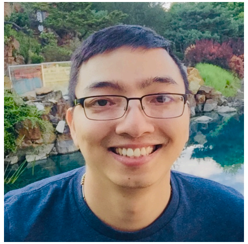
STraTA: Self-Training with Task Augmentation for Better Few-shot Learning

Tu Vu

November, 2021

STraTA: Self-Training with Task Augmentation for Better Few-shot Learning

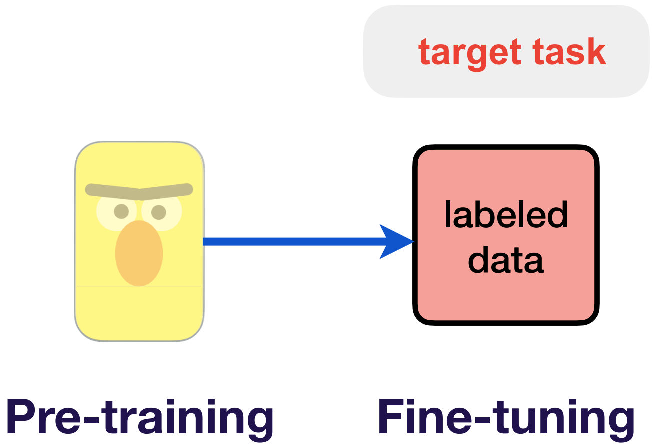
Tu Vu[1,2] Thang Luong[1] Quoc Le[1] Grady Simon[1] Mohit Iyyer[2]

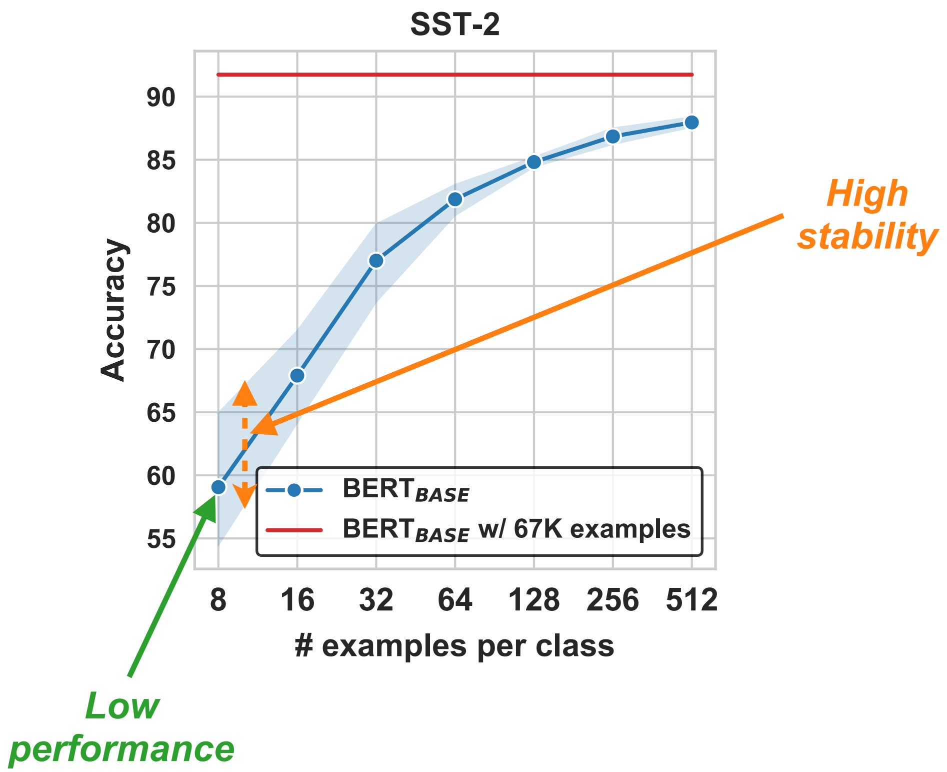
Google AI[1] UMass Amherst[2]

Agenda

- Motivation
- STraTA: Self-training with Task Augmentation
- Results and Discussion
- Conclusion

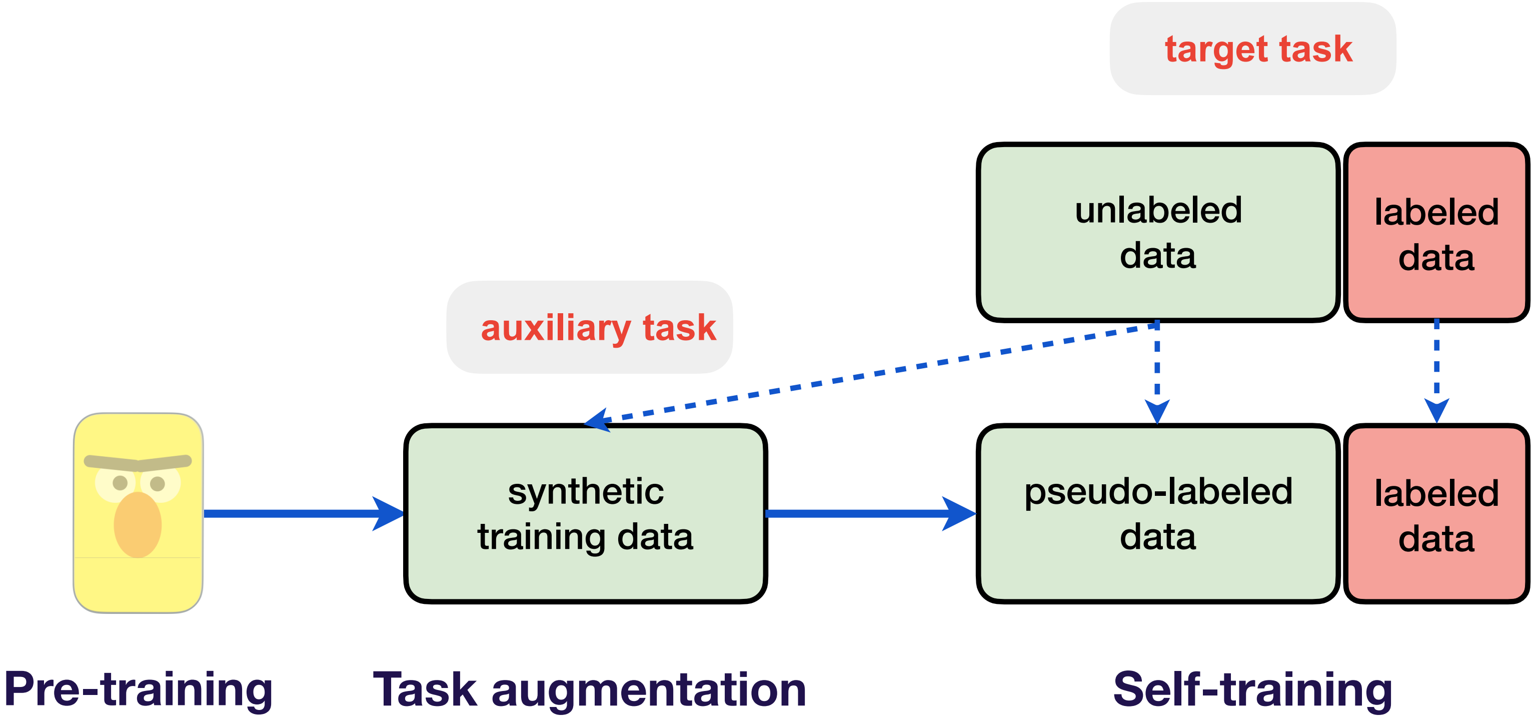
The current dominant learning paradigm





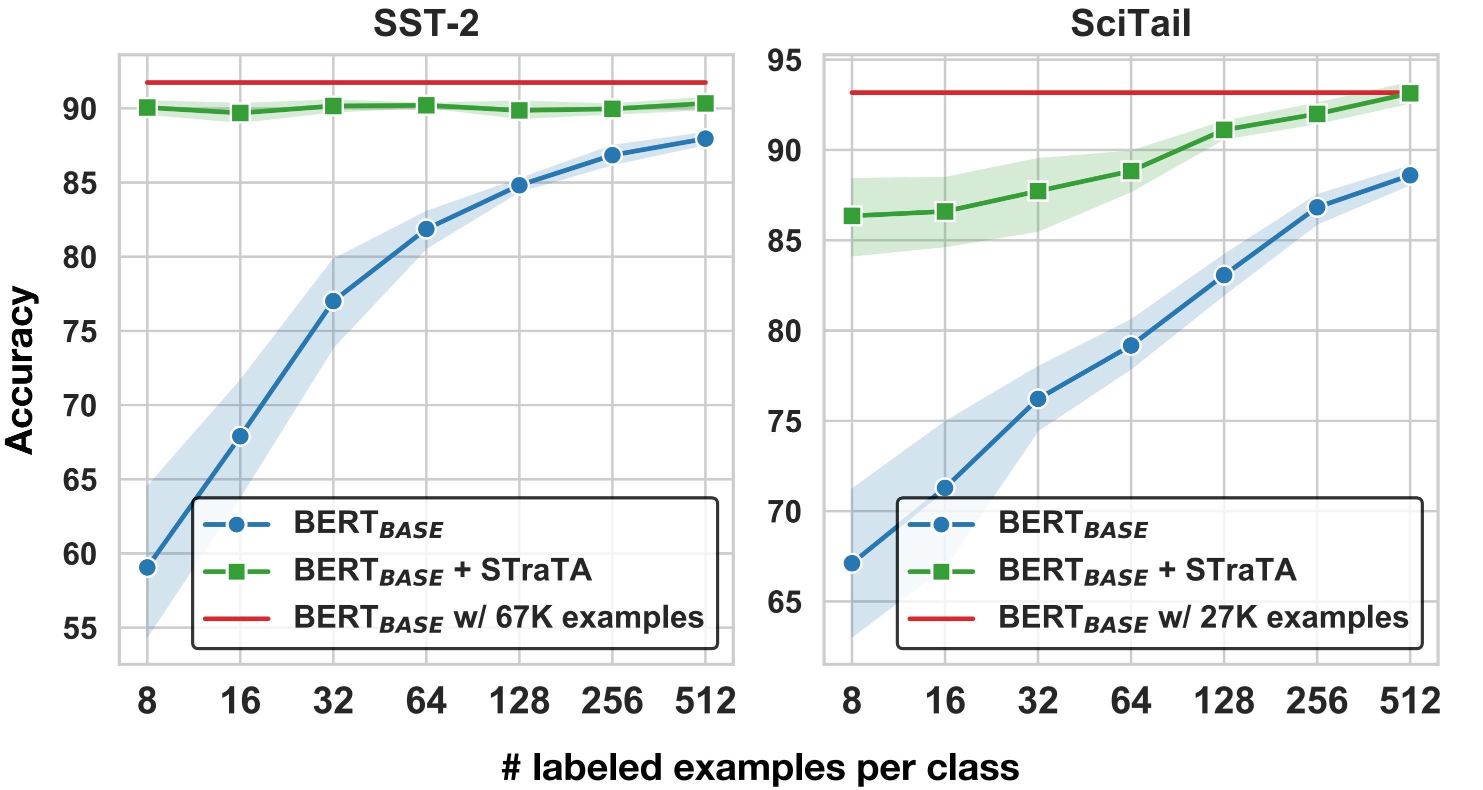
Credit to Jay Alammar for creating the BERT image

Exploiting task-specific unlabeled data

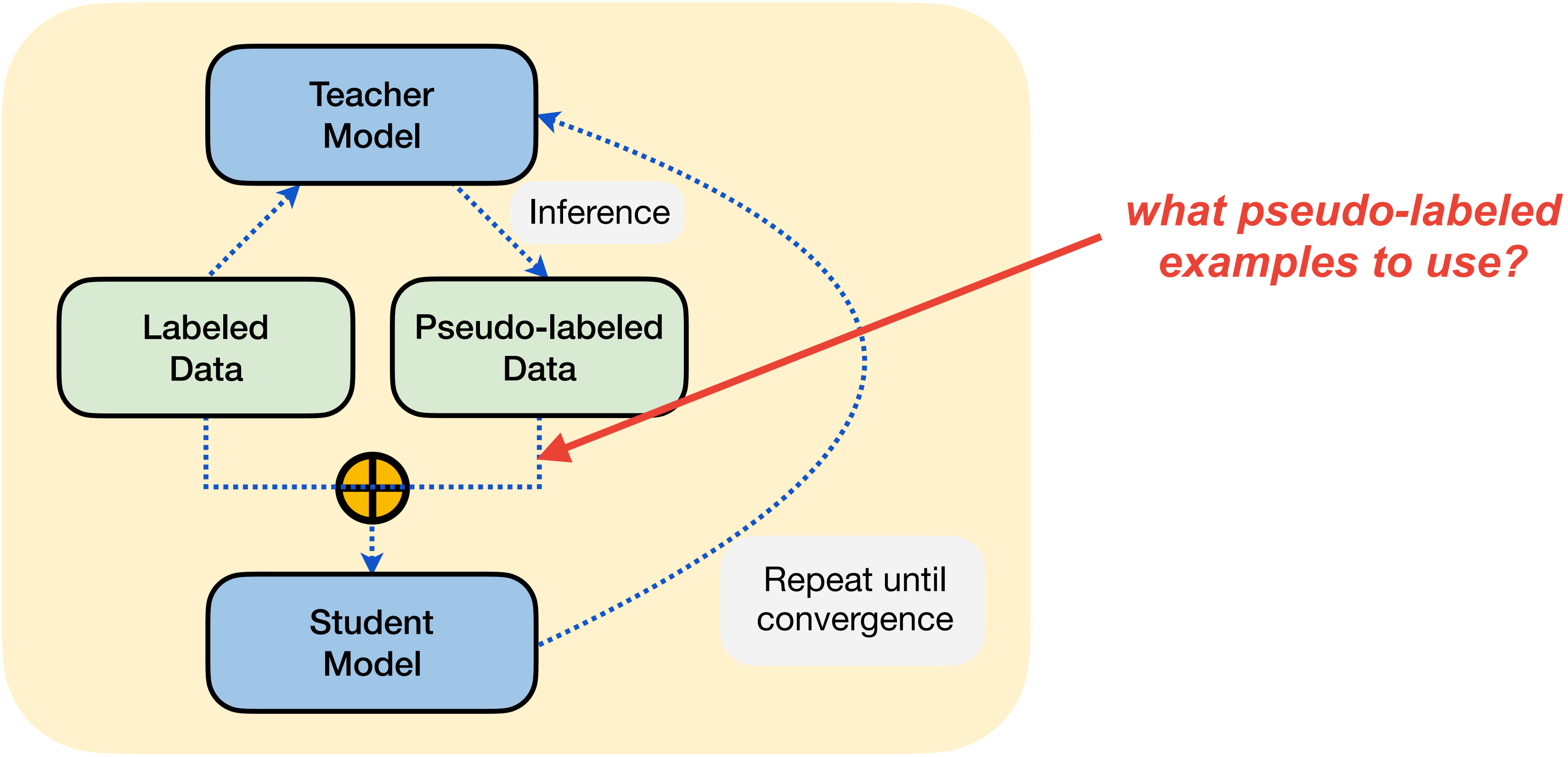


Credit to Jay Alammar for creating the BERT image

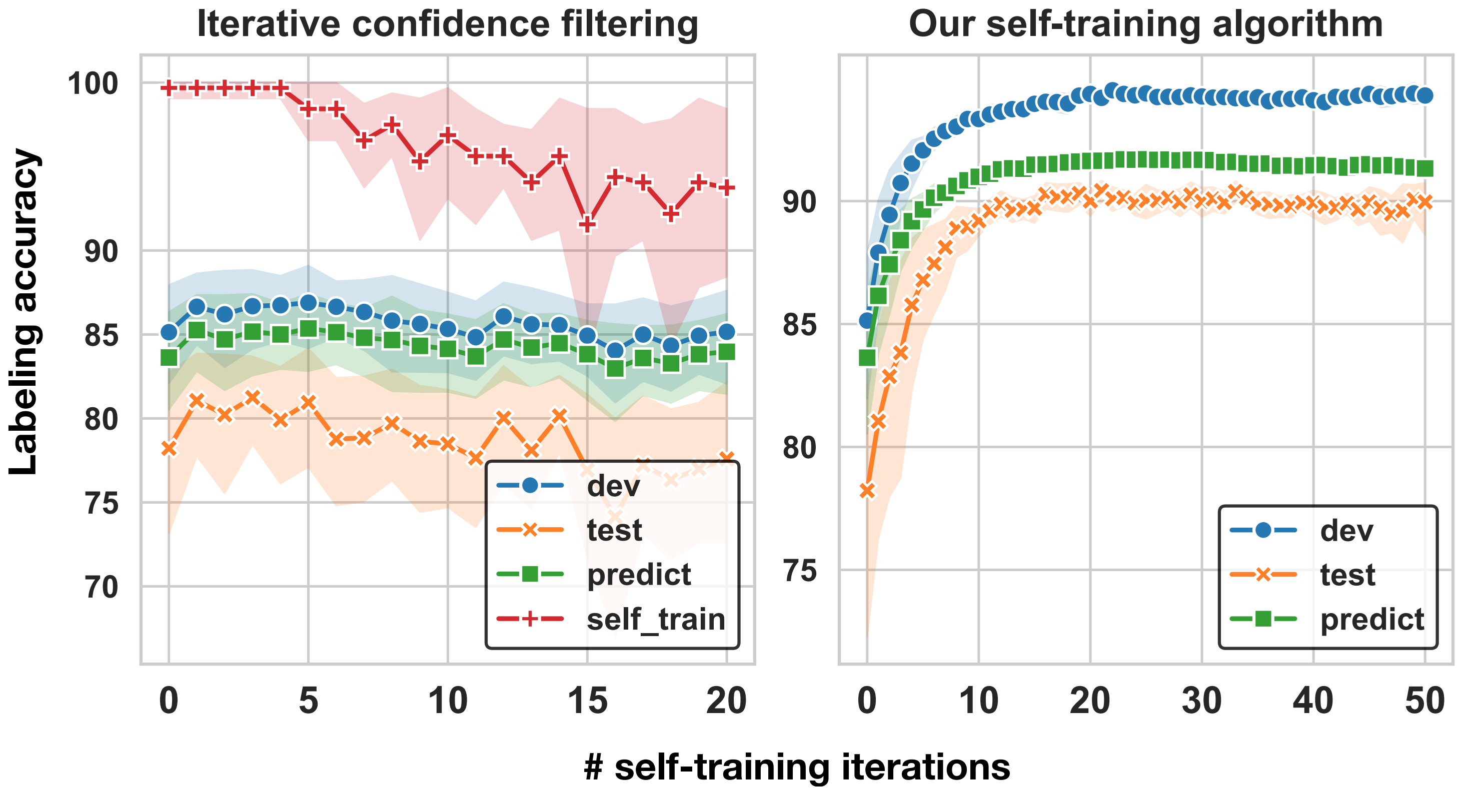
STraTA substantially improves sample efficiency



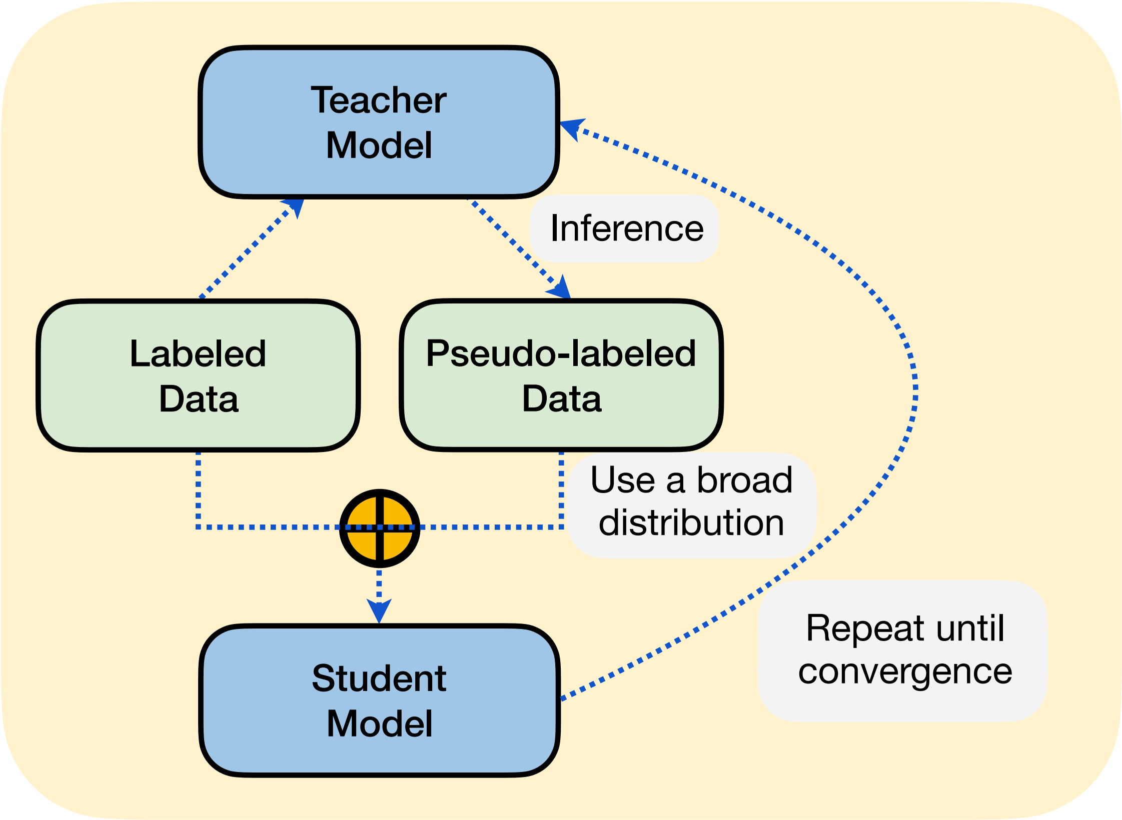
What is self-training?



Self-training on a broad distribution of pseudo-labeled data

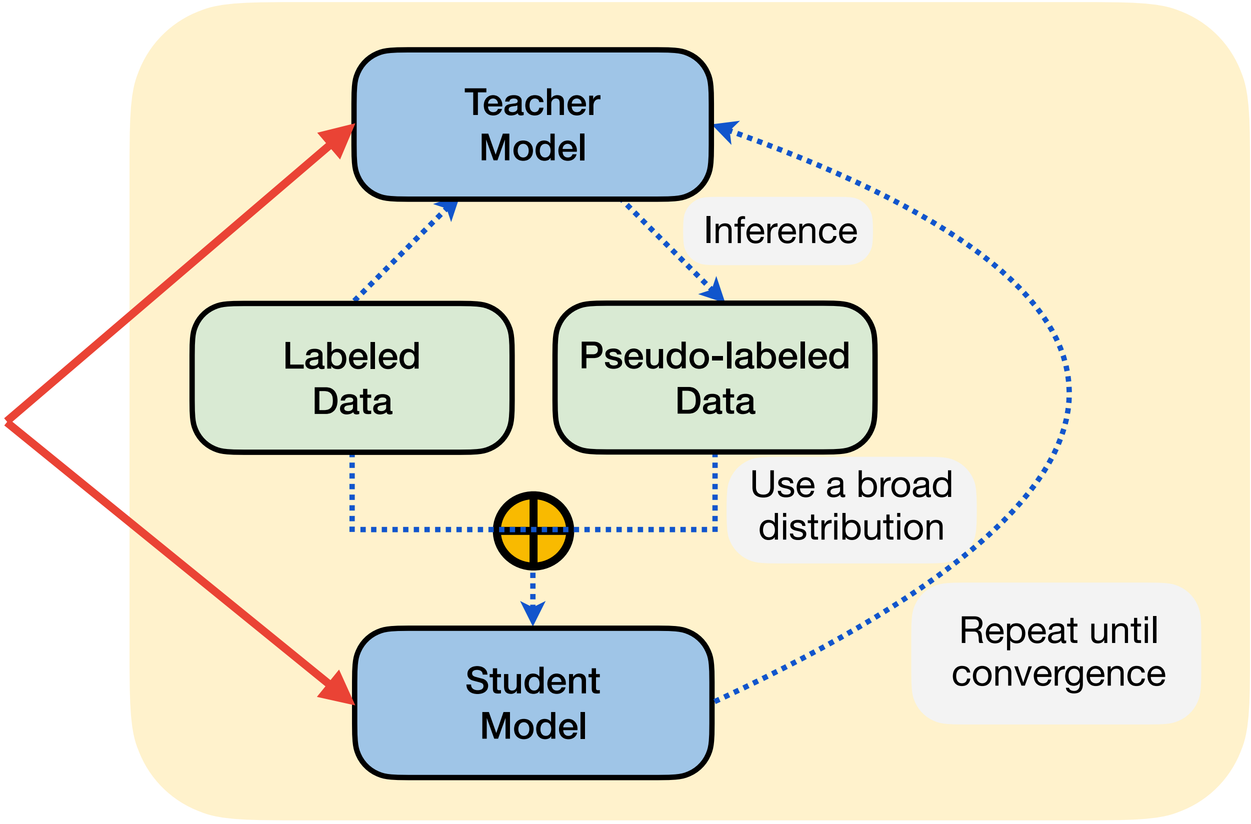


Our self-training algorithm

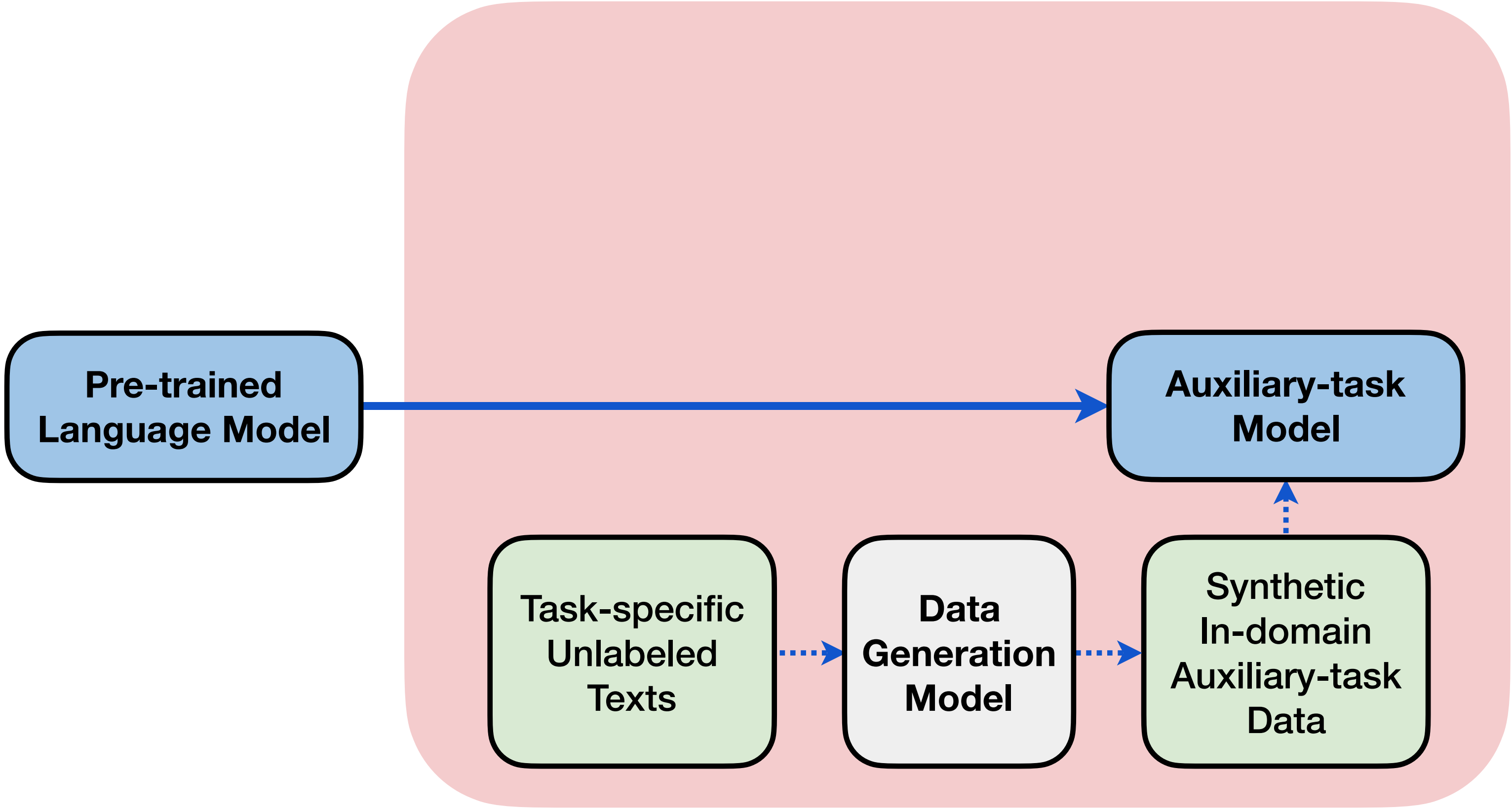


Our self-training algorithm (cont.)

what model to use?



Task Augmentation



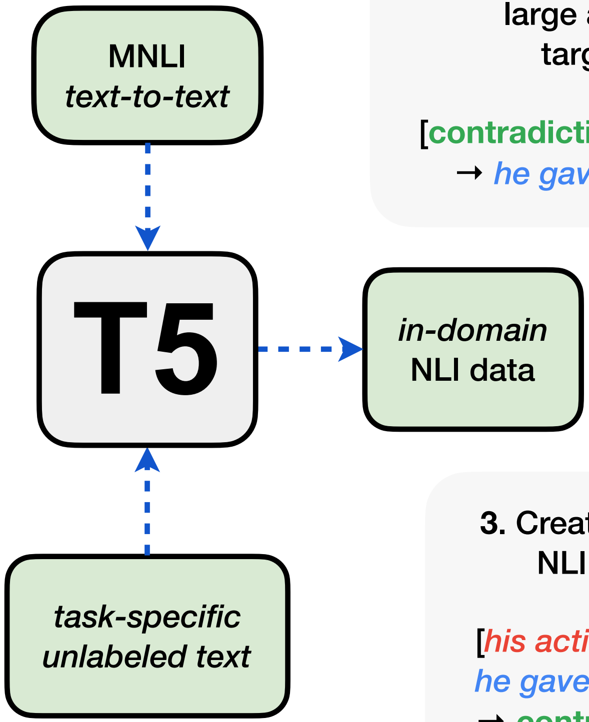
Task augmentation

1. Train an NLI data generator by fine-tuning a pre-trained generative model on the MNLI dataset in a *text-to text* format

[**entailment:** *I have met a woman whom I am attracted to*] → *I am attracted to a woman I met*

2. Use the model to simulate a large amount of NLI data using target-task unlabeled text

[**contradiction:** *his acting was really awful*] → *he gave an incredible performance*



his acting was really awful

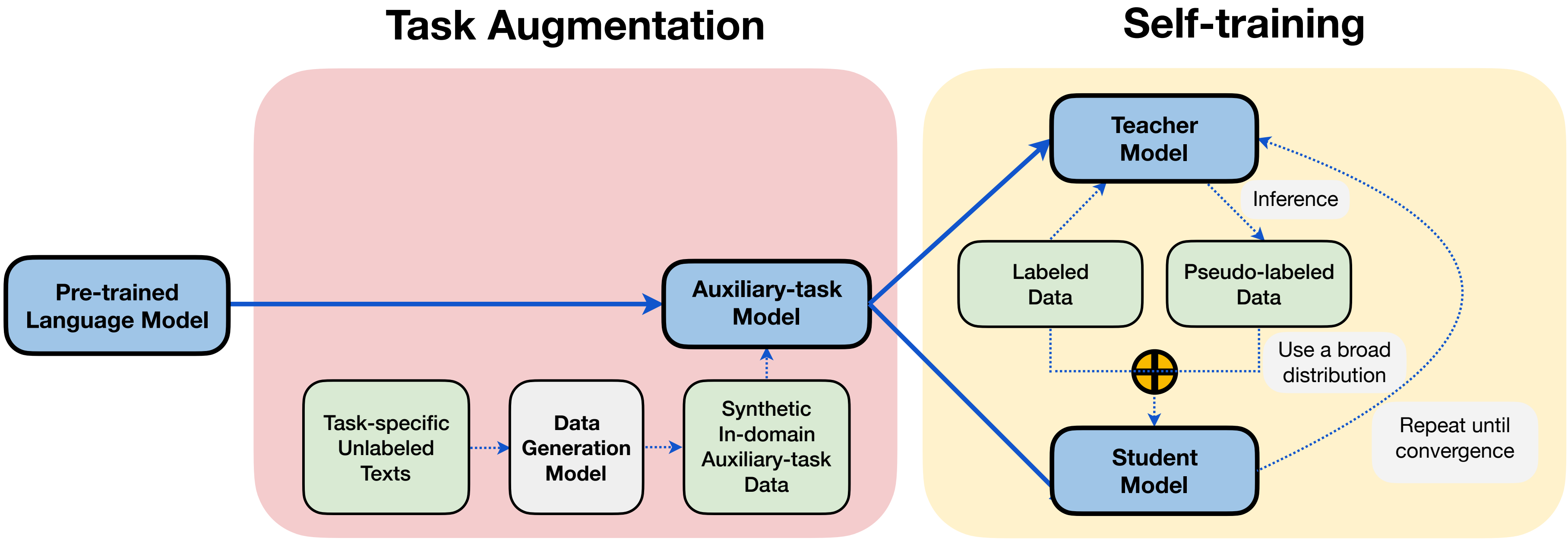
3. Create synthetic in-domain NLI training examples

[*his acting was really awful*, *he gave an incredible performance*] → **contradiction**

Example outputs

unlabeled text from SST-2	
Input: a cast of a-list brit actors	
+ **entailment**	→ **Output:** the cast includes british actors.
+ **neutral**	→ **Output:** the cast includes a-list brit actors such as nicole kidman and charlie chaplin.
+ **contradiction**	→ **Output:** the cast was all new and unknown.
Input: both heartbreaking and heartwarming ... just a simple fable done in an artless sytle , but it ’s tremendously moving	
+ **entailment**	→ **Output:** the story is moving and touching.
+ **neutral**	→ **Output:** the artless style used to make the fable was a favorite of many.
+ **contradiction**	→ **Output:** the fable is artful and sophisticated but it is also heartbreaking.
unlabeled text from SciTail	
Input: Surface tension is responsible for the fact that small insects can walk on water.	
+ **entailment**	→ **Output:** small insects can walk on water due to surface tension.
+ **neutral**	→ **Output:** it’s difficult for large insects to walk on water.
+ **contradiction**	→ **Output:** small insects cannot walk on water because of surface tension.
Input: Tamiflu inhibits spread of virus.	
+ **entailment**	→ **Output:** the virus is less spread when tamiflu is used.
+ **neutral**	→ **Output:** tamiflu is a good way to treat hiv.
+ **contradiction**	→ **Output:** tamiflu promotes viral spread.

STraTA: Self-training with Task Augmentation



Experimental setup: datasets

Task	|Train|	Task type	Domain
text classification/regression			
SNLI (Bowman et al., 2015)	570K	NLI	misc.
MNLI (Williams et al., 2018)	393K	NLI	misc.
QQP (Iyer et al., 2017)	364K	paraphrase identification	social QA
QNLI (Wang et al., 2019b)	105K	QA-NLI	Wikipedia
SST-2 (Socher et al., 2013)	67K	sentiment analysis	movie reviews
SciTail (Khot et al., 2018)	27K	NLI	science QA
SST-5 (Socher et al., 2013)	8.5K	sentiment analysis	movie reviews
STS-B (Cer et al., 2017)	7K	semantic similarity	misc.
SICK-E (Marelli et al., 2014)	4.5K	NLI	misc.
SICK-R (Marelli et al., 2014)	4.5K	semantic similarity	misc.
CR (Hu and Liu, 2004)	4K	sentiment analysis	product reviews
MRPC (Dolan and Brockett, 2005)	3.7K	paraphrase identification	news
RTE (Dagan et al., 2005, et seq.)	2.5K	NLI	news, Wikipedia

Datasets used in our experiments and their characteristics, sorted by training dataset size.

Experimental setup: baselines

LMFT & ITFT

- **LMFT**: target-task language model fine-tuning (Howard and Ruder, 2018; Gururangan et al., 2020)
- **ITFT**: intermediate-task fine-tuning with MNLI (Phang et al., 2019)

Prompt/entailment-based fine-tuning

- **LM-BFF:** prompt-based fine-tuning (Gao et al., 2021)
- **EFL:** entailment-based fine-tuning (Wang et al., 2021)

Du et al. (2021)

- **SentAugST:** Retrieval-based augmentation (SentAug) + self-training (ST)

Main results

STraTA significantly improves results across 12 NLP benchmark datasets (numbers in the subscript indicate the standard deviation across 10 random seeds).

Model	SNLI	QQP	QNLI	SST-2	SciTail	SST-5	STS-B
			FULL				
$\text{BERT}_{\text{LARGE}}$	91.1	88.4	91.9	92.4	95.3	$53.7_{0.9}$	$89.6_{0.2}$
+ LMFT	91.0	88.1	90.4	93.5	95.3	$54.0_{0.4}$	$89.5_{0.2}$
+ $\text{ITFT}_{\text{MNLI}}$	91.1	88.2	91.6	93.5	96.5	$54.0_{0.8}$	$90.3_{0.3}$
+ TA	**91.9**	**88.5**	**92.5**	**94.7**	**96.9**	$\mathbf{55.7}_{0.8}$	$\mathbf{90.9}_{0.2}$
			LIMITED *(1024 total training examples)*				
$\text{BERT}_{\text{LARGE}}$	$77.4_{0.6}$	$74.1_{1.0}$	$81.7_{0.9}$	$89.8_{0.6}$	$90.9_{0.7}$	$49.1_{1.3}$	$88.2_{0.4}$
+ LMFT	$75.8_{1.5}$	$71.6_{0.5}$	$80.5_{2.0}$	$88.9_{0.8}$	$87.7_{2.3}$	$49.2_{3.1}$	$88.4_{0.4}$
+ $\text{ITFT}_{\text{MNLI}}$	$85.2_{0.4}$	$74.0_{0.5}$	$83.5_{0.5}$	$90.0_{0.8}$	$92.1_{1.1}$	$49.4_{1.2}$	$87.8_{0.8}$
+ TA	$\mathbf{87.3}_{0.3}$	$\mathbf{75.7}_{0.5}$	$\mathbf{85.0}_{0.5}$	$\mathbf{91.7}_{0.7}$	$\mathbf{92.3}_{1.1}$	$\mathbf{51.4}_{1.0}$	$\mathbf{89.0}_{0.6}$
			FEW-SHOT *(8 training examples per class)*				
$\text{BERT}_{\text{LARGE}}$	$43.1_{4.4}$	$58.5_{4.7}$	$64.4_{6.1}$	$66.1_{8.7}$	$68.8_{9.5}$	$35.2_{1.3}$	$74.6_{3.8}$
+ LMFT	$39.6_{2.6}$	$52.7_{4.7}$	$52.2_{1.6}$	$66.3_{9.3}$	$66.4_{10.6}$	$36.8_{2.9}$	$75.4_{9.4}$
+ $\text{ITFT}_{\text{MNLI}}$	$79.9_{3.1}$	$62.6_{9.0}$	$64.5_{4.4}$	$80.7_{5.0}$	$72.3_{11.2}$	$36.4_{2.1}$	$75.5_{4.0}$
+ TA	$84.8_{0.7}$	$64.6_{6.3}$	$71.5_{4.0}$	$85.5_{1.4}$	$79.0_{4.5}$	$38.5_{3.0}$	$78.9_{2.4}$
+ ST	$69.3_{9.2}$	$74.3_{1.2}$	$85.4_{1.7}$	$81.9_{12.2}$	$79.9_{4.8}$	$42.0_{1.5}$	$82.8_{2.3}$
+ STraTA	$\mathbf{87.3}_{0.3}$	$\mathbf{75.1}_{0.2}$	$\mathbf{86.4}_{0.8}$	$\mathbf{91.7}_{0.7}$	$\mathbf{87.3}_{2.9}$	$\mathbf{43.0}_{2.3}$	$\mathbf{84.5}_{1.6}$
Prompt-based (LM-BFF; Gao et al., 2021) and entailment-based (EFL; Wang et al., 2021) methods							
$\text{RoBERTa}_{\text{LARGE}}$	$38.4_{1.3}$	$58.8_{9.9}$	$52.7_{1.8}$	$60.5_{3.1}$	–	–	$24.5_{8.4}$
+ LM-BFF	$52.0_{1.7}$	$\mathbf{68.2}_{1.2}$	$61.8_{3.2}$	$79.9_{6.0}$	–	–	$66.0_{3.2}$
+ EFL	$\mathbf{81.0}_{1.1}$	$67.3_{2.6}$	$\mathbf{68.0}_{3.4}$	$\mathbf{90.8}_{1.0}$	–	–	$\mathbf{71.0}_{1.3}$

Main results (cont.)

Compared to Du et al. (2021), our approach leads to better downstream performance, despite using a weaker base model (BERT vs. RoBERTa) and with less labeled examples.

Model	SST-2	SST-5	CR
Ours (8 examples per class)			
$\text{BERT}_{\text{BASE}}$	$69.8_{6.5}$	$32.8_{2.0}$	$73.1_{0.5}$
+ TA	$85.5_{0.6}$	$41.0_{0.8}$	$88.7_{0.2}$
+ ST	$74.9_{9.0}$	$38.3_{0.8}$	$85.6_{1.8}$
+ STraTA	$\mathbf{90.8}_{0.6}$	$\mathbf{43.1}_{1.1}$	$\mathbf{91.4}_{0.2}$
$\text{BERT}_{\text{LARGE}}$	$75.6_{3.3}$	$36.6_{0.4}$	$79.3_{0.7}$
+ TA	$87.3_{0.3}$	$41.7_{1.1}$	$90.0_{0.4}$
+ ST	$90.6_{0.3}$	$43.8_{0.4}$	$89.0_{1.1}$
+ STraTA	$\mathbf{92.4}_{0.1}$	$\mathbf{45.5}_{0.7}$	$\mathbf{90.6}_{0.0}$
Du et al. (2021) (20 examples per class)			
$\text{RoBERTa}_{\text{LARGE}}$	$83.6_{2.7}$	$42.3_{1.6}$	$88.9_{1.7}$
+ SentAugST	$\mathbf{86.7}_{2.3}$	$\mathbf{44.4}_{1.0}$	$\mathbf{89.7}_{2.0}$

STraTA improves a randomly-initialized base model

Model	SST-2	SciTail
$\text{RAND}_{\text{BASE}}$	$50.0_{1.6}$	$50.7_{2.4}$
+ STraTA	$78.6_{0.9}$	$64.4_{3.1}$
$\text{BERT}_{\text{BASE}}$	$59.1_{8.4}$	$67.1_{6.6}$
+ STraTA	$90.1_{0.8}$	$86.3_{3.5}$
$\text{BERT}_{\text{LARGE}}$	$66.1_{8.7}$	$68.8_{9.5}$
+ STraTA	$91.7_{0.7}$	$87.3_{2.9}$

Our approach yields improvements even when starting with a randomly-initialized model, but pre-training helps considerably.

Does self-training work with out-of-domain/ distribution unlabeled data?

Model	SciTail	CR	MRPC	RTE
BERT$_{\text{BASE}}$	$67.1_{6.6}$	$65.2_{8.2}$	$72.4_{10.2}$	$51.4_{2.5}$
BERT$_{\text{BASE}}$+ TA	$78.5_{3.2}$	$86.5_{2.2}$	$74.5_{6.5}$	$67.6_{7.1}$
+ ST$_{\text{IN}}$	$86.3_{3.5}$	$90.5_{0.8}$	$81.0_{0.8}$	$70.6_{2.4}$
+ ST$_{\text{OUT}}$	$81.4_{3.7}$	$88.3_{1.9}$	$80.3_{1.9}$	$71.2_{3.2}$
+ ST$_{\text{IN + OUT}}$	$82.6_{2.6}$	$88.3_{1.5}$	$80.2_{1.1}$	$69.9_{4.0}$

Self-training with out-of-domain unlabeled examples also results in improvements, but using in-domain data works significantly better.

Towards realistic evaluation in few-shot learning

Model	SST-2	SciTail
$\text{BERT}_{\text{BASE}}$	$58.8_{8.4}$ (↓ 0.3)	$61.5_{5.4}$ (↓ 5.6)
+ LMFT	$64.0_{8.1}$ (↓ 0.9)	$59.3_{5.6}$ (↓ 4.7)
+ $\text{ITFT}_{\text{MNLI}}$	$76.5_{7.2}$ (↓ 0.3)	$76.2_{5.4}$ (↑ 0.4)
+ TA	$79.8_{6.3}$ (↓ 0.5)	$77.8_{3.3}$ (↓ 0.7)
+ STraTA	$86.6_{2.6}$ (↓ 3.5)	$80.6_{3.0}$ (↓ 5.7)

In a realistic evaluation without a development set, our **STraTA** approach still leads to significant improvements on top of $\text{BERT}_{\text{BASE}}$. In parentheses, we show the absolute increase (↑) or decrease (↓) in performance compared to the same method used with a development set.

Conclusion

STraTA

- ✦ two *complementary* and and *independently effective* methods to leverage task-specific unlabeled data for improved downstream performance
  - **task augmentation:** synthesizes a large amount of in-domain data for auxiliary-task fine-tuning from target-task unlabeled texts
  - **self-training**: trains on a broad distribution of pseudo-labeled data

- ✦ substantially improves sample efficiency across 12 NLP benchmark datasets

Thank you!

Code will be available at
https://github.com/google-research/google-research/tree/master/STraTA