Does RoBERTa Perform Better than BERT in Continual Learning: An Attention Sink Perspective

Xueying Bai, Yifan Sun, Niranjan Balasubramanian

Stony Brook University

Problem Statement

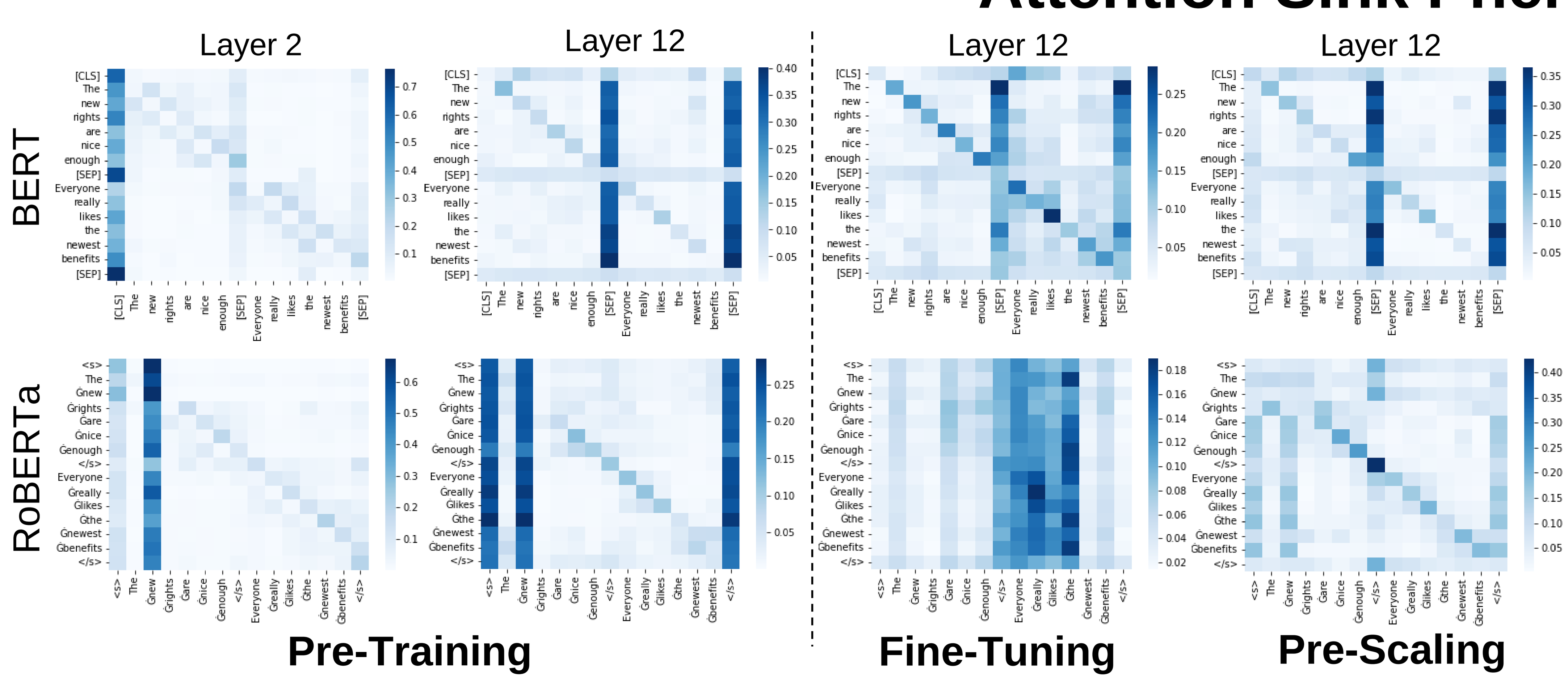
Continual Learning (CL): A Model sequentially learns new tasks without forgetting previous tasks' knowledge.

Question: Does a pre-trained model which has better single-task performance also perform better in CL?

Previous Works' Results: Not always. RoBERTa does not always outperform BERT in CL tasks (Wu et al., 2022).

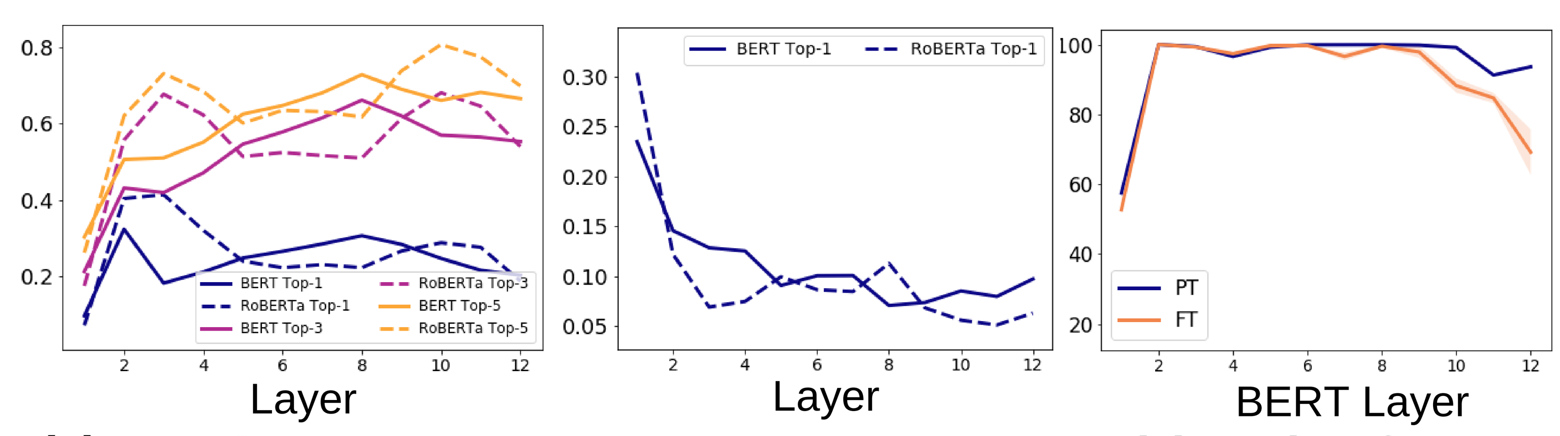
This Paper: Study the influence of attention sinks in models' CL performance.

Attention Sink Phenomenon



Measurements

- Average outer degree: $d_i = \sum_{k=1}^{n} a_{ki} / n$
- Attention deviation: $\Delta_i = \sqrt{\sum_{k=1}^{n} (a_{ki} - d_i)^2 / (n d_i)}$
- Ratio of CST: Ratio of sink tokens that are in the set of common tokens.



(a) Avg Outer Degree (b) Sink Att Deviation (c) Ratio of CST

- ❑ High attention scores are allocated to specific tokens (i.e., sink tokens): (1) high average degrees; (2) small attention deviation.
- ❑ Sink tokens are usually common tokens shared across different tasks (e.g., [SEP], punctuation).

Single Task: Over-Smoothing

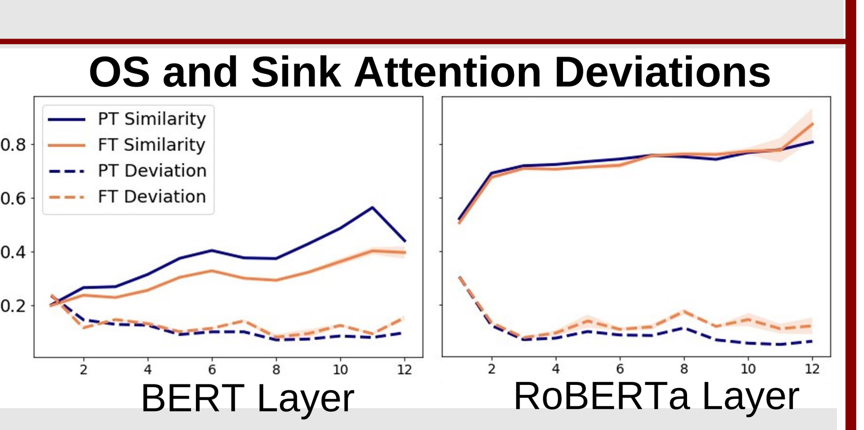
Over-Smoothing (OS): token representations become identical after several self-attention layers.

Over-Smoothing is related to the attention deviations of sink tokens:

$$d_{\mathcal{M}}(\mathbf{AH}) \leq \sqrt{\lambda_{\max}}\, d_{\mathcal{M}}(\mathbf{H})$$

$$\lambda_{\max} \geq \max_i \sum_{k=1}^{n} (a_{ki} - d_i)^2$$

OS and Sink Attention Deviations



⇒ Over-Smoothing may occur with attention sinks above.

⇒ Model distorts pre-trained features -> less generalizable.

Cross Task: Interference

Interference: dot product between a model's (vectorized) gradients on different tasks' losses.

Case study: Given (1) two **irrelevant** tasks; (2) each task's input token embeddings are **orthogonal except embeddings of common sink tokens**; (3) a single-head attention layer:

*If sink attention deviations are **small**, the interference largely depends on dot product between sink token representations.*

⇒ Attention sink on common tokens may propagate unexpected interference across tasks.

Method: Pre-Scaling Mechanism For Diverse Attention

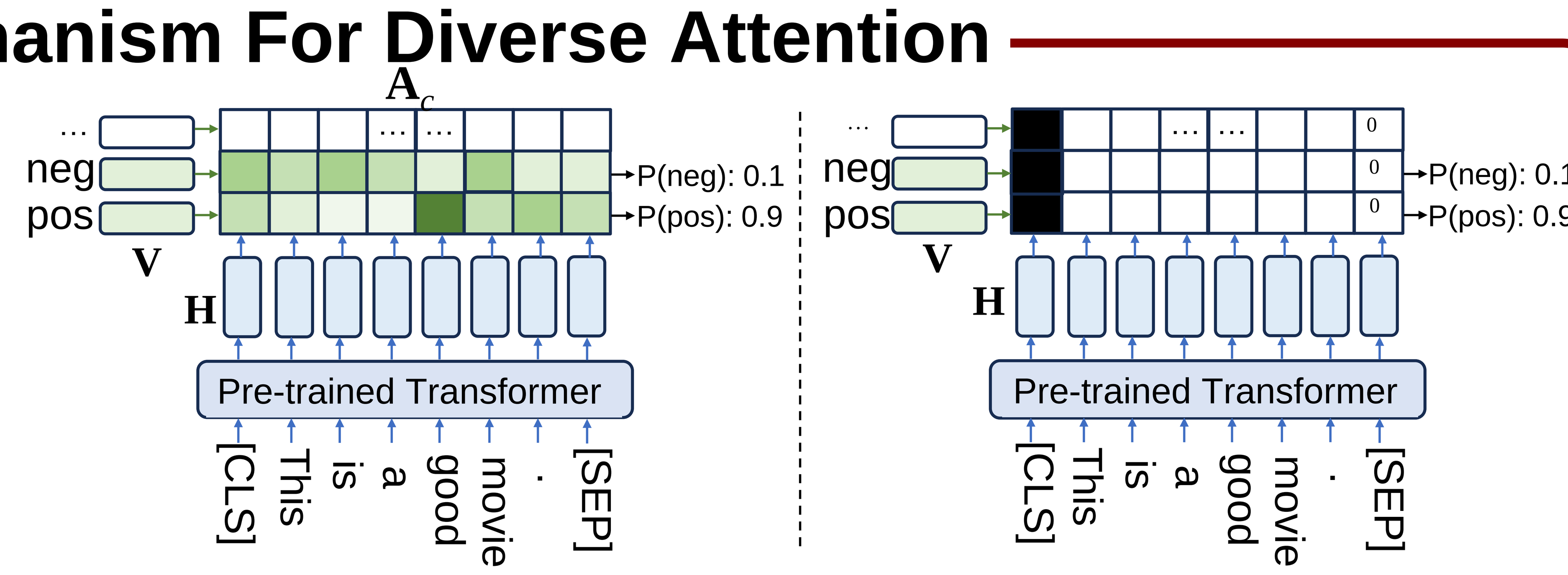
Motivation: pre-trained representations of non-sink tokens (e.g., 'fantastic' in sentiment analysis task) may contain more information about downstream tasks.

Two-Step Training:

- **Pre-Scaling:** pre-scale classes' attentions on tokens.

  $$\mathbf{A}_c = \text{softmax}(\mathbf{V}\, f(\mathbf{H})^T / \sqrt{d}\,) \qquad [\mathbf{V} \text{ and } f(.) \text{ are learnable}]$$

- **Fine-tuning:** fine-tune the whole model, including the encoder and the scaling layer.



(a). Scaling Model (b). Regular Model

Experiments

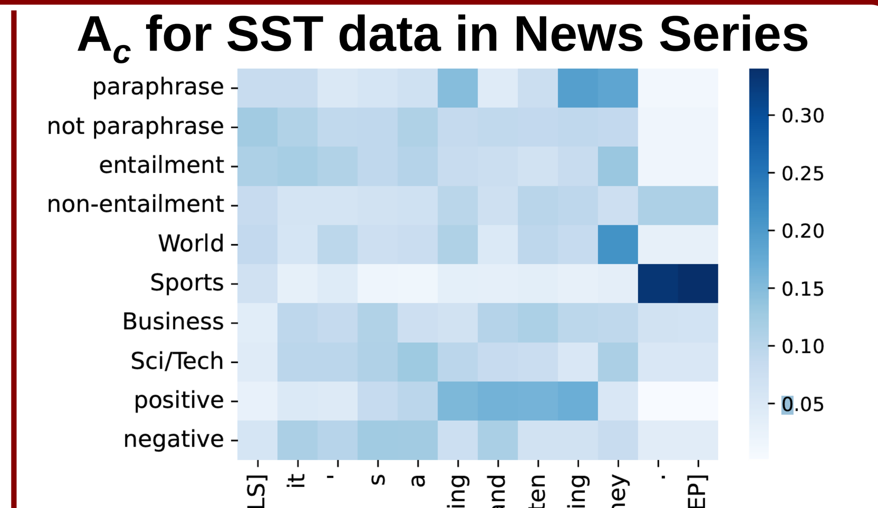
Table 1. CL evaluation by sequentially tuning BERT and RoBERTa

	Model	Yahoo Split ACC std	Yahoo Split FGT std	DB Split ACC std	DB Split FGT std	News Series ACC std	News Series FGT std
BERT	Probing	88.43 0.06	—	99.30 0.03	—	74.81 0.46	—
	FT	86.19 0.92	6.70 1.08	66.22 8.13	39.15 9.47	68.98 5.68	17.13 7.48
	PT+FT	90.24 0.53	2.23 0.77	98.47 2.23	1.64 2.59	77.09 2.11	8.16 2.50
	Prescale (ours)	**90.92** 0.53	**1.47** 0.71	**99.74** 0.05	**0.13** 0.06	**79.76** 0.76	**4.40** 1.18
RoBERTa	Probing	88.06 0.09	—	99.33 0.01	—	68.27 1.32	—
	FT	83.54 4.66	10.91 5.74	71.94 7.48	32.53 8.73	70.61 4.42	18.24 5.21
	PT+FT	90.76 0.86	2.14 1.06	99.68 0.24	0.21 0.29	79.39 2.00	8.01 3.24
	Prescale (ours)	**90.92** 0.77	**1.95** 1.01	**99.78** 0.08	**0.09** 0.11	**81.59** 1.74	**4.16** 2.33

Table 2. CL evaluation compared to other CL models

	Model	Yahoo Split ACC std	Yahoo Split FGT std	DB Split ACC std	DB Split FGT std	News Series ACC std	News Series FGT std
CL	ER	87.42 0.52	5.61 0.68	91.05 10.20	10.20 10.14	75.47 3.93	7.81 5.27
	A-GEM	89.43 0.58	2.95 0.64	94.71 4.70	5.98 5.49	75.90 3.34	6.60 3.84
	MBPA++	86.50 2.78	6.62 2.82	97.17 3.76	3.09 3.68	72.55 5.50	9.64 3.99
	IDBR (-R)	89.32 1.46	2.74 1.35	96.47 4.67	3.95 4.66	72.36 2.93	8.67 4.23
	IDBR	90.48 0.55	1.32 0.64	**99.84** 0.03	0.04 0.03	76.90 1.98	3.24 2.50
	CTR	87.06 1.23	1.28 0.93	99.04 0.95	0.29 0.35	75.12 3.09	3.40 2.92
	L2P	90.82 0.58	0.60 0.56	99.63 0.36	0.29 0.36	73.99 2.36	3.43 2.42
Sequential	FT	86.19 0.92	6.70 1.08	66.22 8.13	39.15 9.47	68.98 5.68	17.13 7.48
	Prescale (ours)	**90.92** 0.53	1.47 0.71	99.74 0.05	0.13 0.06	**79.76** 0.76	4.40 1.18
Non-CL	Separate	92.25 0.04	—	99.87 0.01	—	83.72 0.53	—
	MTL	92.27 0.05	—	99.88 0.01	—	82.04 0.90	—

A_c for SST data in News Series



❑ With pre-scaling, RoBERTa consistently outperforms BERT in CL. The model is powerful even without replay.

❑ Pre-Scaling allocates high attention on non-sink tokens.

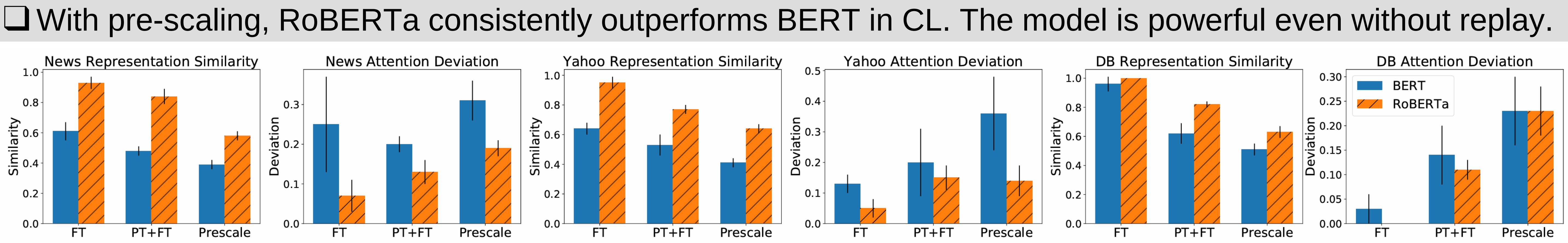


Table 3: *ACC* and *FGT* of different scaling strategies on News Series.

	Model	*ACC*	*FGT*
BERT Prescale	Uniform	78.98	5.32
	Sink	76.08	9.27
	Full	79.76	4.40
RoBERTa Prescale	Uniform	79.44	7.07
	Sink	79.09	9.26
	Full	81.59	4.16

Table 4: *ACC* on task-agnostic evaluations for DB and Yahoo Split.

	Model	*DB*	*Yahoo*
BERT	FT	15.90	36.19
	PT+FT	72.41	53.34
	Prescale	70.38	53.21
RoBERTa	FT	18.71	36.24
	PT+FT	67.32	52.98
	Prescale	77.55	53.51

❑ Pre-Scaling increases attention deviations, and reduces representations' over-smoothing.