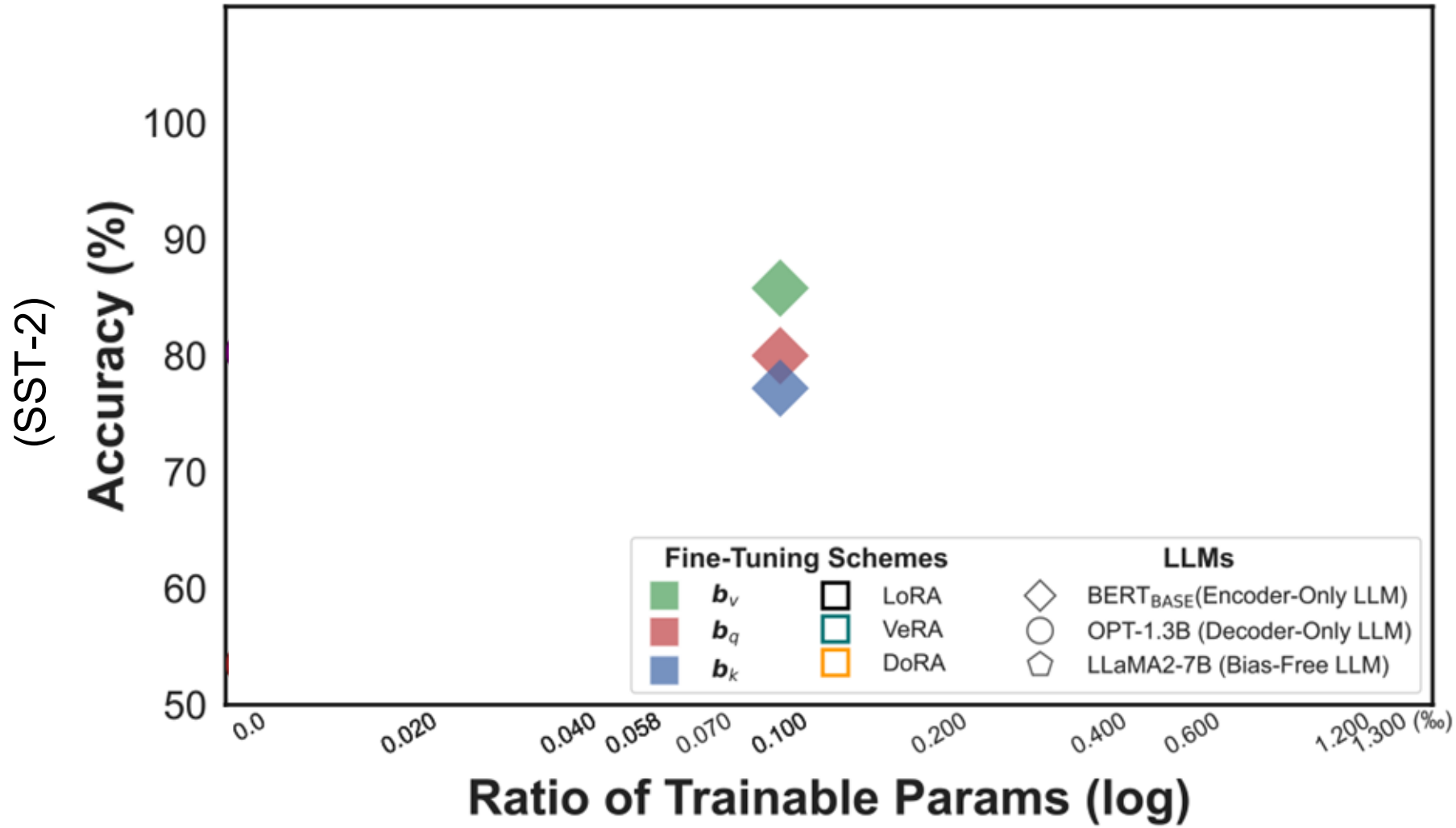
Extension to Bias-Free LLMs

Bias Term	Adding b_v	Adding b_q	Adding b_k
LLaMA2-7B	**94.9%**	90.0%	68.0%
DeepSeek-Coder-Base-1.3B	**76.9%**	67.4%	60.3%
GPT-J-6B	**92.8%**	88.3%	63.8%

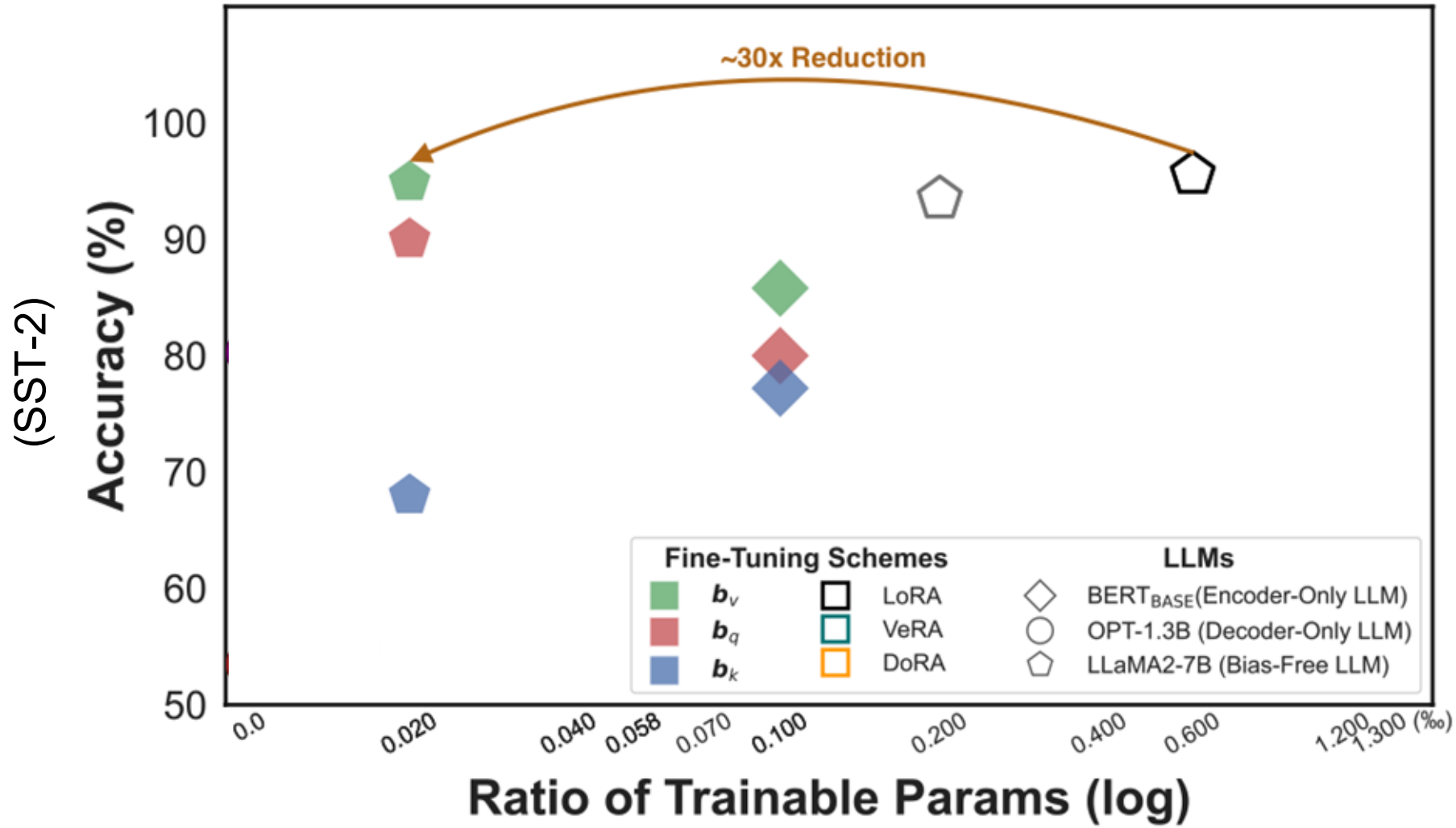
(SST-2)

Directly adding and fine-tuning b_v in low-data regimes!

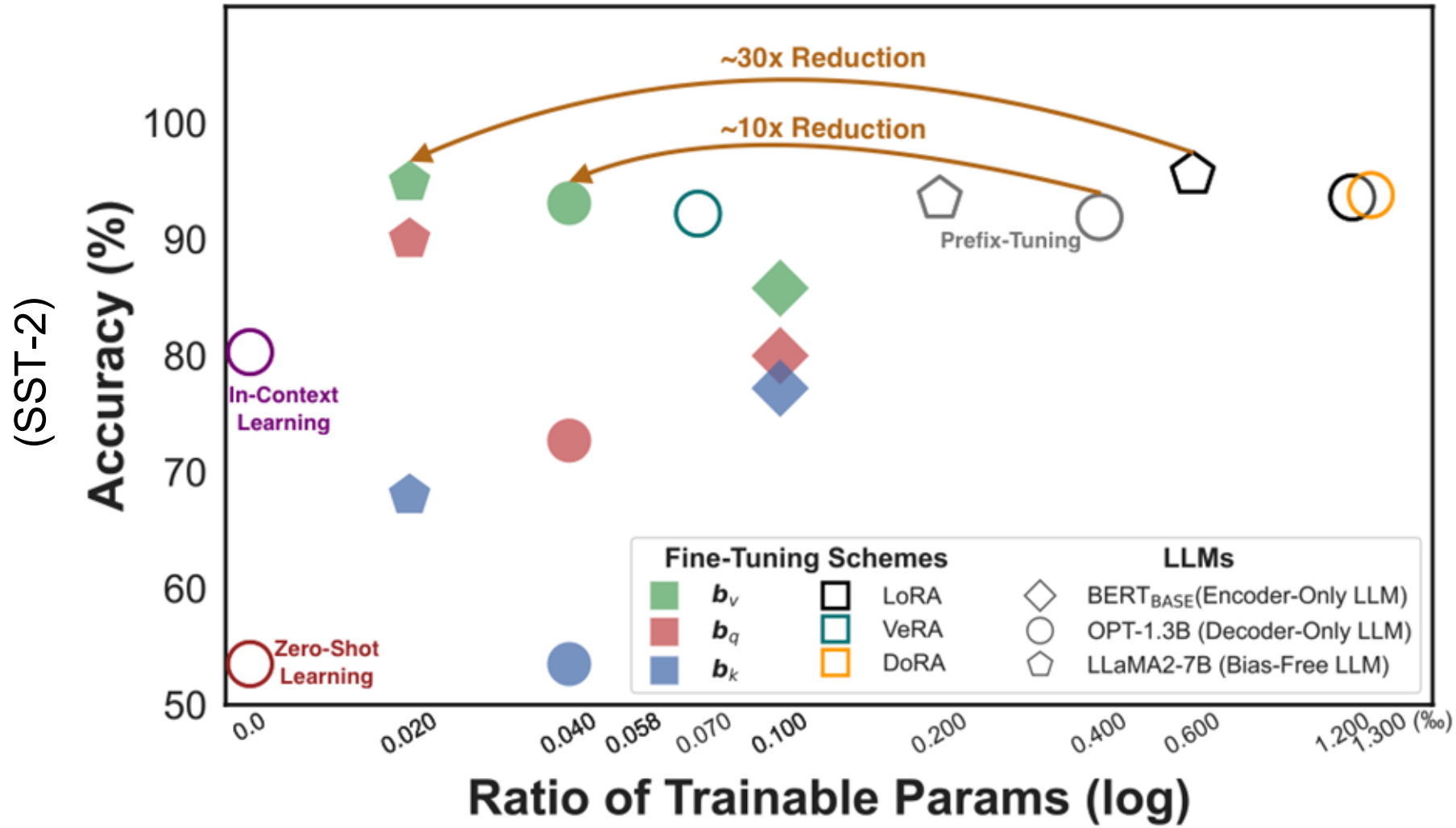
Directly Fine-Tuning b_v in Low-Data Regimes!



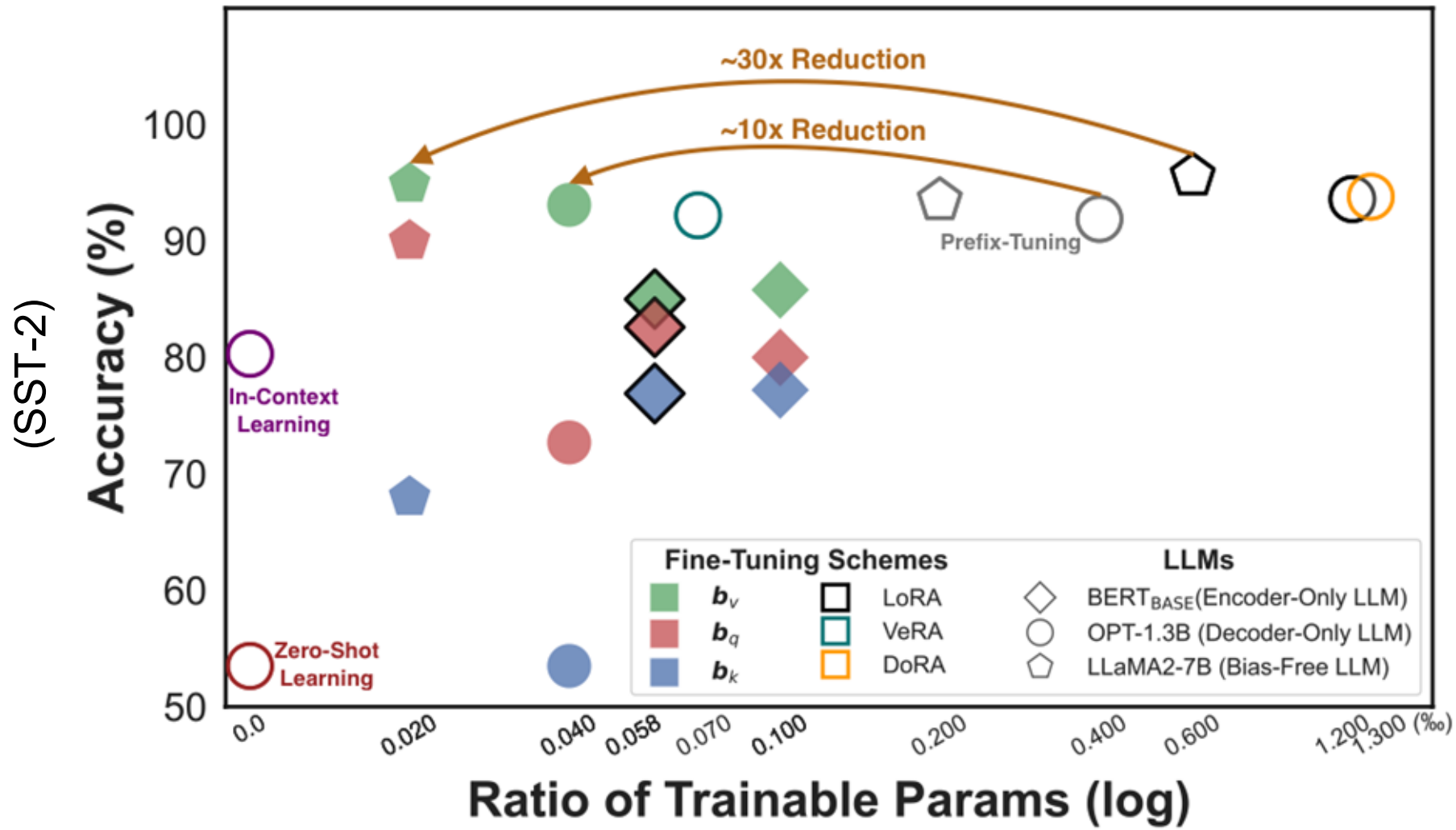
Directly Fine-Tuning b_v in Low-Data Regimes!



Directly Fine-Tuning b_* in Low-Data Regimes!



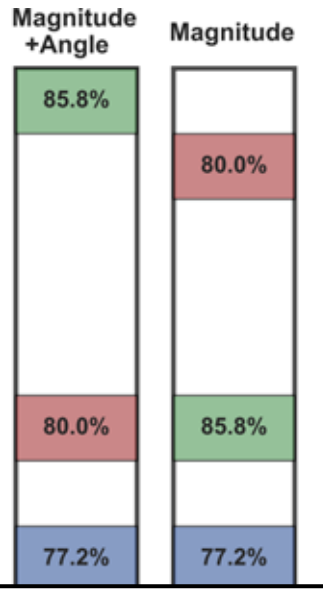
Directly Fine-Tuning b_v in Low-Data Regimes!



Conclusions & Questions

Challenge

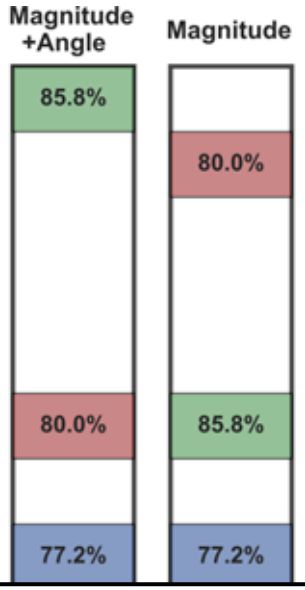
Unclear Bias Terms in Downstream Accuracy



Conclusions & Questions

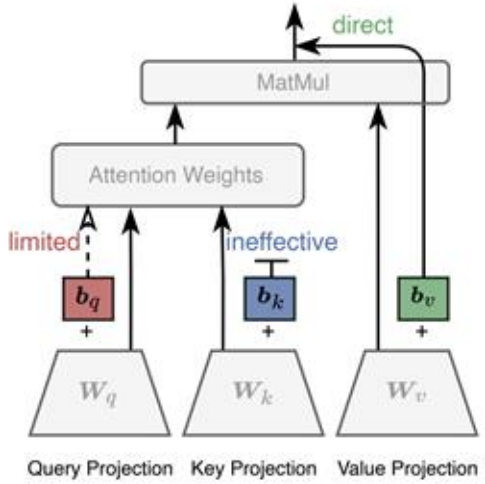
Challenge

Unclear Bias Terms in Downstream Accuracy



Findings

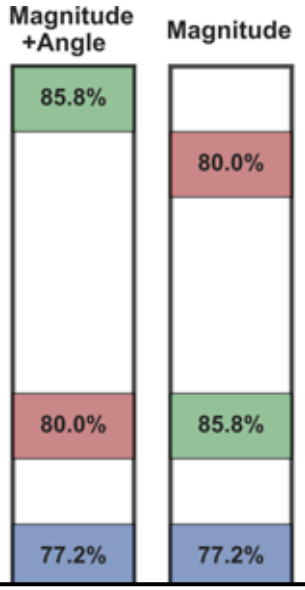
Directly fine-tuning b_v in low-data regimes!



Conclusions & Questions

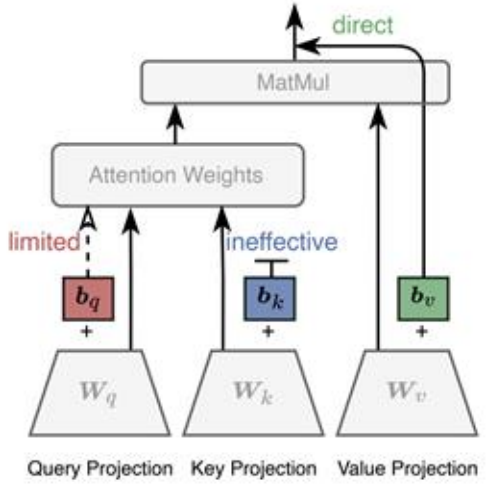
Challenge

Unclear Bias Terms in Downstream Accuracy



Findings

Directly fine-tuning b_v in low-data regimes!



Try BEFT in🤗 !

[huggingface/peft/beft]