Life after BERT: What do Other Muppets Understand about Language?

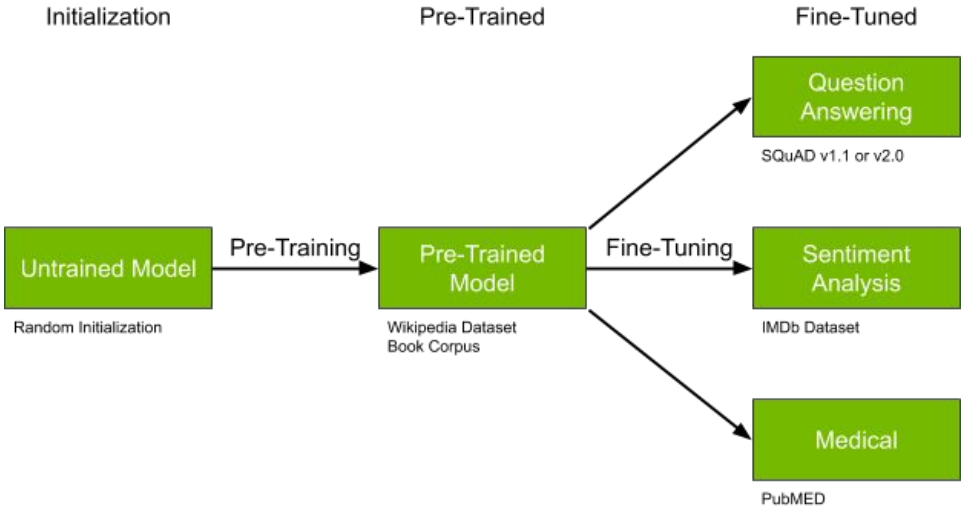
Kevin Zhao

Mentors: Vladislav Lialin, Namrata Shivagunde, Anna Rumshisky

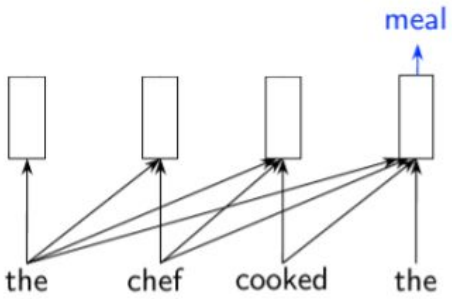
MIT PRIMES Spring Conference (5/22/22)

Pre-training

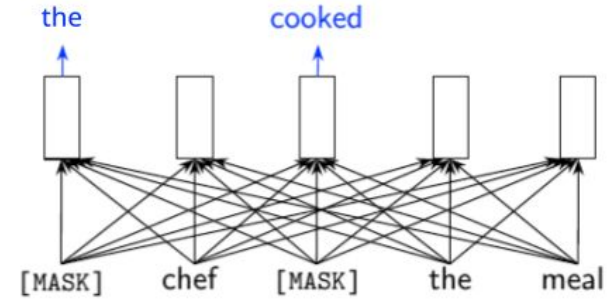
- Pre-train on large collection of texts
  - 10-100s GB of raw, unlabelled text
- Fine-tune on specific task
  - 1-100s MB of text, usually labeled



Language Modeling

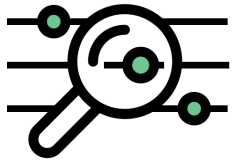


Masked Language Modeling



Model analysis

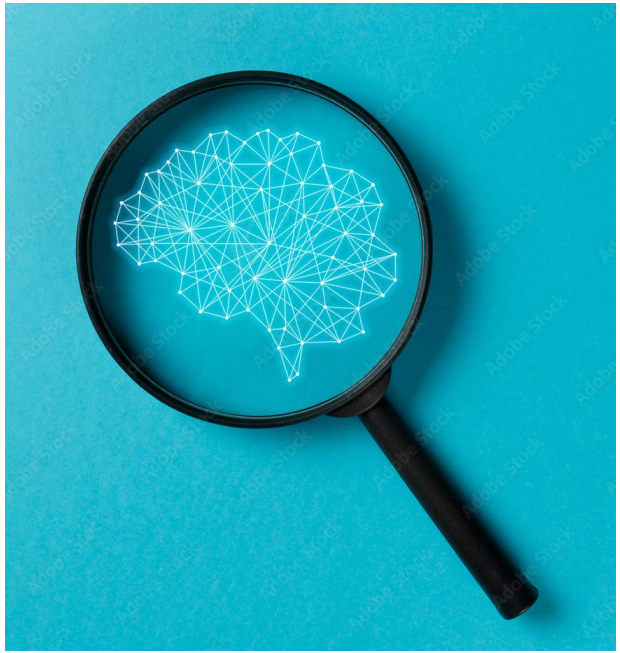
- NLU and NLG benchmarks allow us to **compare** models
- But they do not provide insight into **why** some models are better than others
- Probing tasks evaluate **specific** capabilities of pre-trained models
    - Often zero-shot (no fine-tuning)



Everything is BERT Nowadays

Out of **14** model analysis papers*

- **9** probe only BERT and/or RoBERTa
- **3** probe 2 different model families
- Only **2** papers probe 3 model families
- No paper probes more than 3

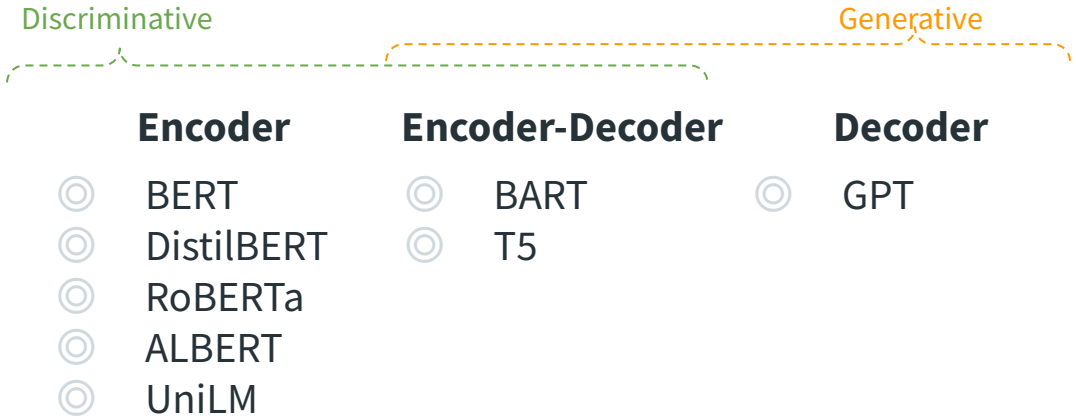


*this selection is not at all exhaustive and most probably skewed towards the papers we knew of in advance

Life After BERT

In our work:

- 8 model families
  - Encoder/Decoder/Seq2seq
  - 5 pre-training objectives
  - + a distilled model
- 29 models in total
- Varying sizes
  - Smallest: 66M
  - Largest: 2.8B



Evaluation framework: oLMpics🥇 + Psycholinguistic Eval👩‍💼

oLMpics	Object Comparison	The size of a **nail** is usually [MASK] than the size of a **fork**	**smaller**, larger
	Age Comparison	A **41** year old person age is [MASK] than a **42** year old person.	**younger**, older
	Antonym Negation	It was [MASK] a **fracture**, it was really a **break**	not, **really**
	Taxonomy Conjunction	A **ferry** and a **biplane** are both a type of [MASK]	airplane, **craft**, boat
	Always-Never	A chicken [MASK] has horns	**never**, rarely, sometimes, often, always
	Multi-hop Composition	When comparing a **21** year old, **15** year old, and **19** year old, the [MASK] is oldest.	**first**, second, third
Psycholinguistic	CPRAG	Justin put a **house on Park Place**. He and his sister spent hours **playing** [MASK]	**monopoly**, chess, baseball
	ROLE	the birthday **boy** saw which friend the **clown** had [MASK]	**scared**
	NEG	Salmon is **not** a [MASK]	fish, **dog**

🥇: Talmor et al, ACL 2020, arxiv.org/abs/1912.13283
👩‍💼: Ettinger, ACL 2020, arxiv.org/abs/1907.13528

Evaluation framework: oLMpics🥇 + Psycholinguistic Eval👩‍🔬

- Evaluate **specific** linguistic capabilities
- Zero-shot
  - Evaluate what model learns during pre-training
  - MLM or LM-based
- Left-to-right extension for MLM task
  - Compare the probabilities of a sentence with answer A vs sentence with answer B

$$p = \prod_{j=0}^{n} P(x_j \mid x_{<j})$$

	Age Comp.	Always Never	Object Comp.	Antonym Negation	Taxonomy Conj.	Multi-hop Comp.
Majority	50.6	36.1	50.6	50.2	34.0	34.0
$\text{BERT}_{\text{base}}$	49.4 ± 0.2	13.2 ± 1.2	55.4 ± 1.0	53.8 ± 1.0	46.8 ± 0.6	33.4 ± 0.6
$\text{BERT}_{\text{large}}$	50.6 ± 0.2	22.5 ± 1.3	52.4 ± 1.6	50.8 ± 0.8	**53.9 ± 0.9**	33.8 ± 0.7
$\text{BERT}_{\text{large}}$ **WWM**	76.4 ± 1.7	10.7 ± 1.5	55.8 ± 1.1	57.2 ± 0.7	46.4 ± 0.8	33.8 ± 0.7
$\text{RoBERTa}_{\text{large}}$	98.6 ± 0.1	13.5 ± 1.6	87.4 ± 0.9	**74.6 ± 0.8**	45.4 ± 0.4	28.0 ± 1.0
$\text{DistilBERT}_{\text{base}}$	49.4 ± 0.2	15.0 ± 1.2	51.0 ± 1.3	50.8 ± 0.7	46.8 ± 0.8	34.0 ± 1.0
$\text{DistilRoBERTa}_{\text{base}}$	45.4 ± 1.2	13.9 ± 1.3	50.8 ± 0.7	51.0 ± 1.0	50.6 ± 1.1	34.0 ± 1.0
$\text{AlBERT}_{\text{base}}$	47.0 ± 0.6	23.2 ± 1.2	50.6 ± 0.7	52.6 ± 1.0	-	34.0 ± 1.0
$\text{AlBERT}_{\text{large}}$	52.8 ± 1.2	30.7 ± 1.0	49.2 ± 0.7	50.2 ± 1.0	-	34.0 ± 1.0
$\text{AlBERT}_{\text{xlarge}}$	39.8 ± 0.3	26.1 ± 1.5	50.4 ± 0.8	44.6 ± 1.4	-	32.2 ± 1.2
$\text{AlBERT}_{\text{xxlarge}}$	95.4 ± 0.4	22.9 ± 0.5	61.0 ± 0.7	66.4 ± 0.5	-	34.0 ± 1.0
$\text{AlBERTv2}_{\text{base}}$	50.6 ± 0.2	21.4 ± 0.9	49.4 ± 0.7	54.2 ± 1.7	-	34.0 ± 1.0
$\text{AlBERTv2}_{\text{large}}$	51.4 ± 0.6	31.7 ± 1.5	50.6 ± 0.6	55.2 ± 1.3	-	34.0 ± 1.0
$\text{AlBERTv2}_{\text{xlarge}}$	46.2 ± 0.7	37.9 ± 1.9	50.6 ± 0.7	62.4 ± 0.9	-	32.4 ± 0.8
$\text{AlBERTv2}_{\text{xxlarge}}$	93.8 ± 0.5	23.9 ± 0.7	78.8 ± 0.8	64.8 ± 0.5	-	34.0 ± 1.0
$\text{BART}_{\text{large}}$	49.4 ± 0.2	23.2 ± 1.2	49.4 ± 0.7	49.8 ± 1.0	48.8 ± 0.9	33.8 ± 0.7
T5_{small}	49.4 ± 0.2	16.1 ± 1.6	48.2 ± 0.8	47.0 ± 0.9	49.3 ± 0.4	33.8 ± 0.7
T5_{base}	49.4 ± 0.2	10.7 ± 1.2	59.0 ± 0.7	53.4 ± 0.8	46.6 ± 0.9	33.6 ± 0.7
T5_{large}	94.0 ± 0.4	25.7 ± 0.7	83.2 ± 0.5	64.6 ± 1.4	42.2 ± 1.0	33.8 ± 0.7
T5_{xl}	**100.0 ± 0.0**	20.4 ± 1.0	**90.0 ± 0.5**	68.4 ± 0.8	41.2 ± 0.8	34.4 ± 0.6
$\text{T5v1.1}_{\text{small}}$	49.4 ± 0.2	34.3 ± 1.8	50.6 ± 0.7	51.4 ± 1.1	48.2 ± 0.7	**37.8 ± 0.9**
$\text{T5v1.1}_{\text{base}}$	50.6 ± 0.2	11.8 ± 1.6	56.0 ± 1.5	45.0 ± 0.8	49.9 ± 0.7	37.6 ± 0.9
$\text{T5v1.1}_{\text{large}}$	49.6 ± 0.3	15.7 ± 0.8	50.6 ± 0.8	47.1 ± 1.1	41.7 ± 1.0	33.8 ± 0.7
$\text{T5v1.1}_{\text{xl}}$	49.4 ± 0.2	23.9 ± 1.8	49.4 ± 0.7	54.2 ± 1.2	**53.9 ± 0.5**	33.8 ± 0.7
$\text{UniLM}_{\text{base}}$	47.9±1.6	16.1±0.8	48.0±2.7	43.6±1.3	45.1±1.2	34.8±0.9
$\text{UniLM}_{\text{large}}$	47.9±1.6	19.9±1.3	61.4±1.8	51.2±1.4	50.2±2.1	33.6±0.7
$\text{GPT2}_{\text{base}-0.1\text{B}}$	47.6±1.2	**50.1±1.5**	50.1±1	52.8±1.9	48.4±1.0	32.2±2.4
$\text{GPT2}_{\text{medium}-0.3\text{B}}$	50.1±1.3	40.8±2.2	49.6±0.9	54.7±2.4	49.1±1.7	29.6±2.1
$\text{GPT2}_{\text{large}-0.8\text{B}}$	69.6±1.0	20.2±1.7	50.4±1.0	50.1±2.7	46.9±1.5	33.5±1.3
$\text{GPT}_{\text{NEO}-1.3\text{B}}$	58.6±0.7	29.0±1.0	52.1±0.7	65.2±1.1	50.6±1.5	33.3±1.0
$\text{GPT2}_{\text{xl}-1.5\text{B}}$	51.9±1.5	26.6±0.7	52.6±0.7	60.6±1.2	45.8±1.3	34.0±1.0

Findings

Language models are not **universal** multitask learners

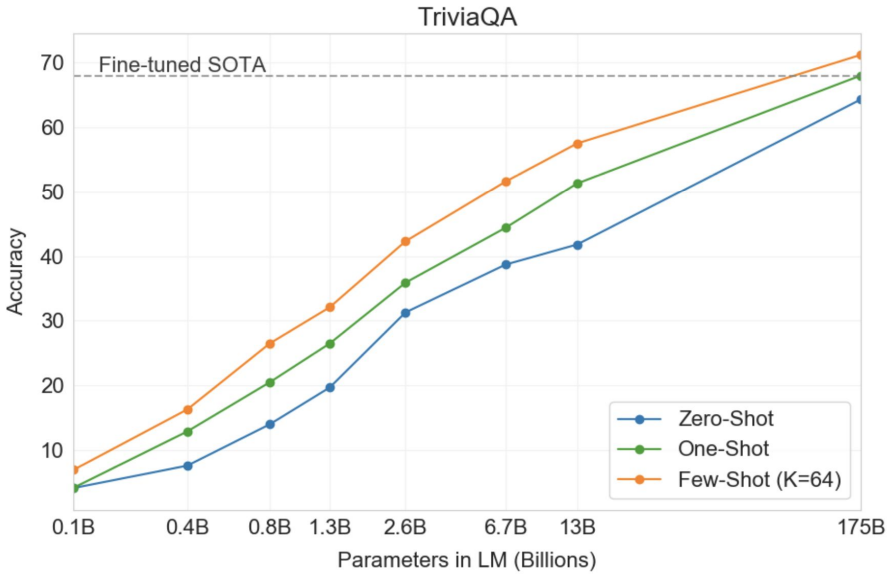
No model, out of 29, significantly outperforms majority baselines for these tasks

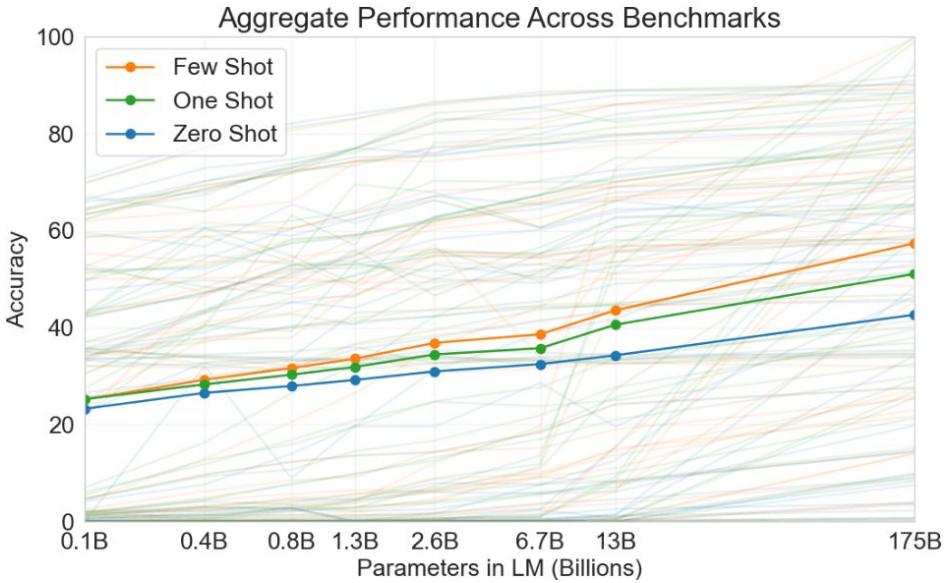
	Majority baseline	BERT base-large	RoBERTa large	DistilBERT	ALBERT base-XXL	BART large	T5 small-XL	UniLM base-large	GPT base-XL
Always-Never	36.1	13.3 - 22.5	13.5	15.0	15.0 - **37.9**	14.3	10.7 - 34.3	15.5 - 19.2	9.0 - 31.3
Multi-hop Composition	34.0	33.2 - 33.8	28.0	33.4	14.0 - 34.0	33.8	33.8 - **37.8**	33.3 - 34.9	32.6 - 34.0

Always-Never: "A chicken [MASK] has horns"

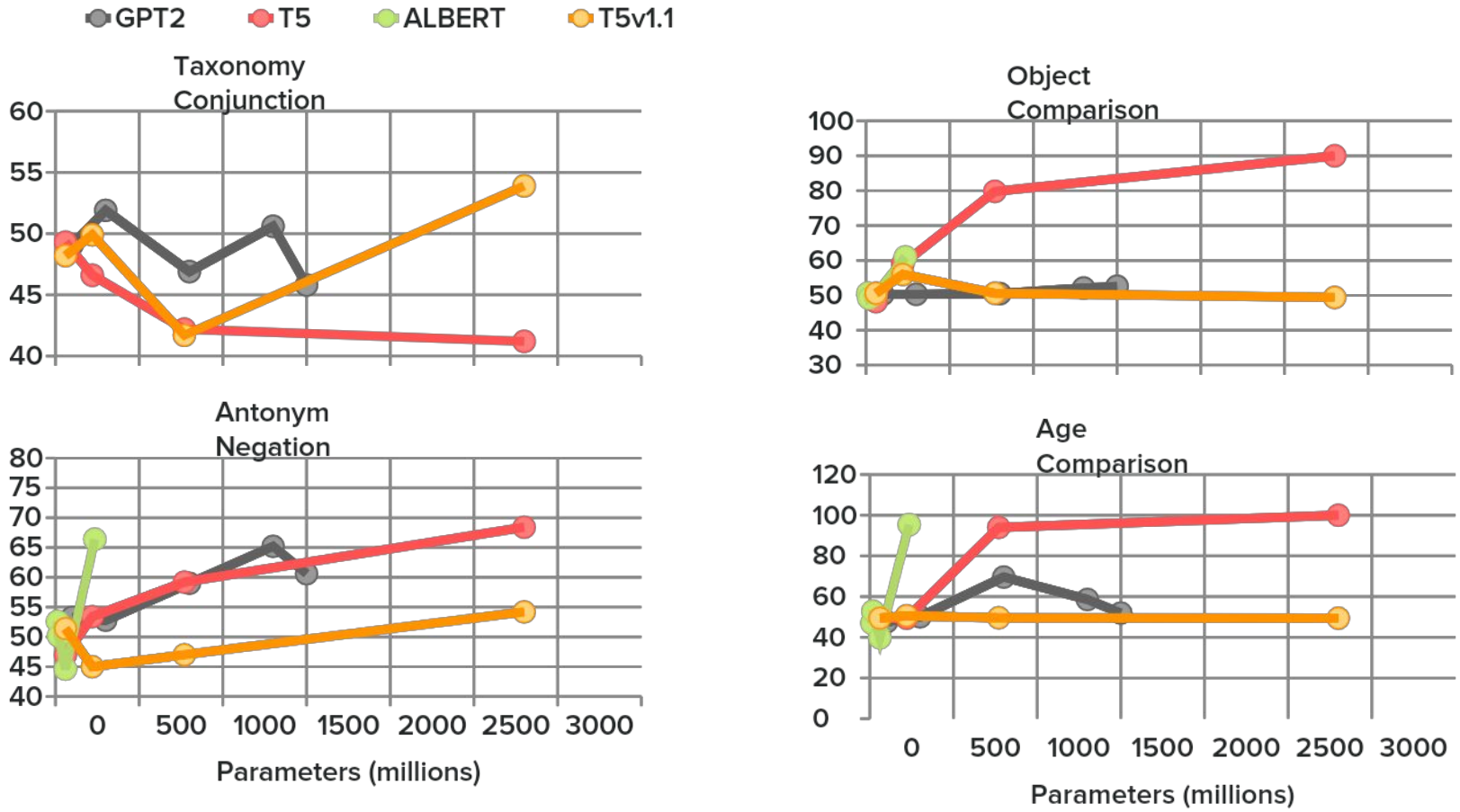
Multi-hop Composition: "When comparing a 23, a 38 and a 31 year old, the [MASK] is oldest"

Scale in NLP


TriviaQA
Fine-tuned SOTA
70
60
50
40
30
20
10
Accuracy
Zero-Shot
One-Shot
Few-Shot (K=64)
0.1B
0.4B
0.8B
1.3B
2.6B
6.7B
13B
175B
Parameters in LM (Billions)


Aggregate Performance Across Benchmarks
100
80
60
40
20
0
Accuracy
Few Shot
One Shot
Zero Shot
0.1B
0.4B
0.8B
1.3B
2.6B
6.7B
13B
175B
Parameters in LM (Billions)

Bigger does not mean **better**



Model properties are not predictive of model linguistic capabilities

In general, none of

- Parameter count
- Pre-training dataset size
- Model architecture
- Directionality
- Vocabulary size
- ...

seem to matter...?

	Params	Objective	Data	Architecture	oLMpics score
BERT Large	340M	MLM + NSP	16Gb	Enc	44.0
RoBERTa Large	355M	MLM	160Gb	Enc	57.9
BART Large	406M	Sentence restoration	160Gb	Enc-Dec	46.9
T5 XL	**2.8B**	Seq2seq MLM + Sup	**750Gb**	Enc-Dec	**59.1**
ALBERT XXL	235M	MLM + SOP	16Gb	Enc	**59.1**
T51.1 XL	**2.8B**	Seq2seq MLM	**750Gb**	Enc-Dec	44.1

Some models **are** sensitive to negation

A robin is a... bird

A robin is not a... bird

	Majority baseline	BERT large	RoBERTa large	T5 XL	ALBERT XXL	GPT2 M (0.3B)	GPT2 XL (1.5B)	GPT NEO (1.3B)
Antonym Negation	50	51.0	**74.4**	68.4	66.4	52.8	60.06	65.2
NEG-LNAT (AFF/NEG)	50/50	**75.0**/0.0	**75.0**/12.5	37.5/50	**75.0**/12.5	50/**62.5**	62.5/37.5	65.2/25.0

ALBERT changes its predictions on NEG-LNAT after negation in only 5% of cases
RoBERTa in **33%**

Conclusion

- Not much evidence of improved linguistic capabilities of pre-trained models in the last few years
- Model size, architecture, pre-training objective, dataset size, etc. don't seem to be predictive of linguistic capabilities
- Slight differences in optimization or masking strategy might be more important

Acknowledgements

Thanks to:

- My mentors Vladislav Lialin and Anna Rumshisky for their guidance and support
- Namrata Shivagunde for performing the GPT2 and UniLM experiments and helping adapt oLMpics to autoregressive models
- Dr. Gerovitch and the PRIMES program for giving me this research opportunity
- My parents

Questions?