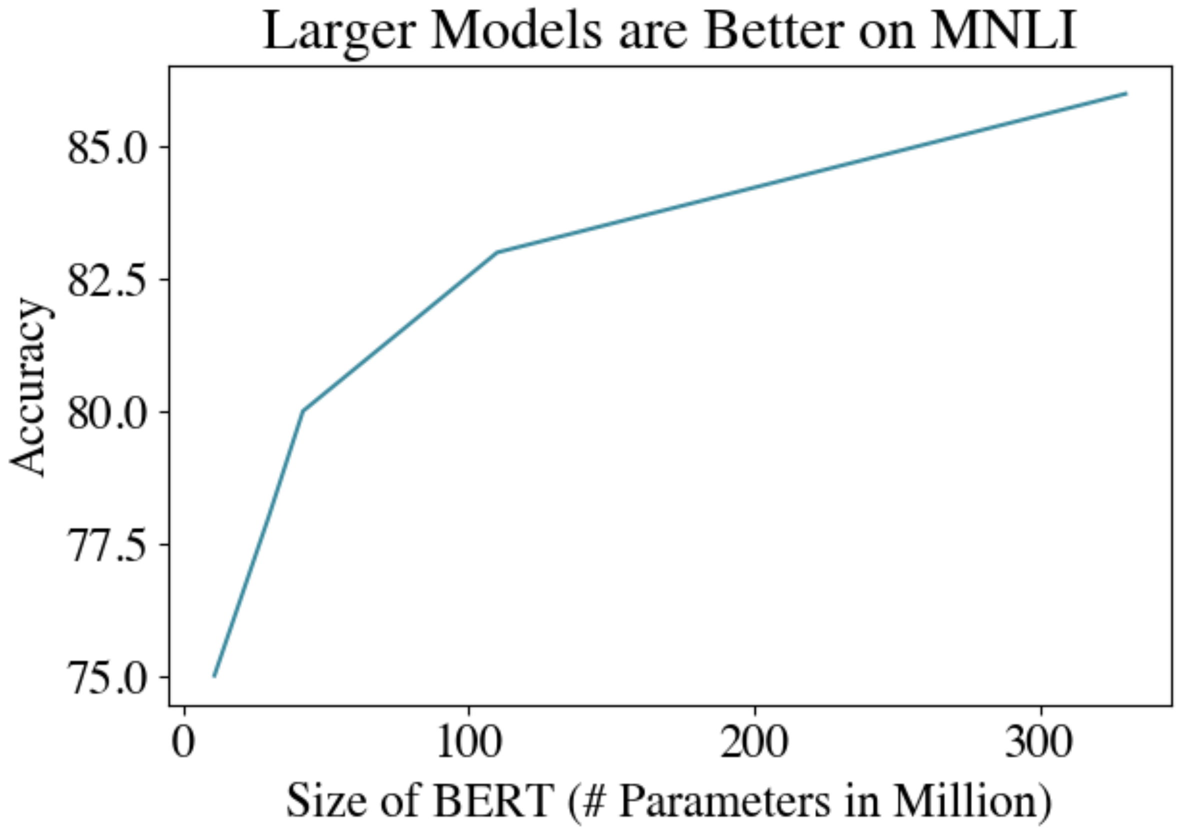
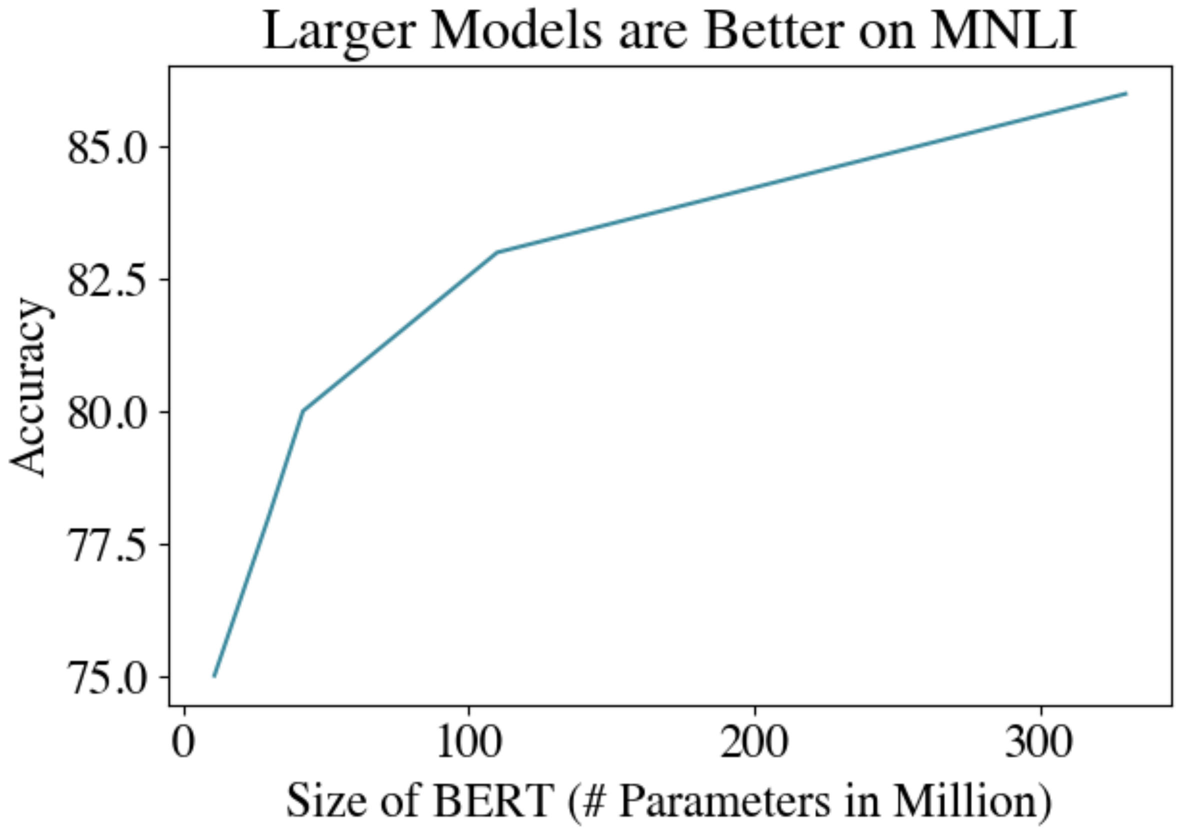
Are Larger Pre-trained Language Models Uniformly Better? Comparing Performance at the Datapoint Level

Ruiqi Zhong, Dhruba Ghosh, Dan Klein, and Jacob Steinhardt

Larger → Uniformly Better?



Larger → Uniformly Better?



How many datapoints are smaller models better at?

A Naïve Attempt

MNLI
Datapoint 1
Datapoint 2
Datapoint 3
Datapoint 4
Datapoint 5
…

A Naïve Attempt

MNLI	BERT-Base
Datapoint 1	✓
Datapoint 2	✓
Datapoint 3	X
Datapoint 4	✓
Datapoint 5	X
…	…
Accuracy	81.0%

A Naïve Attempt

MNLI	BERT-Base	BERT-Large
Datapoint 1	✓	X
Datapoint 2	✓	✓
Datapoint 3	X	✓
Datapoint 4	✓	✓
Datapoint 5	X	✓
…	…	…
Accuracy	81.0%	83.5%

A Naïve Attempt

MNLI	BERT-Base	BERT-Large	
Datapoint 1	✓	X	**4.5%**
Datapoint 2	✓	✓	
Datapoint 3	X	✓	
Datapoint 4	✓	✓	
Datapoint 5	X	✓	
…	…	…	
Accuracy	81.0%	83.5%	

Wait a Second …

MNLI	BERT-Base Seed 1	BERT-Base Seed 2	
Datapoint 1	✓	X	4.0%???
Datapoint 2	✓	✓	
Datapoint 3	X	✓	
Datapoint 4	✓	✓	
Datapoint 5	X	X	
…	…	…	
Accuracy	81.0%	81.2%	

Naïvely comparing models at the datapoint level is extremely noisy!

Denoising

Denoising

- New concept: “better” if more likely to be correct.

Denoising

- New concept: “better” if more likely to be correct.
- Denoise:

Denoising

- New concept: “better” if more likely to be correct.
- Denoise:
  - 10 pre-training seeds × 5 fine-tuning seeds

Denoising

- New concept: “better” if more likely to be correct.
- Denoise:
  - 10 pre-training seeds × 5 fine-tuning seeds
  - an easy and efficient statistical tool to upper-bound the noises

Findings

Findings

- BERT-Mini is better than BERT-Large on at least 1-4% datapoints across MNLI, SST-2, QQP

Findings

- BERT-Mini is better than BERT-Large on at least 1-4% datapoints across MNLI, SST-2, QQP
- Larger models have higher variance w.r.t. fine-tuning seeds.

Findings

- BERT-Mini is better than BERT-Large on at least 1-4% datapoints across MNLI, SST-2, QQP
- Larger models have higher variance w.r.t. fine-tuning seeds.
- (many others)

Check out Our Paper for …

Check out Our Paper for …

- New concepts & statistical tools for datapoint level understanding.

Check out Our Paper for …

- New concepts & statistical tools for datapoint level understanding.
- Predictions from > 500 models.

Check out Our Paper for …

- New concepts & statistical tools for datapoint level understanding.
- Predictions from > 500 models.
- Pre-trained models with different random seeds.