Revisiting Few-sample BERT Fine-tuning

Tianyi Zhang*

Felix Wu*

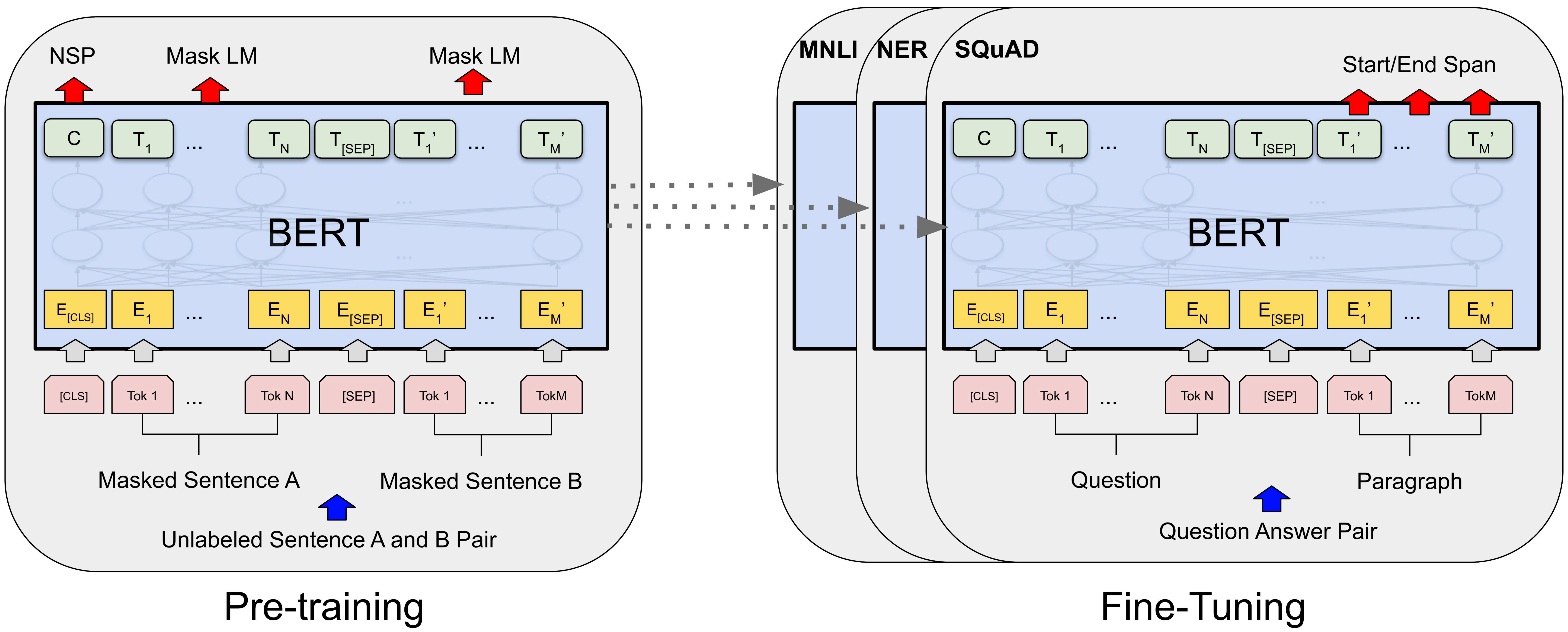
Arzoo Katiyar

Kilian Q. Weinberger

Yoav Artzi

*: Equal contribution

Fine-tuning BERT



ASAPP

Fine-tuning BERT on Small Datasets is Unstable

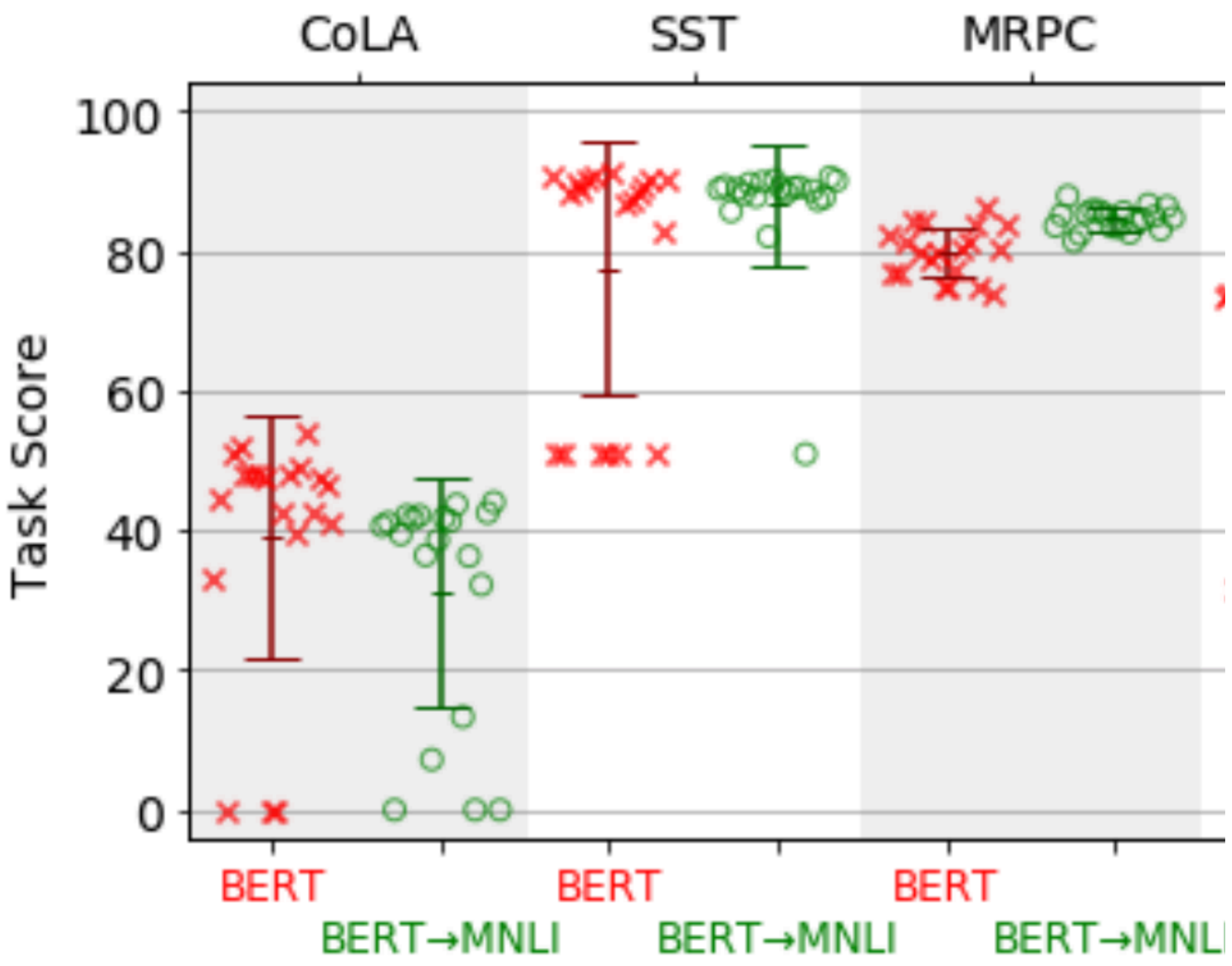


Figure from Phang et al, 2018

Why is fine-tuning so unstable? How can we fix it?

Analyzing from Three Aspects of Optimization

- Optimization Algorithm
- Initialization
- Number of iterations

Optimization Algorithm: Debiasing Omission in BERTAdam

Adam Algorithm (Kingma & Ba, 2014)

Optimization Algorithm: Debiasing Omission in BERTAdam

Adam Algorithm (Kingma & Ba, 2014)

In each iteration t, we have mini-batch $\mathbf{x}_t$

Gradients: $g_t \leftarrow \nabla_\theta L(\theta_{t-1}, \mathbf{x}_t)$

Optimization Algorithm: Debiasing Omission in BERTAdam

Adam Algorithm (Kingma & Ba, 2014)

In each iteration t, we have mini-batch $\mathbf{x}_t$

Gradients:	$g_t \leftarrow \nabla_\theta L(\theta_{t-1}, \mathbf{x}_t)$
First moment estimate:	$m_t \leftarrow \beta_1 m_{t-1} + (1 - \beta_1) g_t$
Second moment estimate:	$v_t \leftarrow \beta_2 v_{t-1} + (1 - \beta_2) g_t^2$

Optimization Algorithm: Debiasing Omission in BERTAdam

Adam Algorithm (Kingma & Ba, 2014)

Initialization: $m_0 \leftarrow 0$

$v_0 \leftarrow 0$

In each iteration t, we have mini-batch $\mathbf{x}_t$

Gradients: $g_t \leftarrow \nabla_\theta L(\theta_{t-1}, \mathbf{x}_t)$

First moment estimate: $m_t \leftarrow \beta_1 m_{t-1} + (1 - \beta_1) g_t$

Second moment estimate: $v_t \leftarrow \beta_2 v_{t-1} + (1 - \beta_2) g_t^2$

Optimization Algorithm: Debiasing Omission in BERTAdam

Adam Algorithm (Kingma & Ba, 2014)

Initialization: $m_0 \leftarrow 0$

$v_0 \leftarrow 0$

In each iteration t, we have mini-batch $\mathbf{x}_t$

Gradients: $g_t \leftarrow \nabla_\theta L(\theta_{t-1}, \mathbf{x}_t)$

First moment estimate: $m_t \leftarrow \beta_1 m_{t-1} + (1 - \beta_1) g_t$

Second moment estimate: $v_t \leftarrow \beta_2 v_{t-1} + (1 - \beta_2) g_t^2$

Bias-corrected first moment estimate: $\hat{m}_t \leftarrow \dfrac{m_t}{1 - \beta_1^t}$

Bias-corrected second moment estimate: $\hat{v}_t \leftarrow \dfrac{v_t}{1 - \beta_2^t}$

ASAPP

Optimization Algorithm: Debiasing Omission in BERTAdam

Adam Algorithm (Kingma & Ba, 2014)

Initialization: $m_0 \leftarrow 0$

$v_0 \leftarrow 0$

In each iteration t, we have mini-batch $\mathbf{x}_t$

Gradients:	$g_t \leftarrow \nabla_\theta L(\theta_{t-1}, \mathbf{x}_t)$
First moment estimate:	$m_t \leftarrow \beta_1 m_{t-1} + (1 - \beta_1) g_t$
Second moment estimate:	$v_t \leftarrow \beta_2 v_{t-1} + (1 - \beta_2) g_t^2$
Bias-corrected first moment estimate:	$\hat{m}_t \leftarrow \dfrac{m_t}{1 - \beta_1^t}$
Bias-corrected second moment estimate:	$\hat{v}_t \leftarrow \dfrac{v_t}{1 - \beta_2^t}$
Parameter update:	$\theta_t \leftarrow \theta_{t-1} - \alpha \cdot \dfrac{\hat{m}_t}{\sqrt{\hat{v}_t}}$

ASAPP

Optimization Algorithm: Debiasing Omission in BERTAdam

Adam Algorithm (Kingma & Ba, 2014)

Initialization: $m_0 \leftarrow 0$

$v_0 \leftarrow 0$

In each iteration t, we have mini-batch $\mathbf{x}_t$

Gradients: $g_t \leftarrow \nabla_\theta L(\theta_{t-1}, \mathbf{x}_t)$

First moment estimate: $m_t \leftarrow \beta_1 m_{t-1} + (1-\beta_1) g_t$

Second moment estimate: $v_t \leftarrow \beta_2 v_{t-1} + (1-\beta_2) g_t^2$

~~Bias-corrected first moment estimate:~~ $\cancel{\hat{m}_t \leftarrow \frac{m_t}{1-\beta_1^t}}$

~~Bias-corrected second moment estimate:~~ $\cancel{\hat{v}_t \leftarrow \frac{v_t}{1-\beta_2^t}}$

Parameter update: $\theta_t \leftarrow \theta_{t-1} - \alpha \cdot \frac{\hat{m}_t}{\sqrt{\hat{v}_t}}$

BERTAdam omits bias-correction

Parameter update: $\theta_t \leftarrow \theta_{t-1} - \alpha \cdot \frac{m_t}{\sqrt{v_t}}$

ASAPP

Optimization Algorithm: Debiasing Omission in BERTAdam

Bias ratio:

$$\frac{\text{BERTAdam update magnitude}}{\text{Adam update magnitude}} = \frac{\dfrac{m_t}{\sqrt{v_t}}}{\dfrac{\hat{m}_t}{\sqrt{\hat{v}_t}}} = \frac{1-\beta_1^t}{\sqrt{1-\beta_2^t}} = \frac{1-0.9^t}{\sqrt{1-0.999^t}}$$

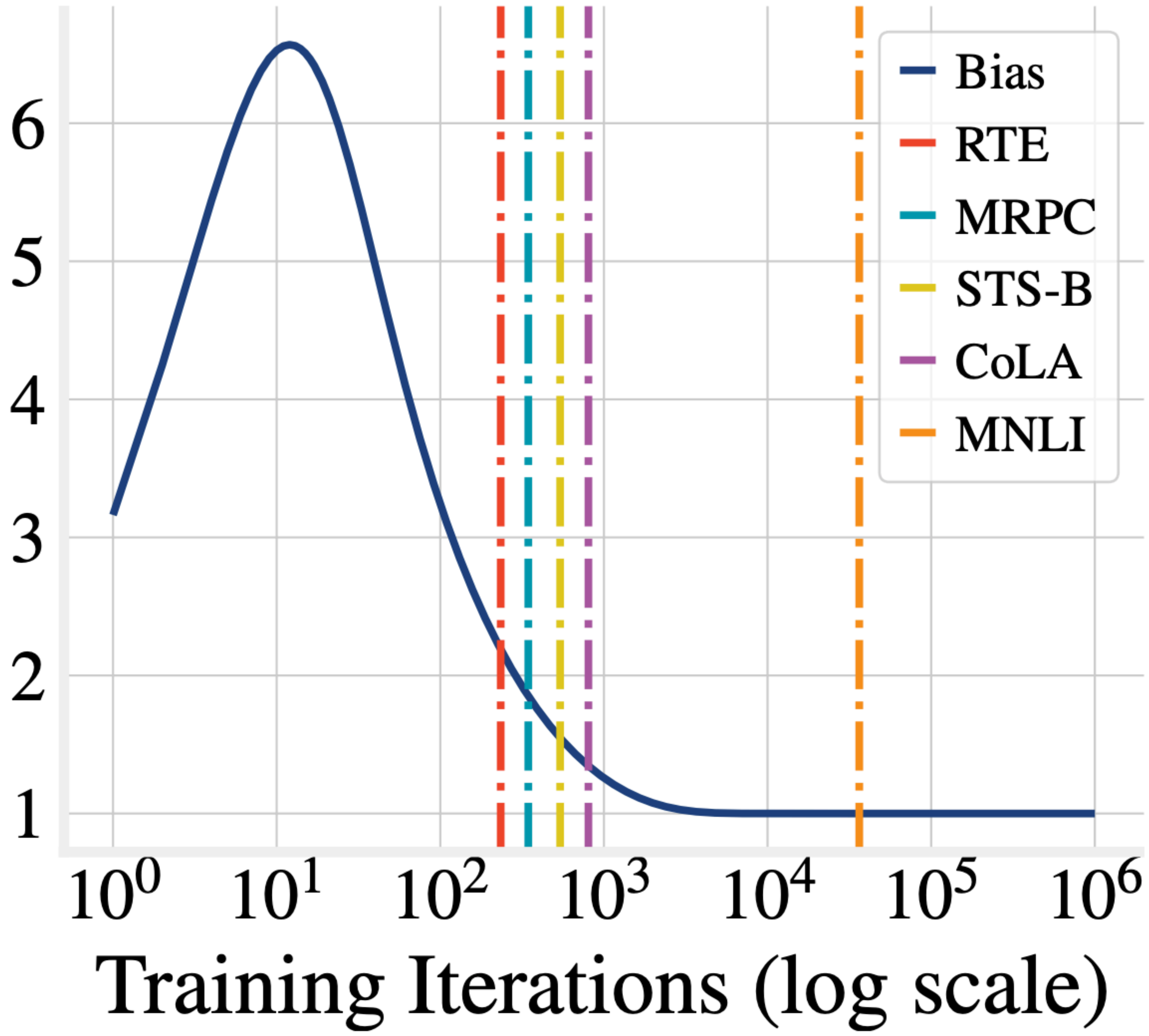
ASAPP

Optimization Algorithm: Debiasing Omission in BERTAdam

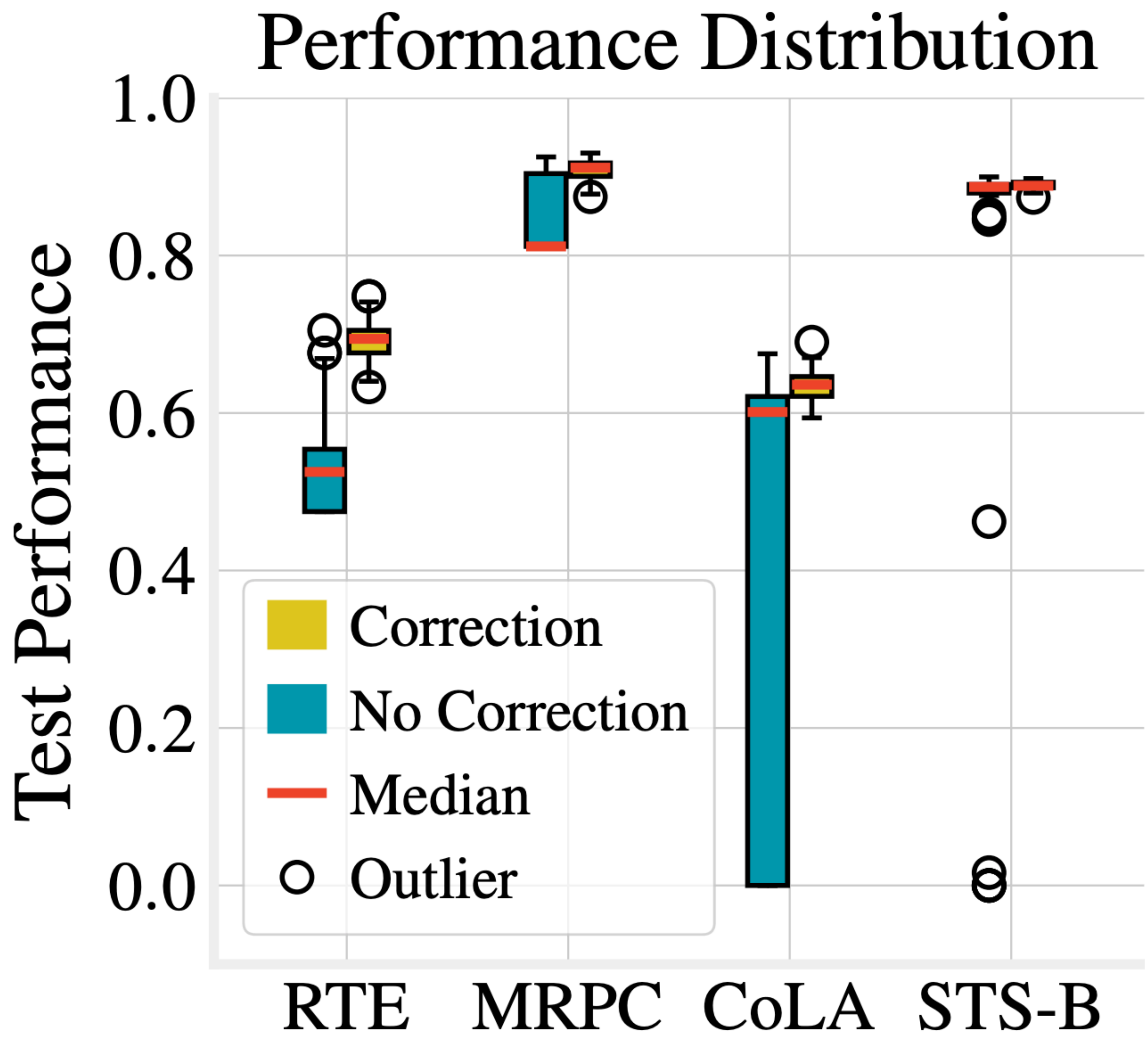
Bias ratio:

$$\frac{\text{BERTAdam update magnitude}}{\text{Adam update magnitude}}$$

Small datasets



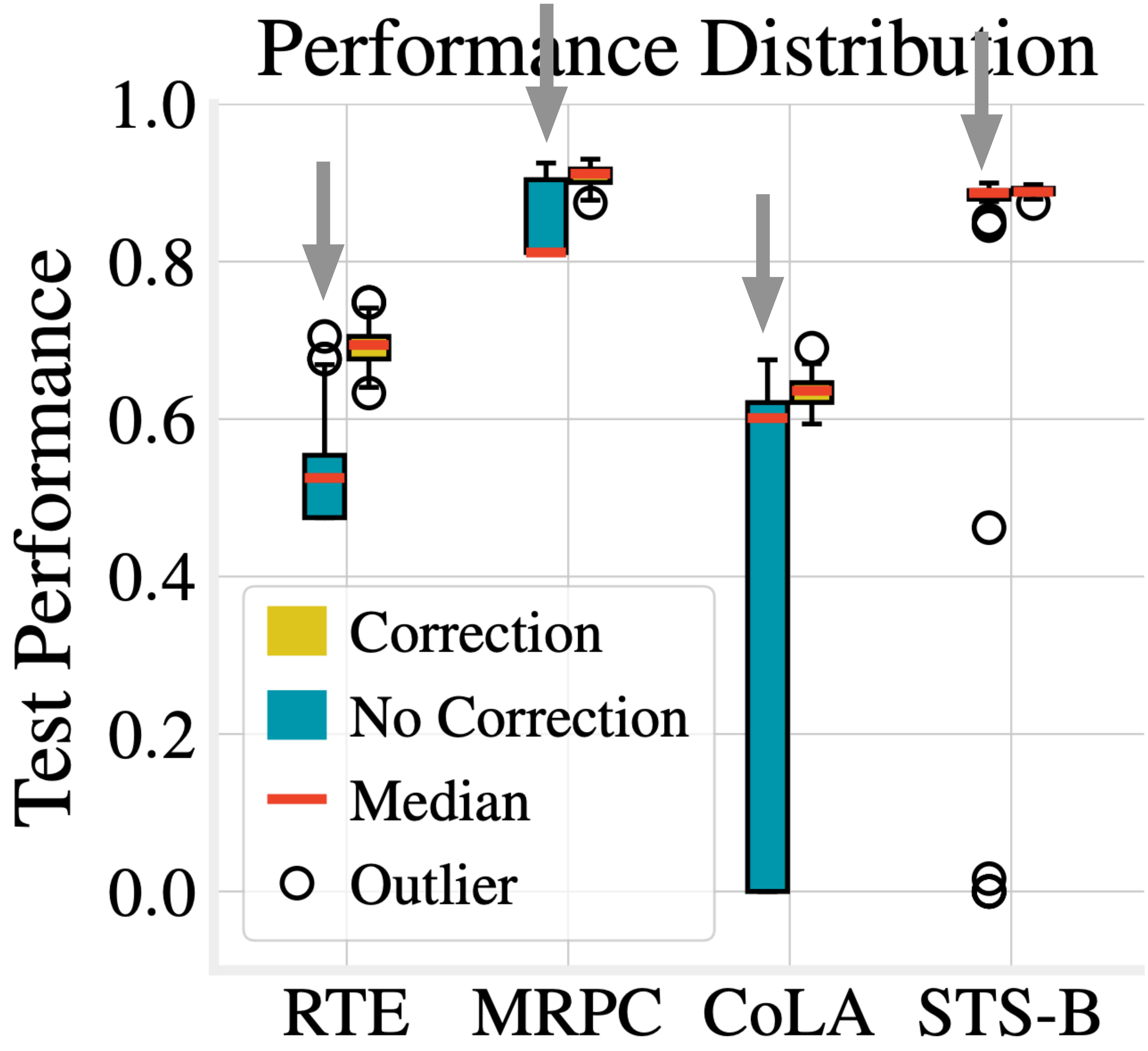
Optimization Algorithm: Debiasing Omission in BERTAdam



Use the standard Adam (with bias-correction)!

ASAPP

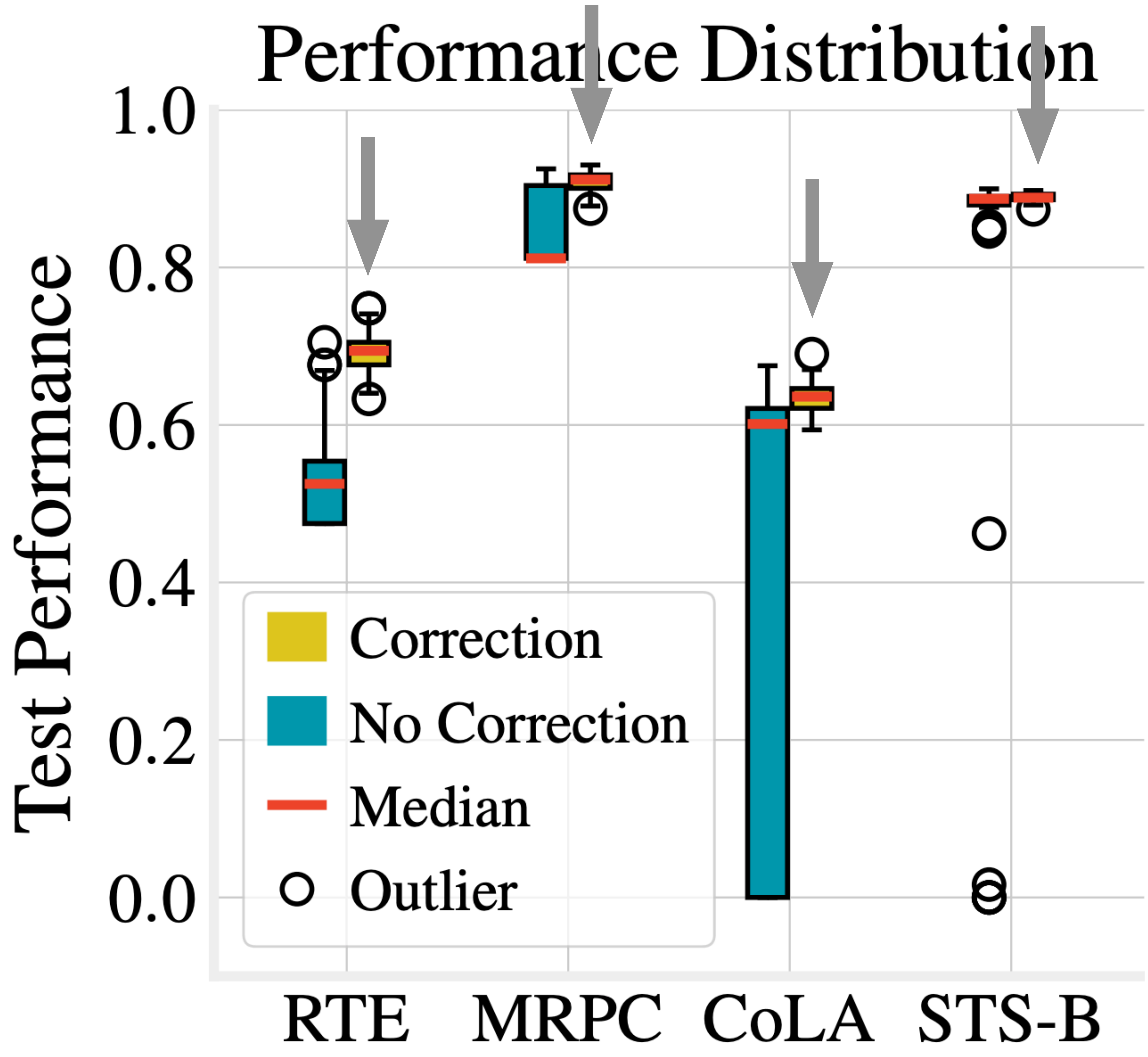
Optimization Algorithm: Debiasing Omission in BERTAdam



Use the standard Adam (with bias-correction)!

ASAPP

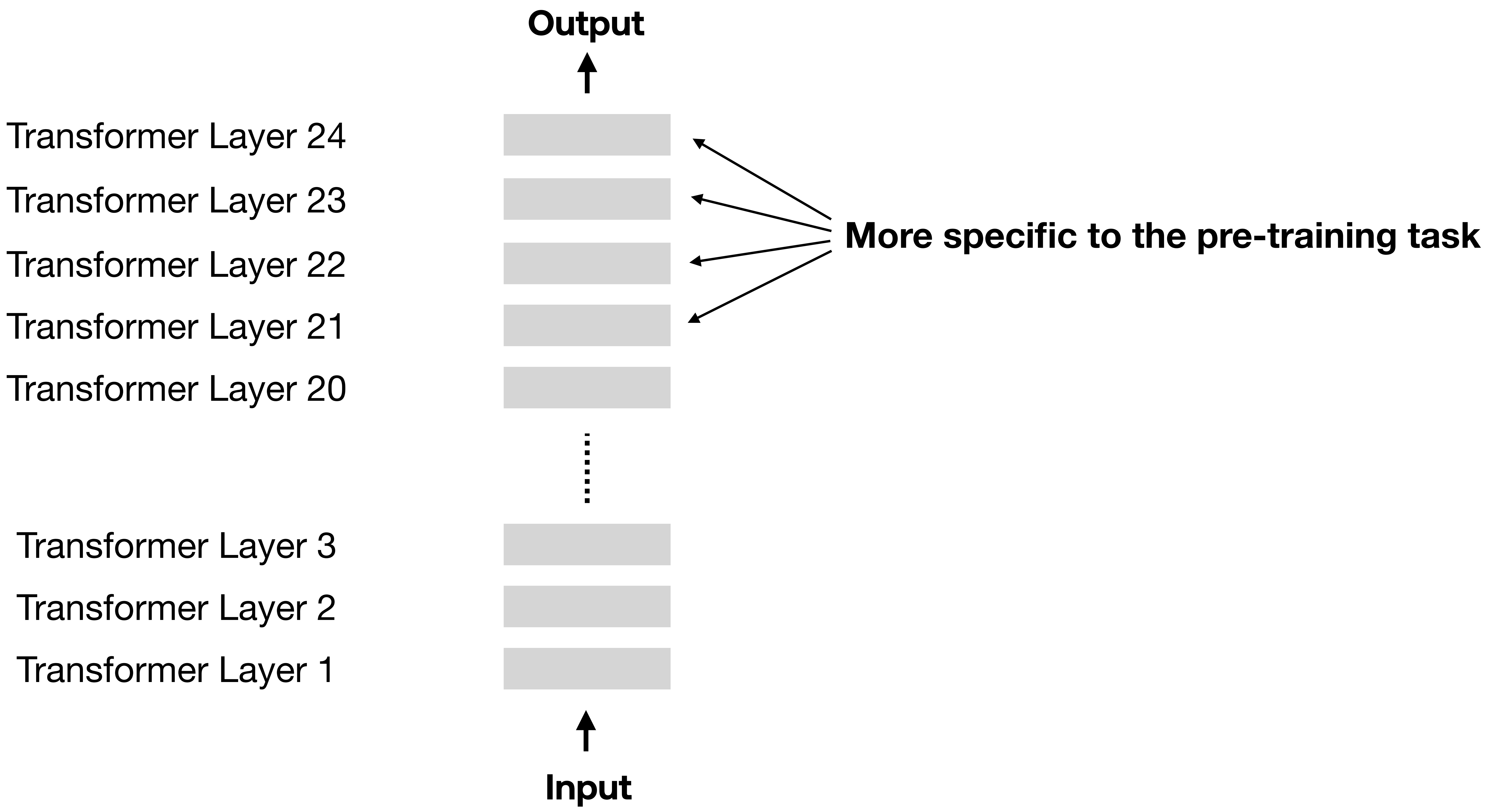
Optimization Algorithm: Debiasing Omission in BERTAdam



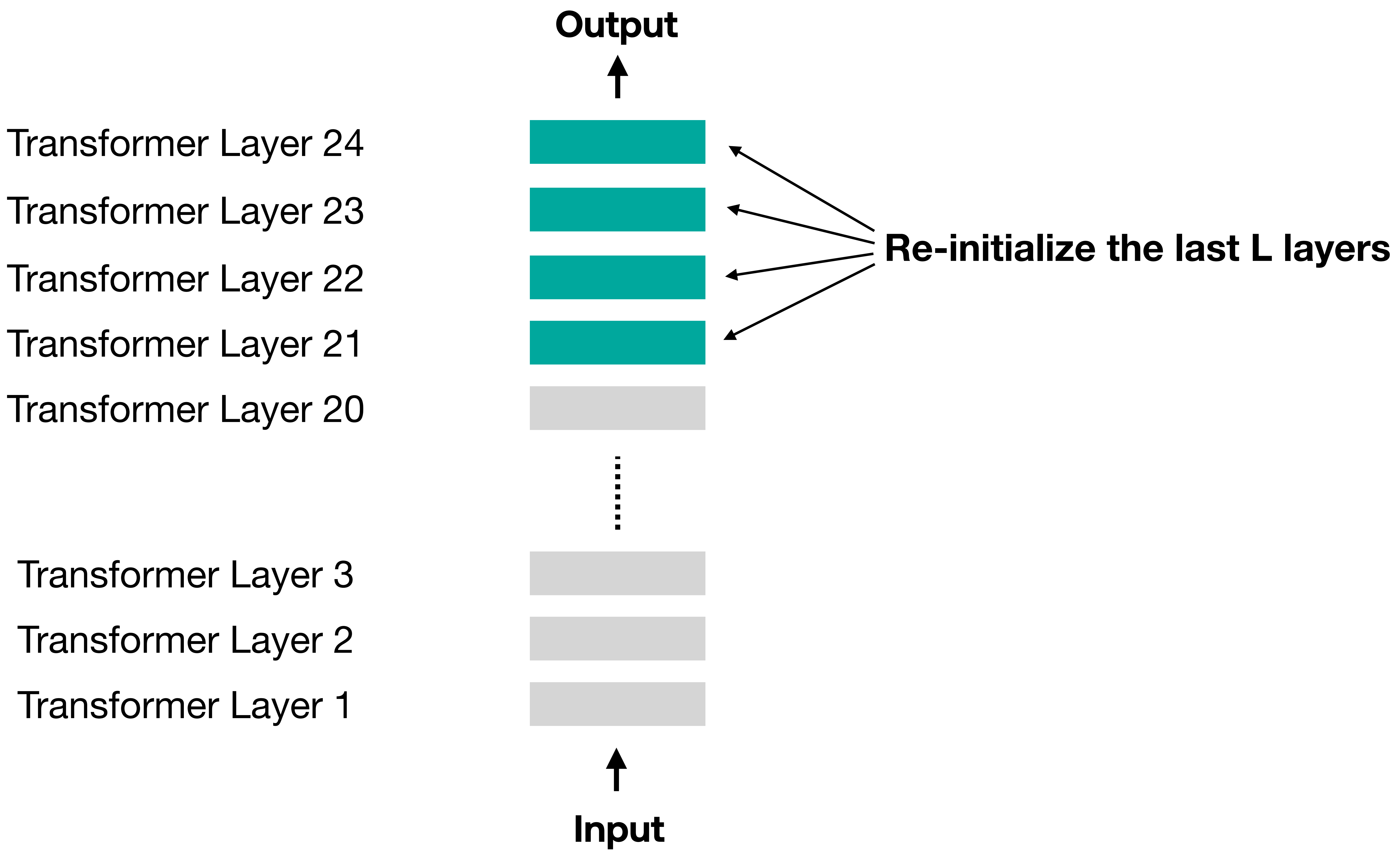
Use the standard Adam (with bias-correction)!

ASAPP

Initialization: Re-initializing BERT Pre-trained Layers

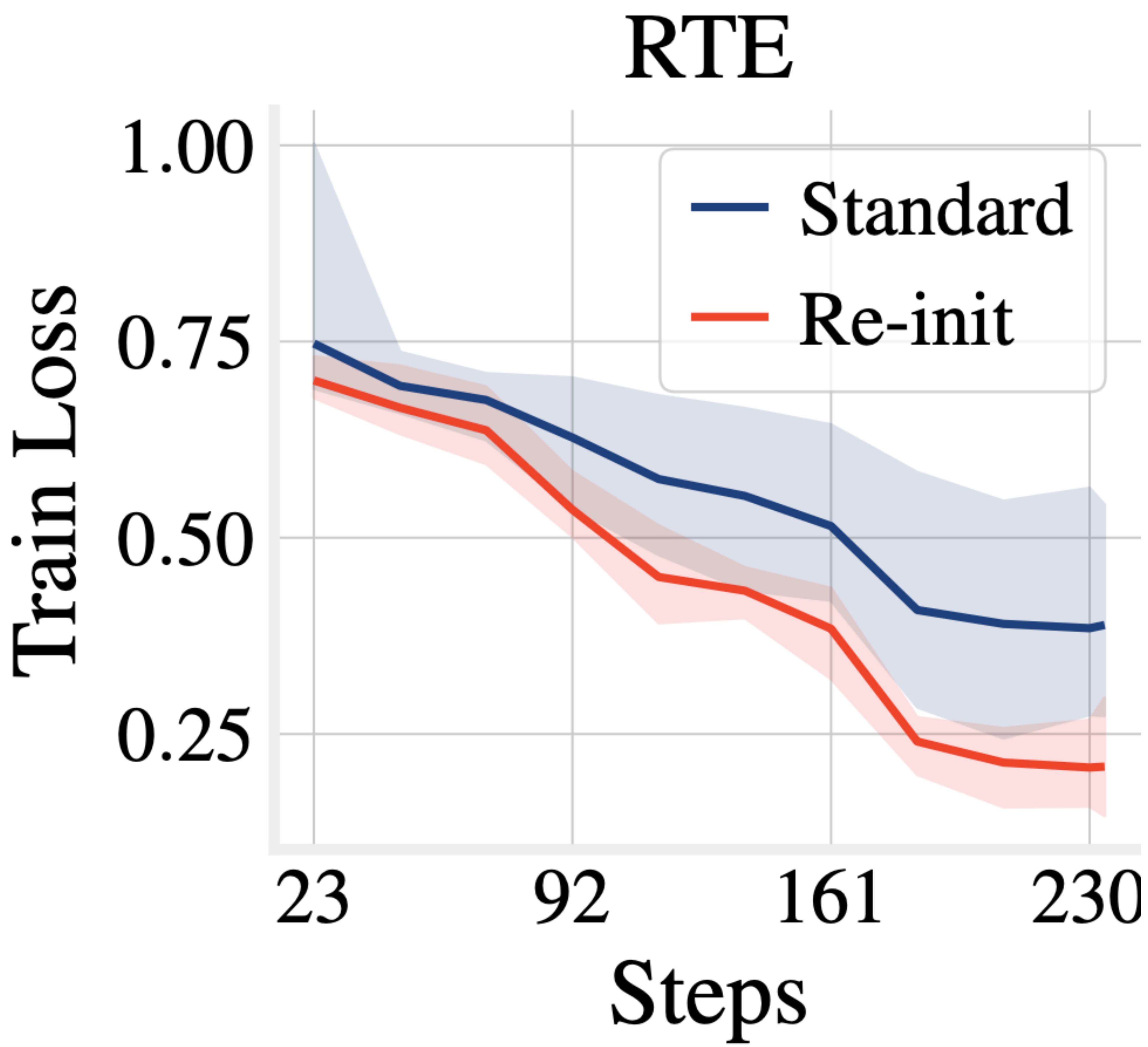


Initialization: Re-initializing BERT Pre-trained Layers

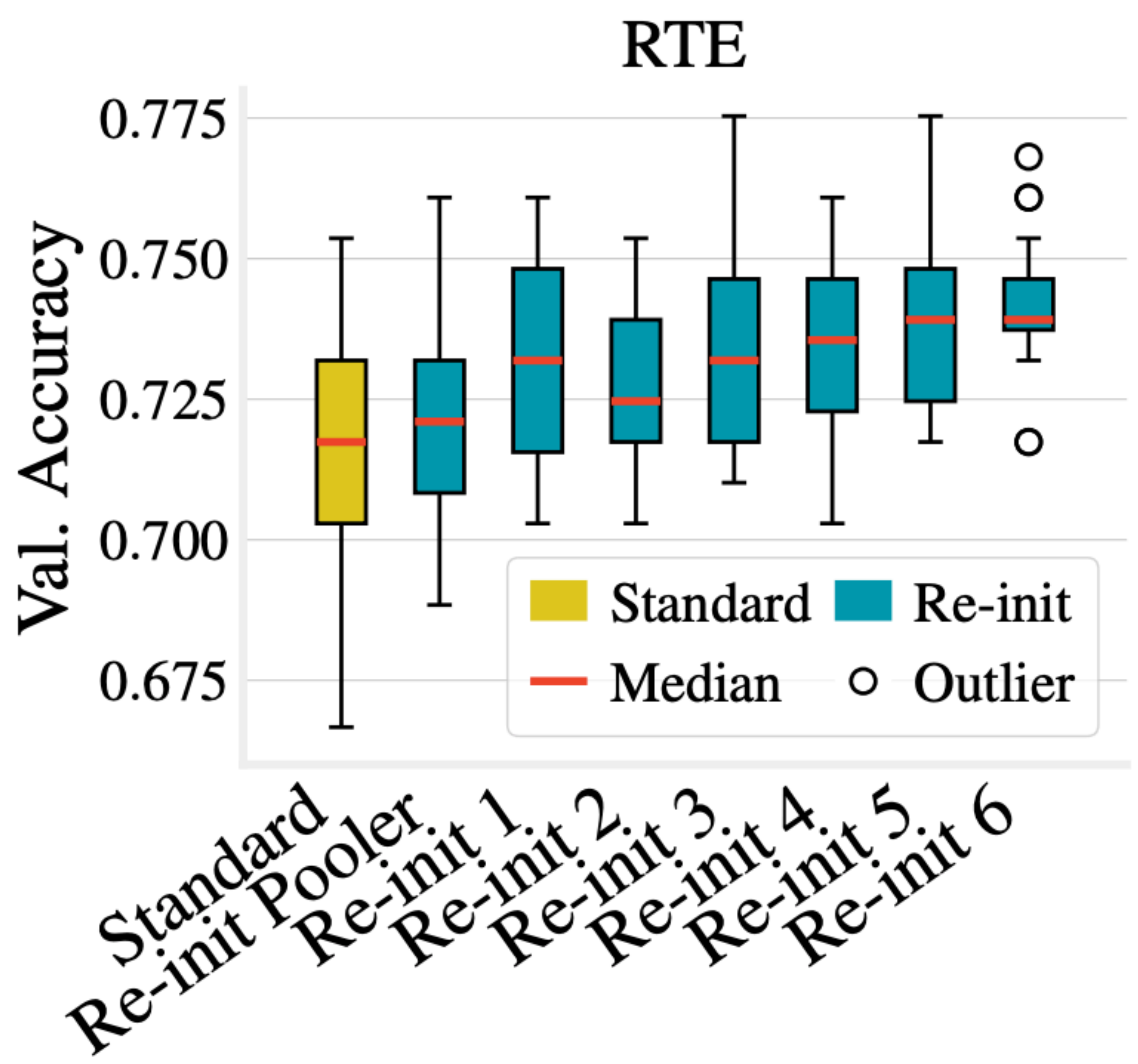


ASAPP

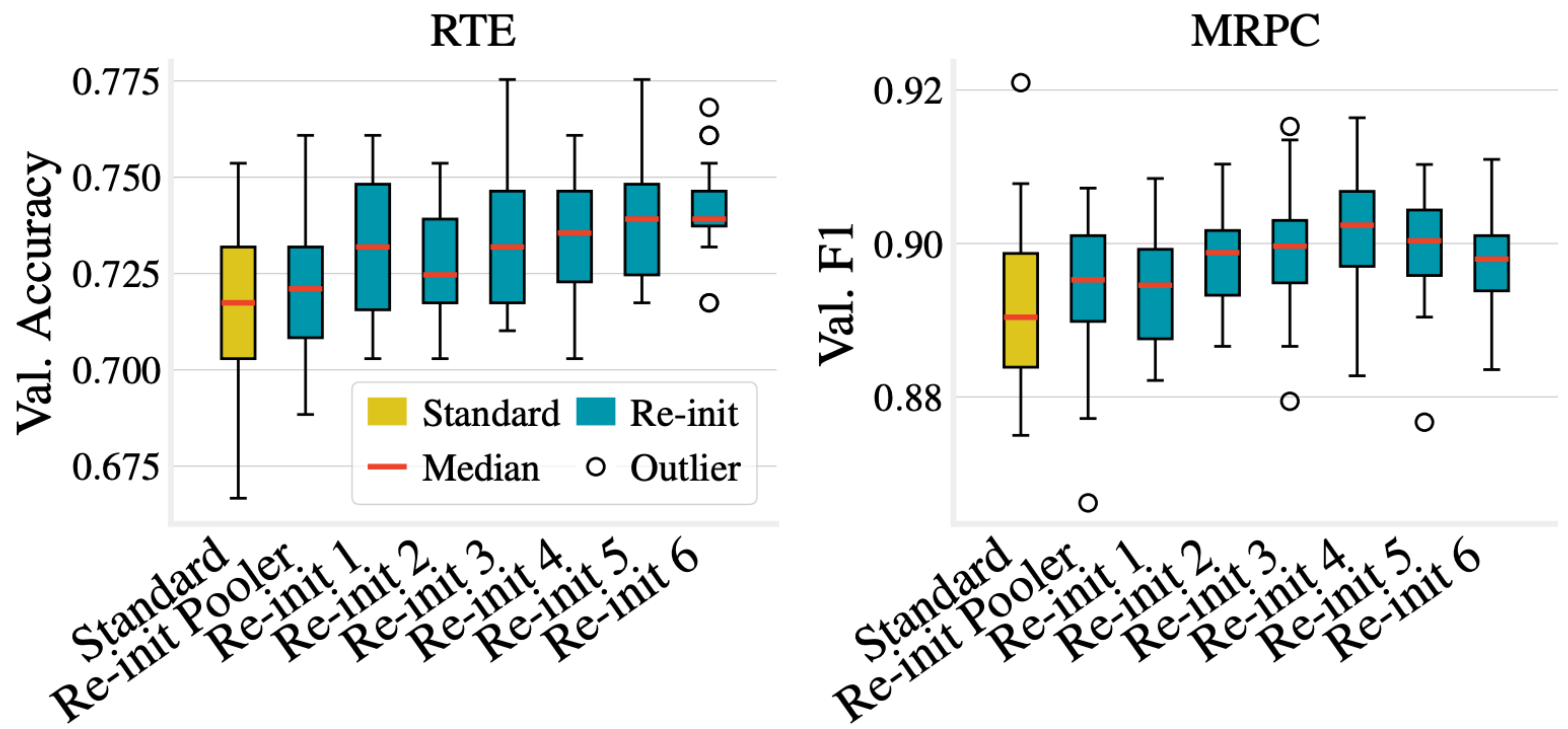
Initialization: Re-initializing BERT Pre-trained Layers



Initialization: Re-initializing BERT Pre-trained Layers



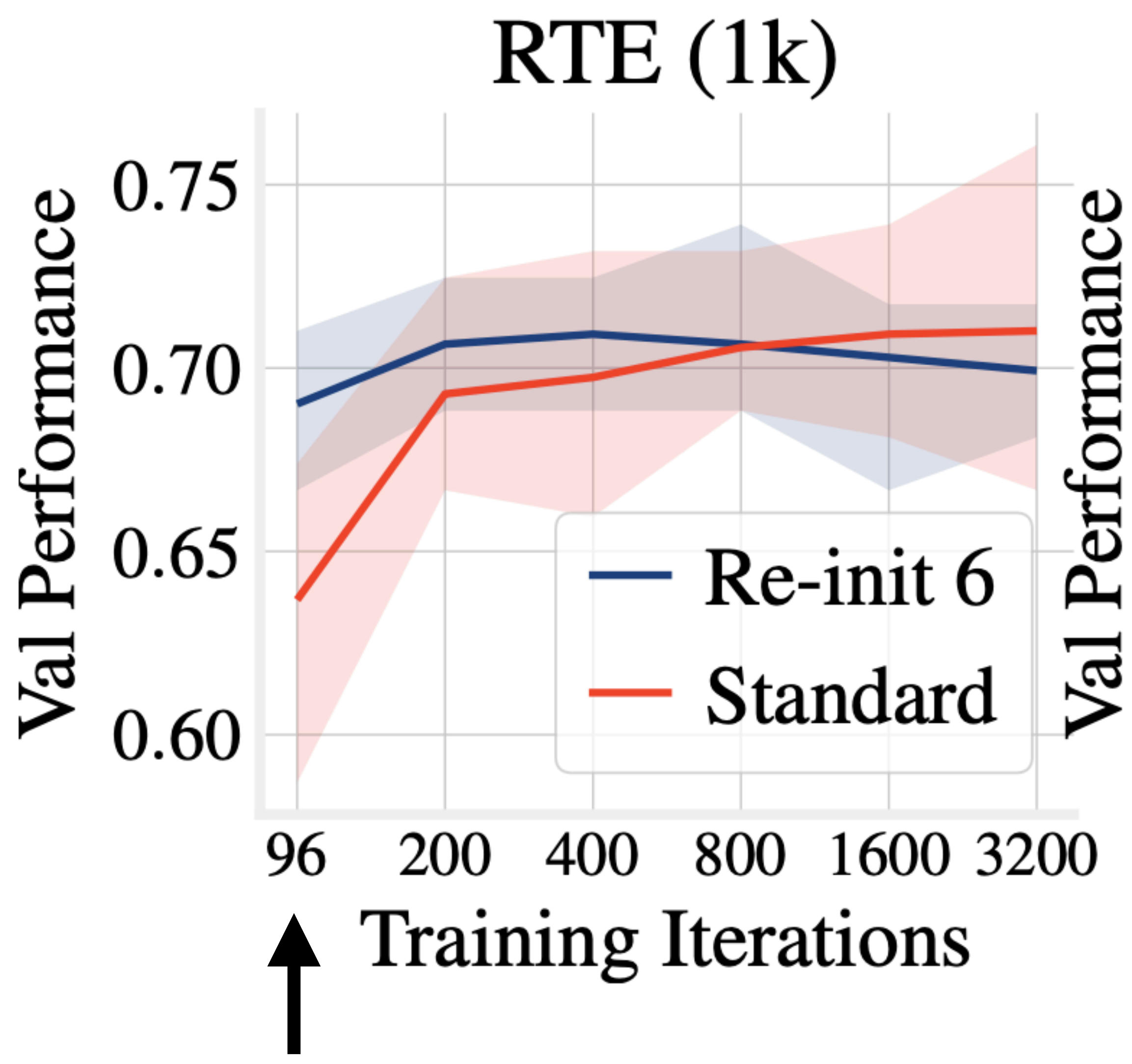
Initialization: Re-initializing BERT Pre-trained Layers



ASAPP

Number of Iterations: Fine-tuning BERT for Longer

The conventional 3-epoch recipe is insufficient!

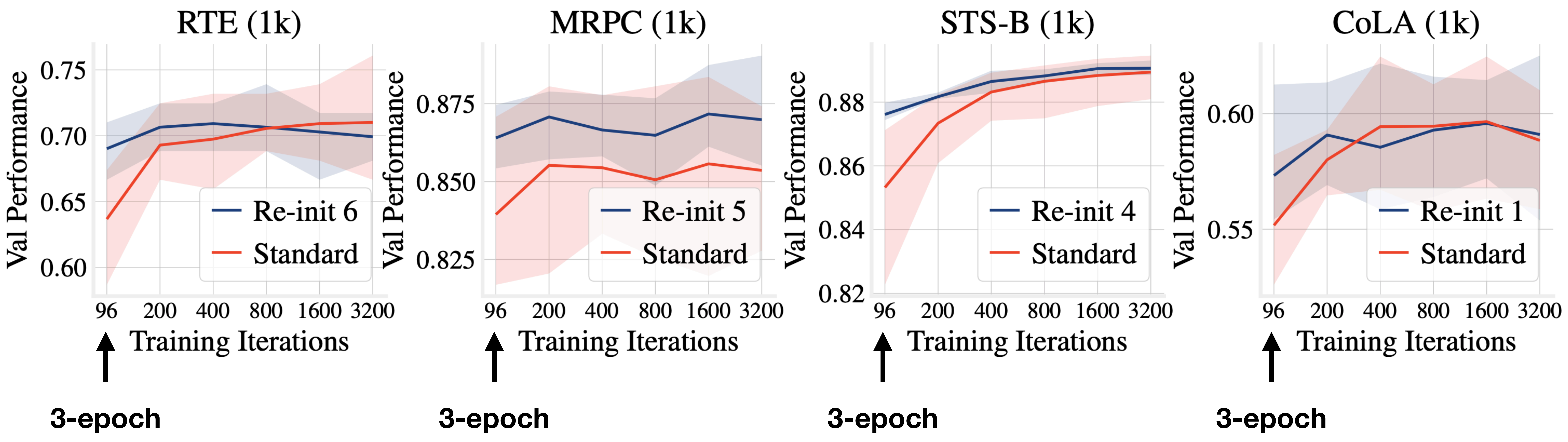


3-epoch

ASAPP

Number of Iterations: Fine-tuning BERT for Longer

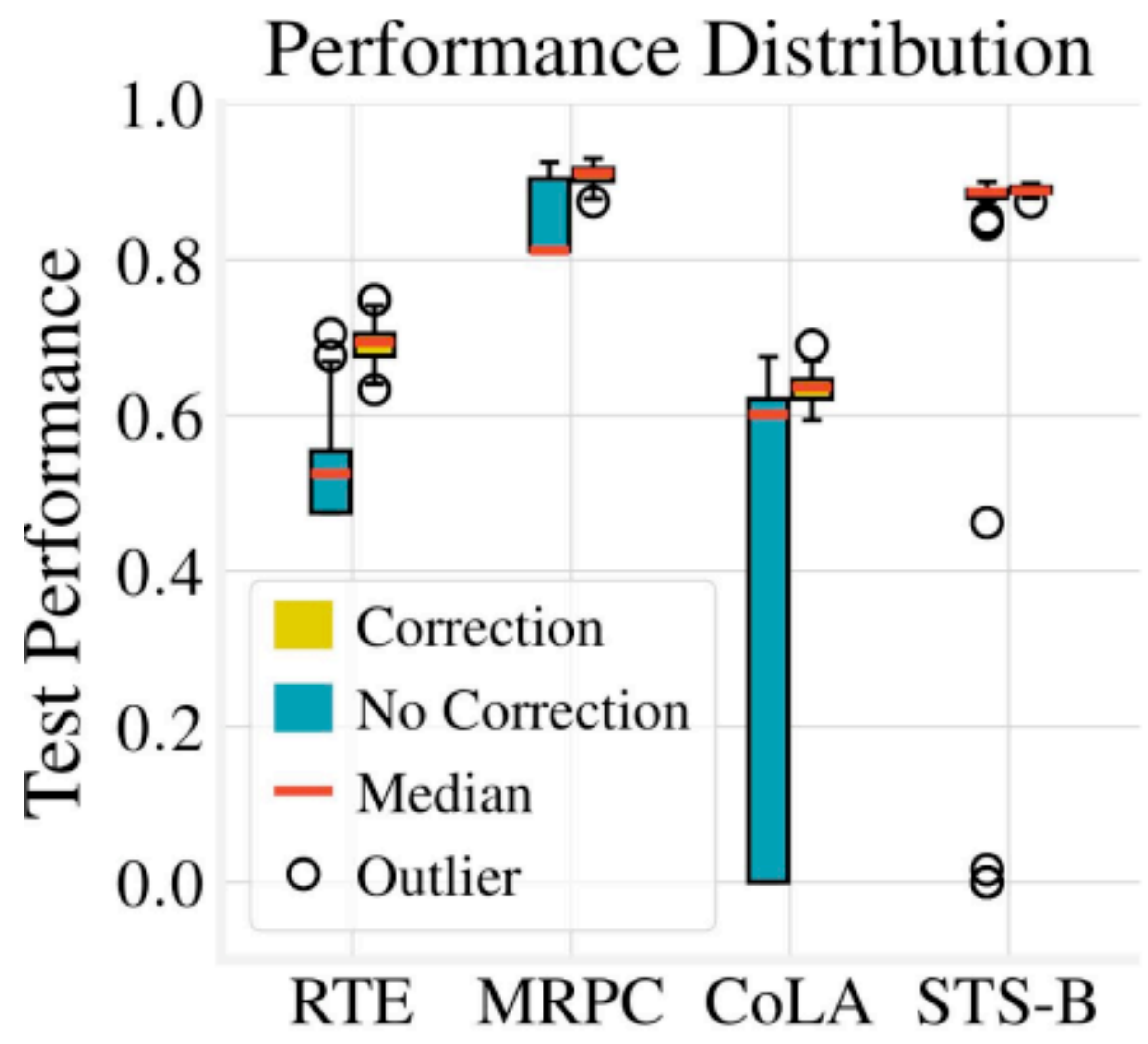
The conventional 3-epoch recipe is insufficient!



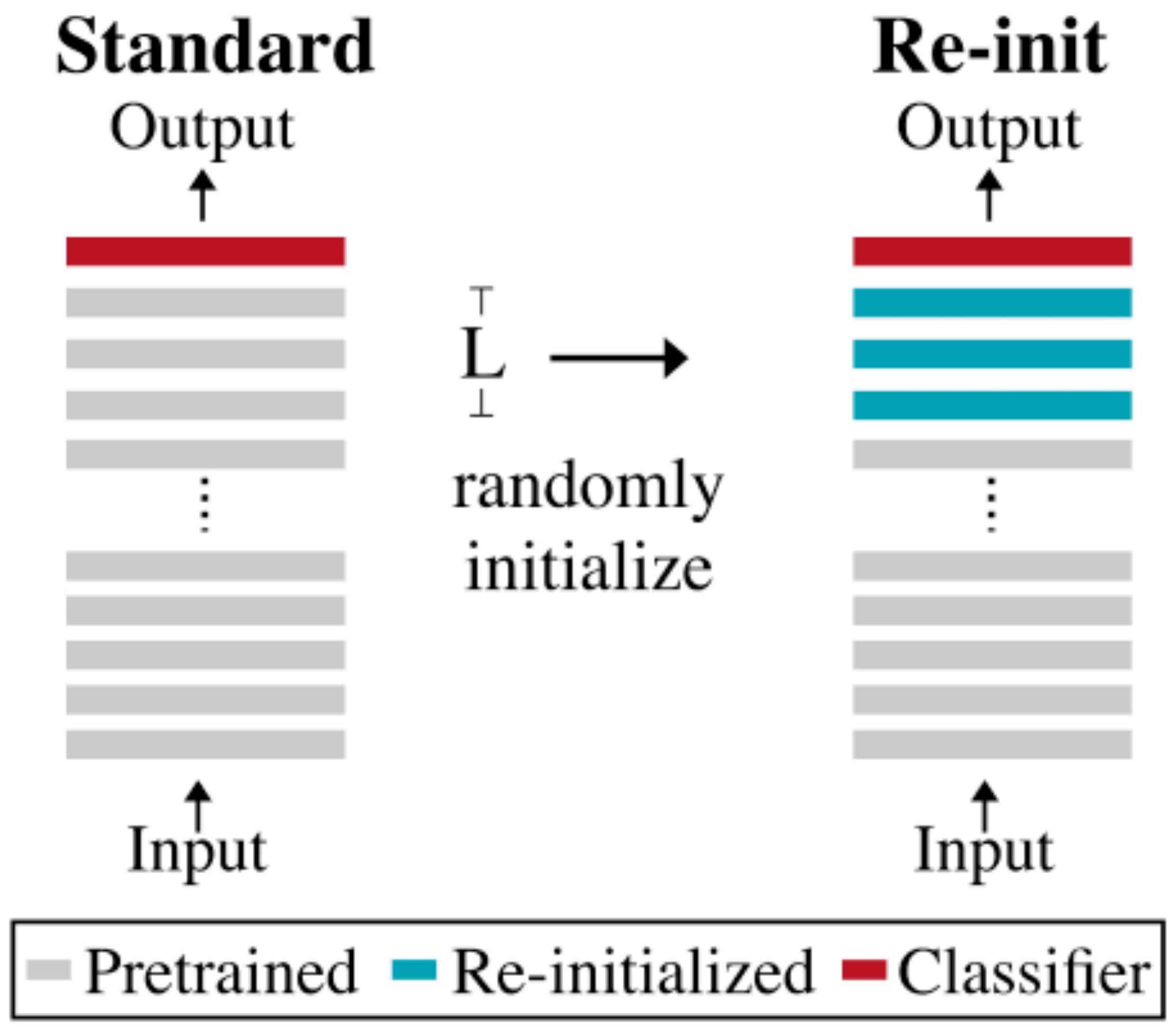
ASAPP

Take-aways

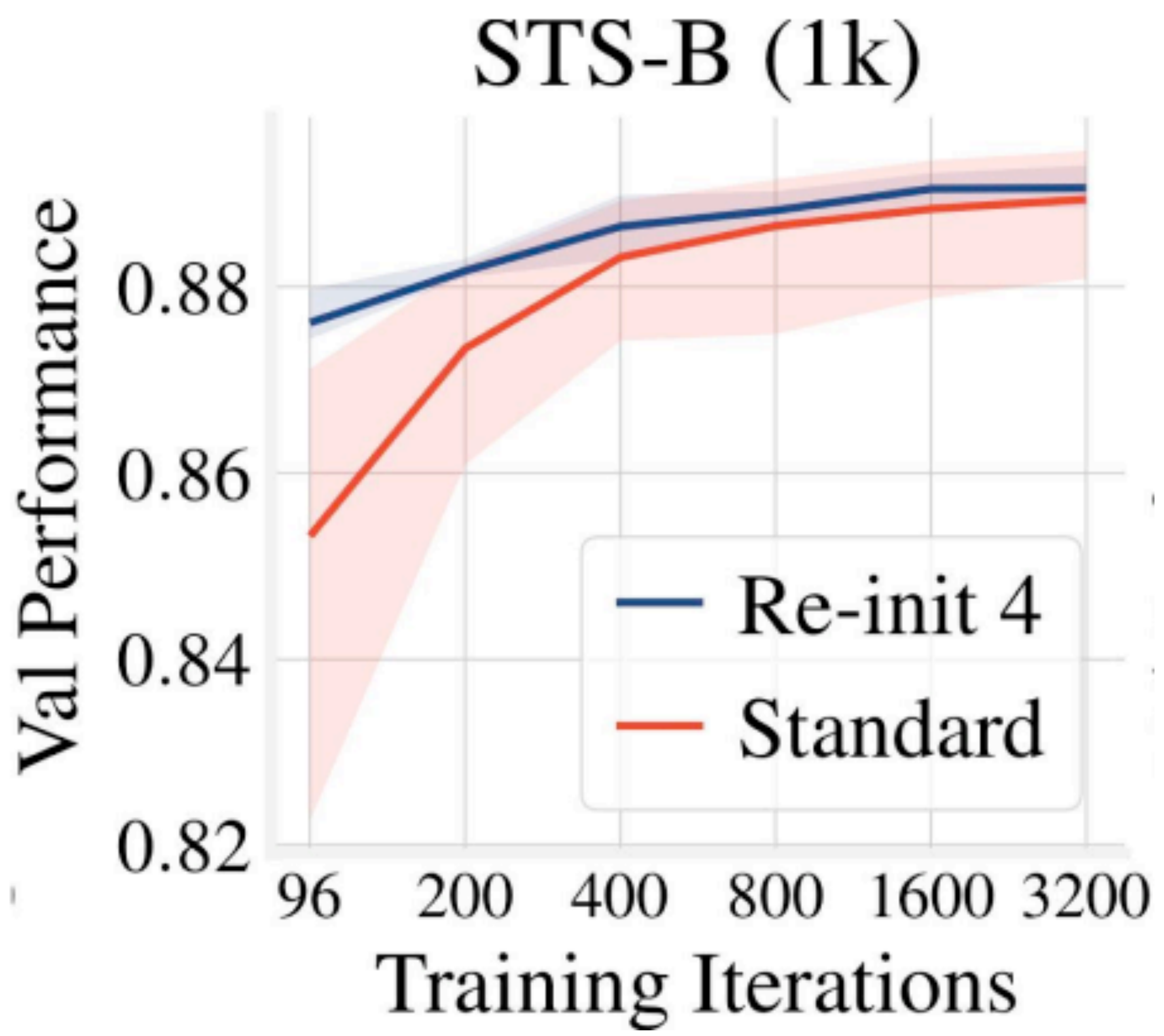
Bias Correction



Re-initialization



Training Longer



Thank you!

Paper:

Code:

ASAPP