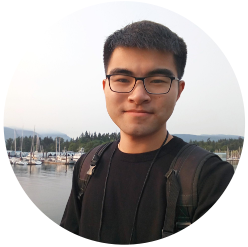
Probing Across Time:

What Does RoBERTa Know and When?

Leo Z. Liu*, Yizhong Wang*, Jungo Kasai, Hannaneh Hajishirzi, Noah A. Smith



GloVe

ELMo

GPT

BERT

RoBERTa

......

Leaderboard

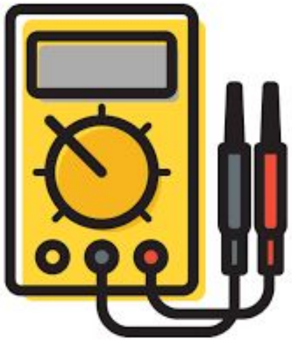
SQuAD2.0 tests the ability of a system to not only answer reading comprehension questions, but also abstain when presented with a question that cannot be answered based on the provided paragraph. How will your system compare to humans on this task?

Rank	Model	EM	F1
	Human Performance *Stanford University* (Rajpurkar & Jia et al. '18)	86.831	89.452
1 Mar 20, 2019	BERT + DAE + AoA (ensemble) *Joint Laboratory of HIT and iFLYTEK Research*	**87.147**	**89.474**
2 Mar 15, 2019	BERT + ConvLSTM + MTL + Verifier (ensemble) *Layer 6 AI*	86.730	89.286
3 Mar 05, 2019	BERT + N-Gram Masking + Synthetic Self-Training (ensemble) *Google AI Language* https://github.com/google-research/bert	86.673	89.147

Why?

Probes

Linguistic, factual,
commonsense, etc.

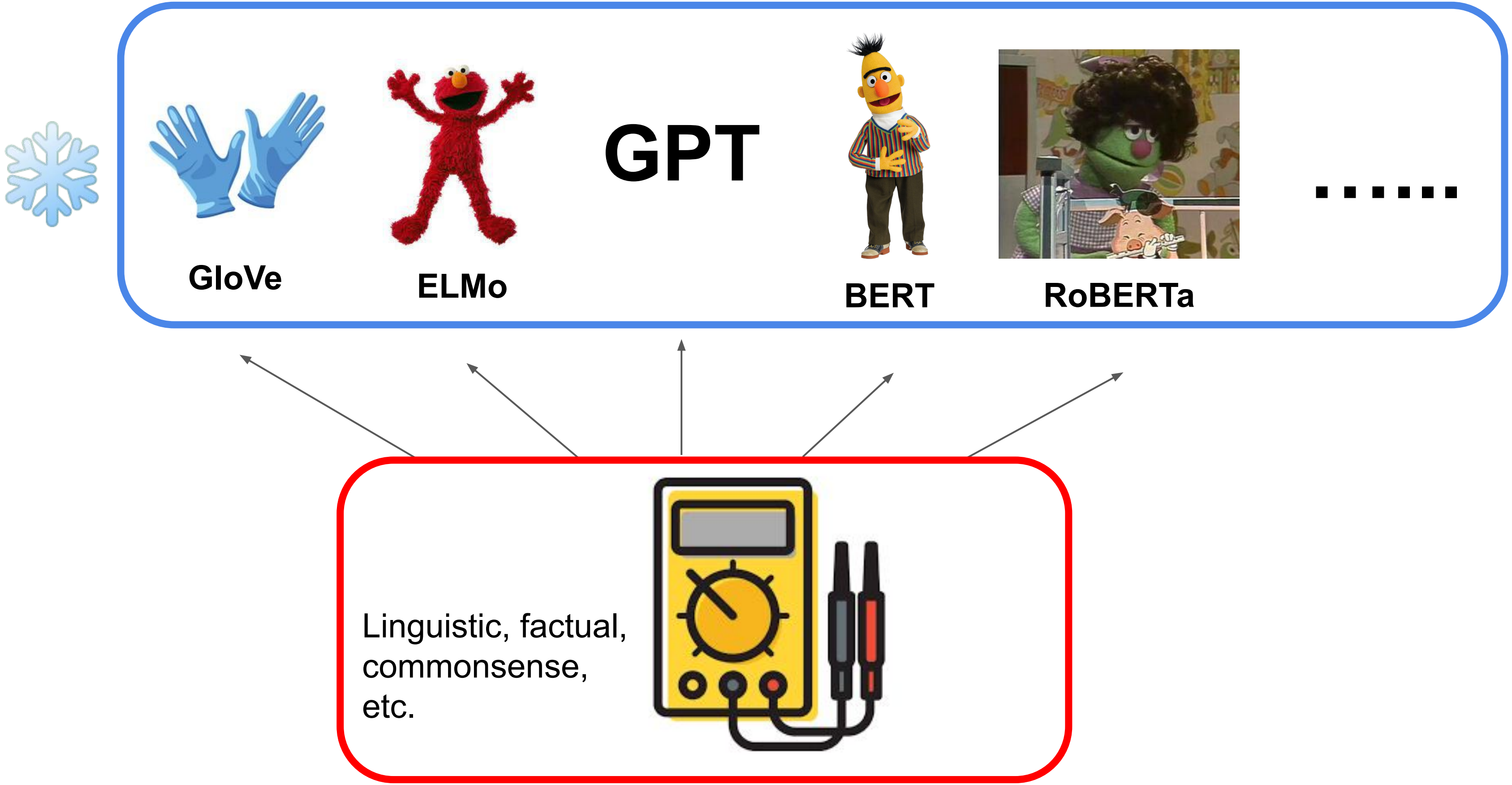


- Well motivated tests that encode and measure correspondence to human knowledge/intelligence (e.g. linguistic annotation, factual query, etc.)
- Better test score

  → better learned ability

  → better explain the **“why?”**

Current Probes



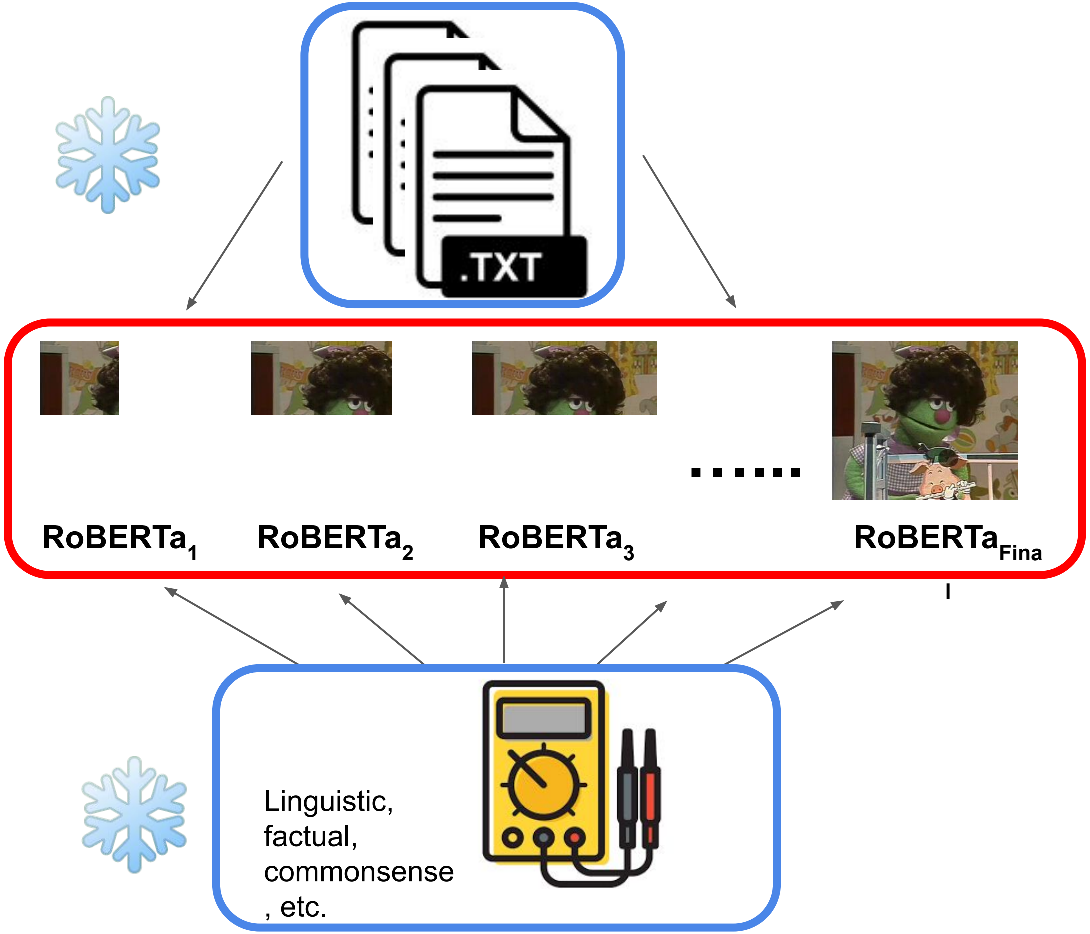
😄

- Compare models “shoulder-to-shoulder” by an interpretable metric

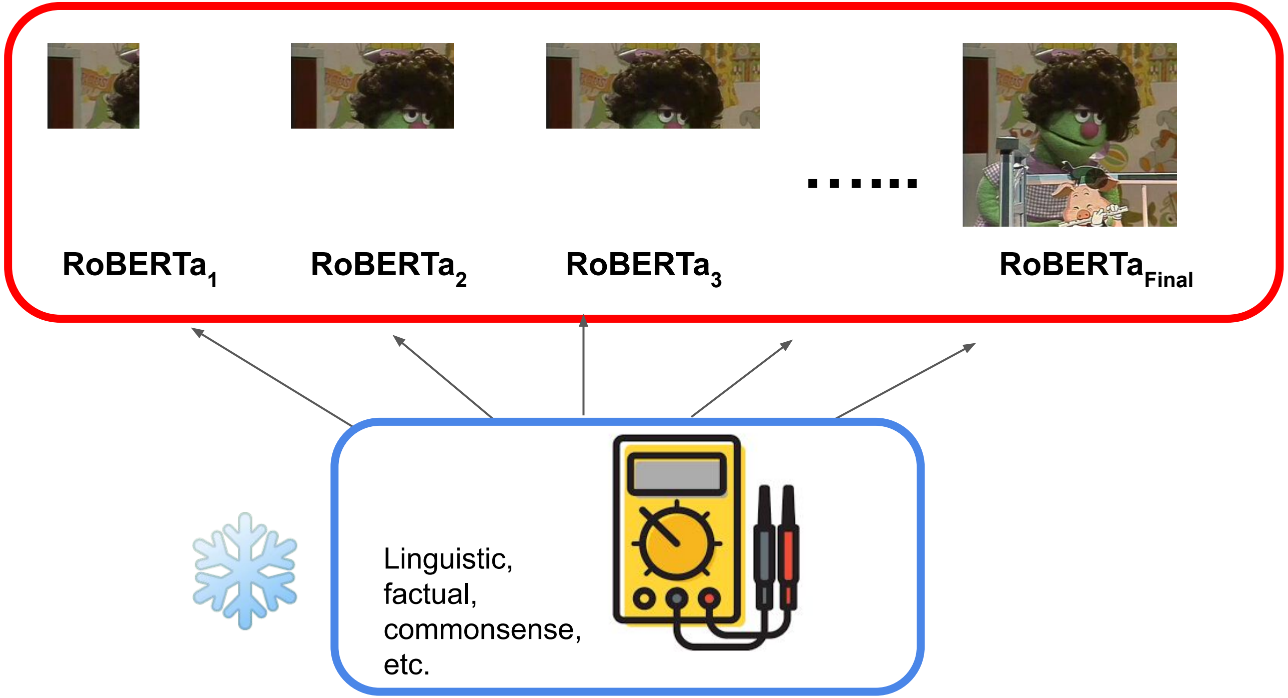
☹️

- Would model perform better in the next optimization update?
- How does the model learn?

Probe Across Time



Probing Across Time



😄

- Understand the underlying learning curriculum
- Longer observation increases our confidence in concluding how well model learns tested knowledge

Choice of probes

Diverse probed knowledge

- Linguistic
- Factual
- Commonsense
- Reasoning

Diverse probe formulation

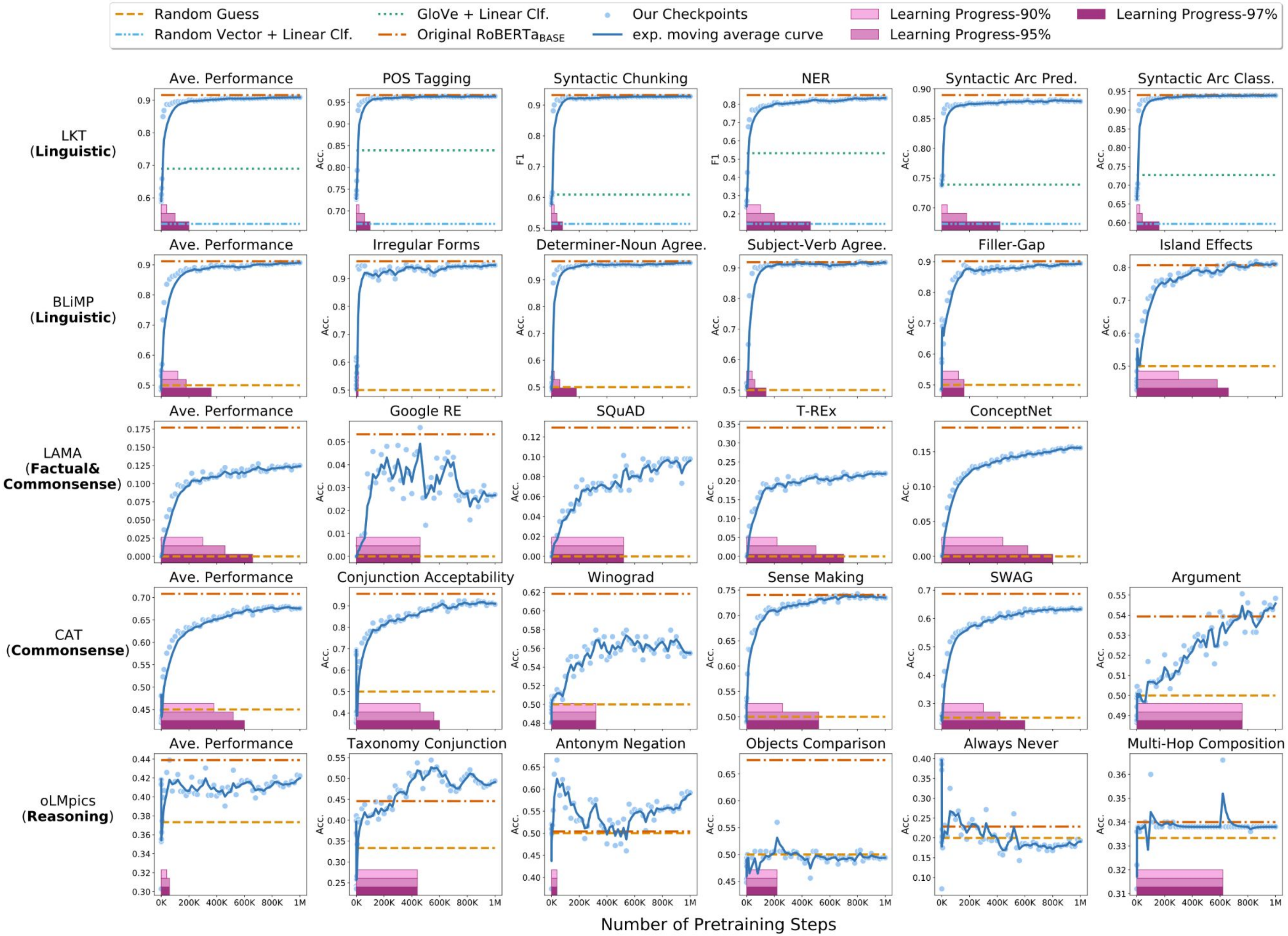
- (Contextual) embedding → $\mathbf{f}_{\text{linear}}$ → annotation
- Score(✅) > Score(❌), e.g. Score=perplexity of a sentence
- Pr(✅ | slot-filling query) > Pr(❌ | slot-filling query)
- A ferry and a floatplane are both a type of ***[MASK]***.

  ✅ vehicle ❌ airplane ❌ boat

Baseline*

- **Random Guess**: 1 / (# labels)
- **{Random, GloVe} Vector** → $\mathbf{f}_{\text{linear}}$ → annotation
- **Original RoBERTa** probes the officially released checkpoint of RoBERTa base to see if our checkpoints are pretrained properly and can achieve reasonable performance

* applicable to different types of probes



Learning curriculum

TL;DR

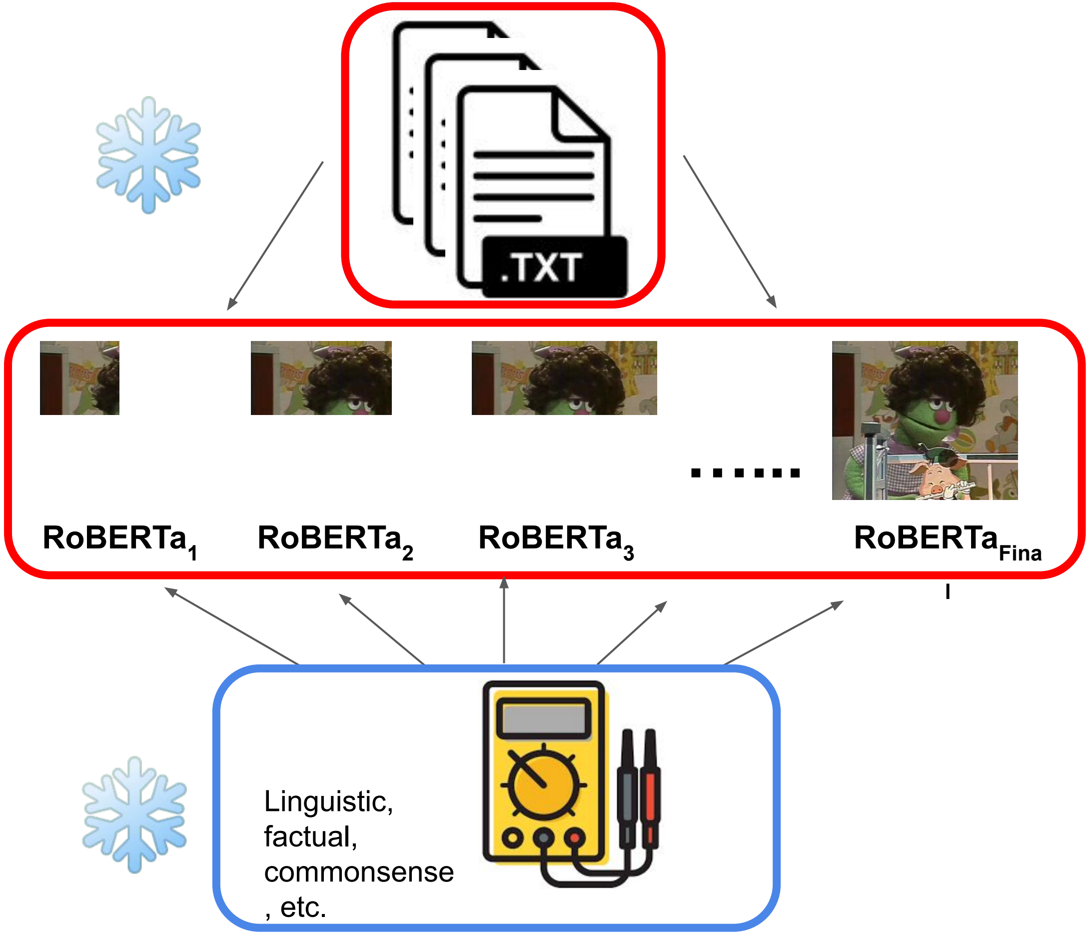
Linguistic (knowledge)

⌄

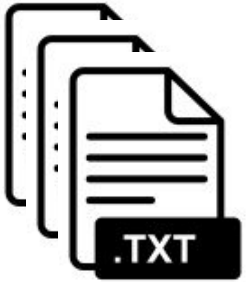
Factual ≅ Commonsense

Reasoning

In fact,
we didn’t mention…


.TXT
....
RoBERTa$_1$
RoBERTa$_2$
RoBERTa$_3$
RoBERTa$_{Fina}$
l
Linguistic,
factual,
commonsense
, etc.

Varying Pretraining Corpus

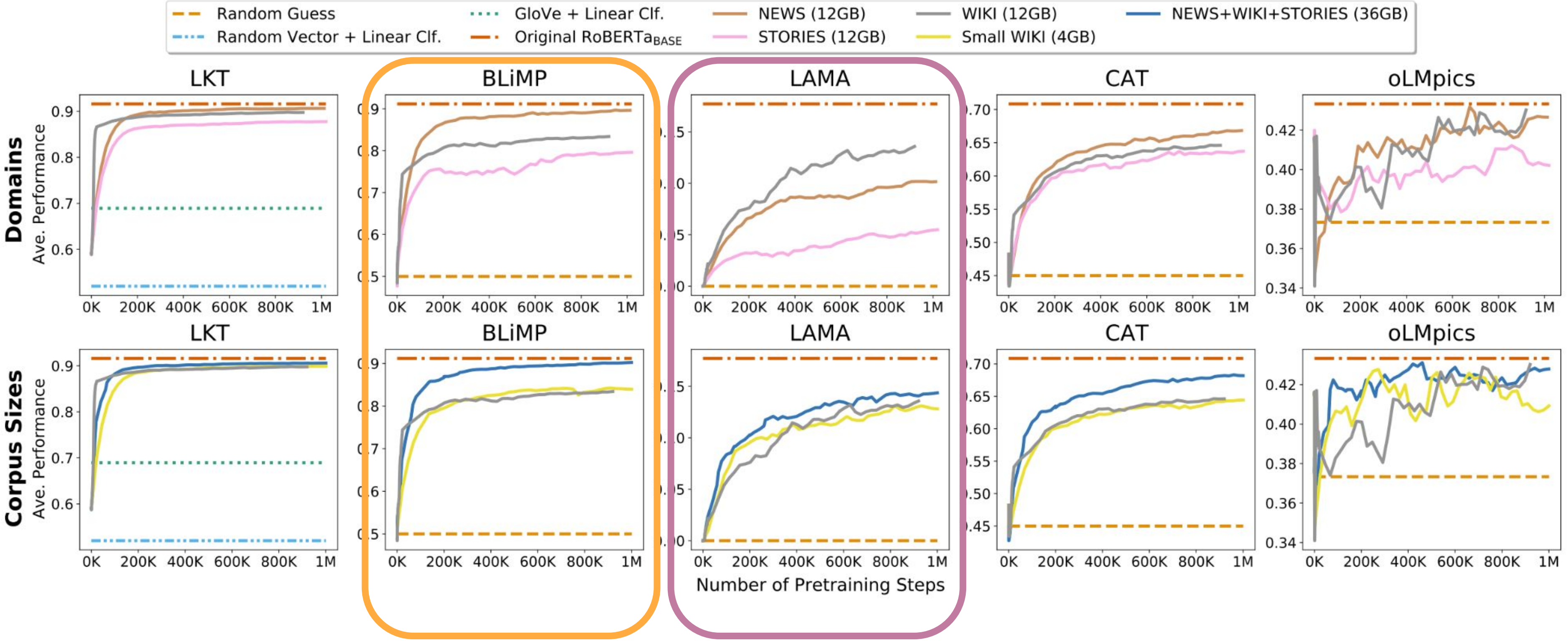


Domains:

- English WIKI (12 GB)
- NEWS (12 GB)
- STORIES (12 GB)

Corpus Size:

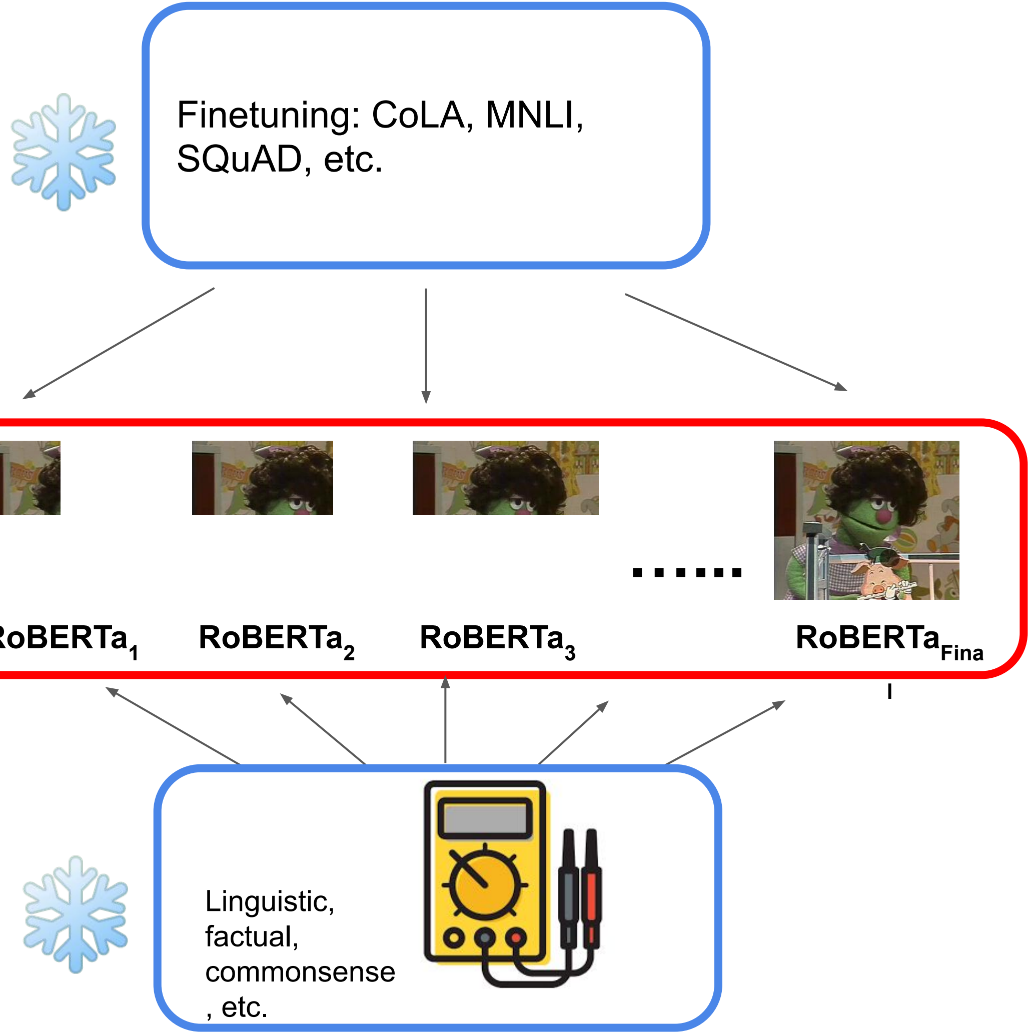
- Small English WIKI (4 GB)
- English WIKI (12 GB)
- English WIKI + NEWS + STORIES (36 GB)

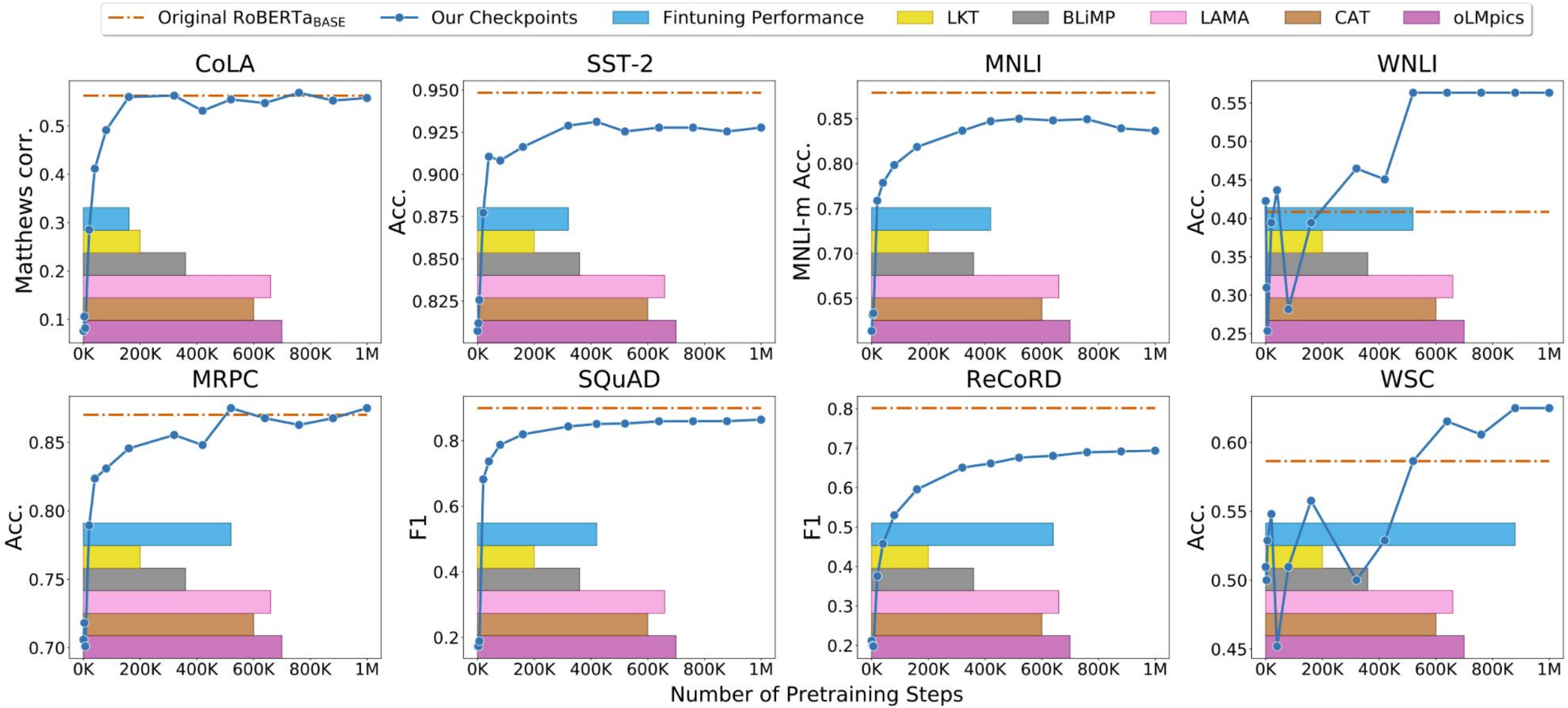


TL;DR:

- Observed learning curriculum remains the same
- Domains affect learning more than corpus sizes

Research Benchmarks





TL;DR:

Among finetuning tasks, **ordering of difficulties exists --** more knowledge required, more difficult

Main Contribution:

- Most systematic work of learning dynamics for pretraining yet
- Learning curriculum:

Linguistic (knowledge) $>$ Factual $\cong$ Commonsense $\gg$ Reasoning

- Domain diversity matters more than just corpus size
- Ordering of difficulties among downstream tasks
- As models evolve and new probes emerge, ***probing across time*** framework can serve as a general framework to inform progress on both fronts

Thanks!

Check our paper for more details and discussion!