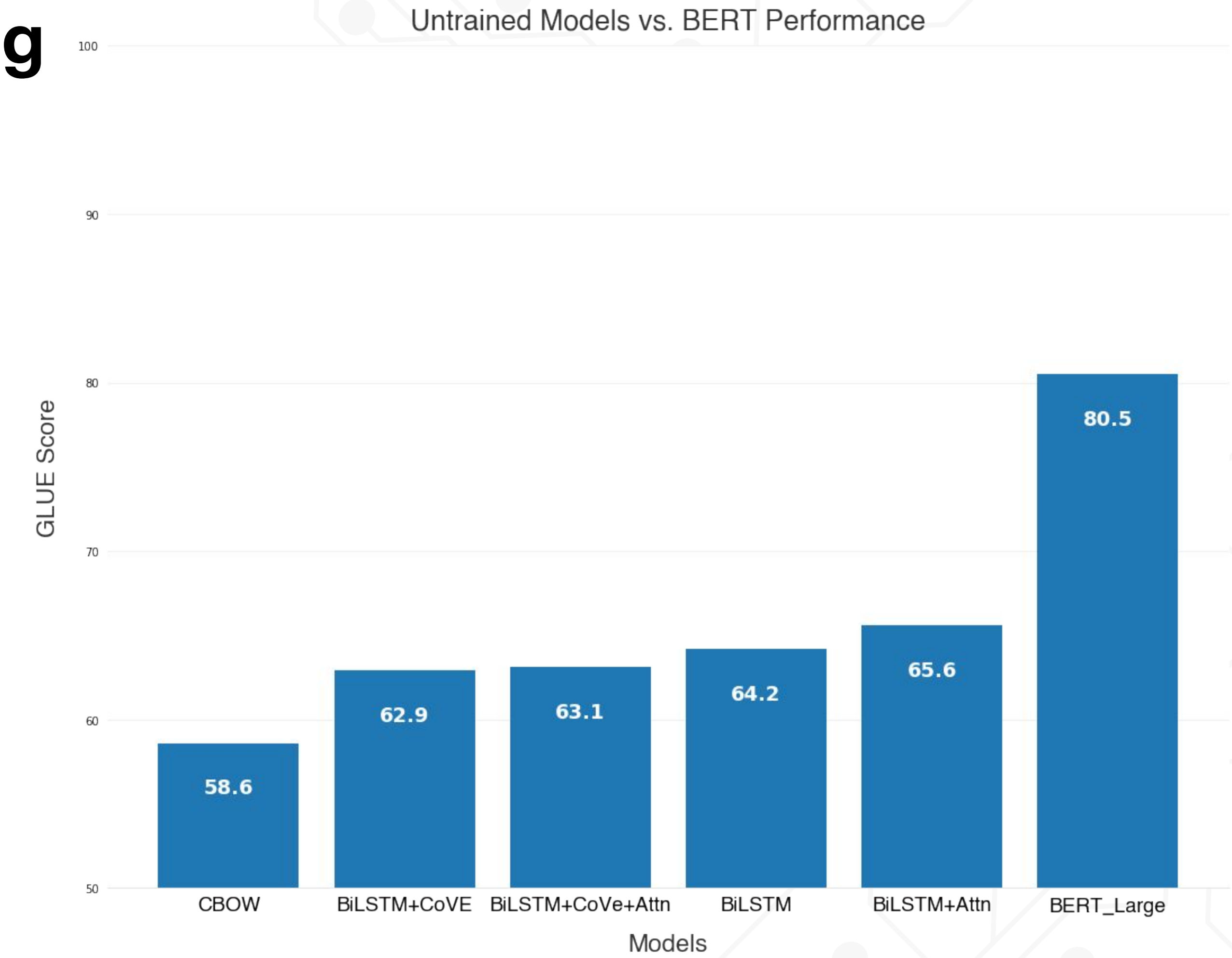
More than BERT: oLMpics on diverse language models

Kevin Zhao

Mentored by Vladislav Lialin, Namratar Shivagunde, Anna Rumshisky

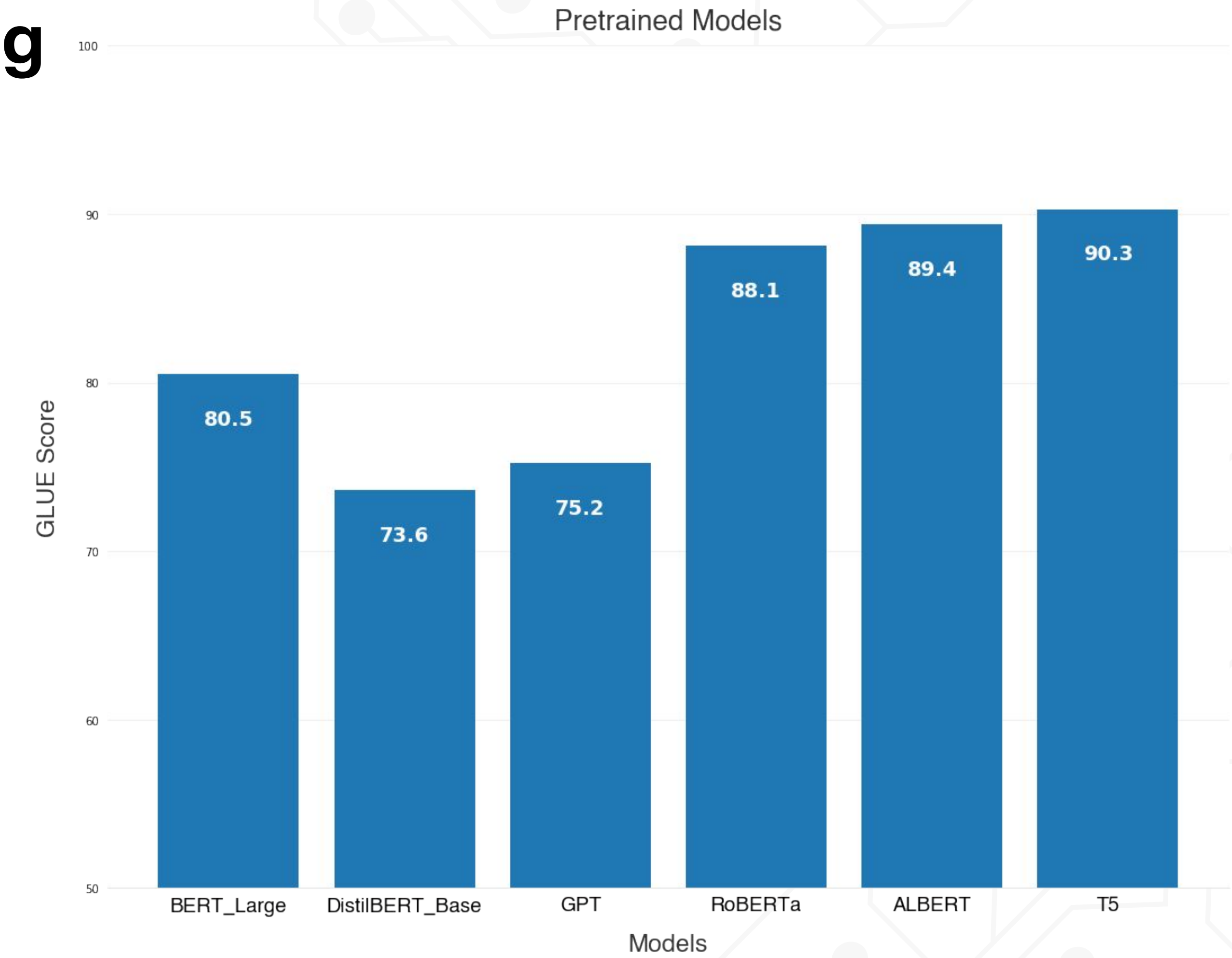
Transfer Learning

- Pre-BERT: most of the model is trained from scratch
- BERT: pre-trained on vast amounts of generic text



Transfer Learning

- Pre-BERT: most of the model is trained from scratch
- BERT: pre-trained on vast amounts of generic text
- Post-BERT: pre-training is one of the pillars of NLP
- The number of pre-training methods has rocketed



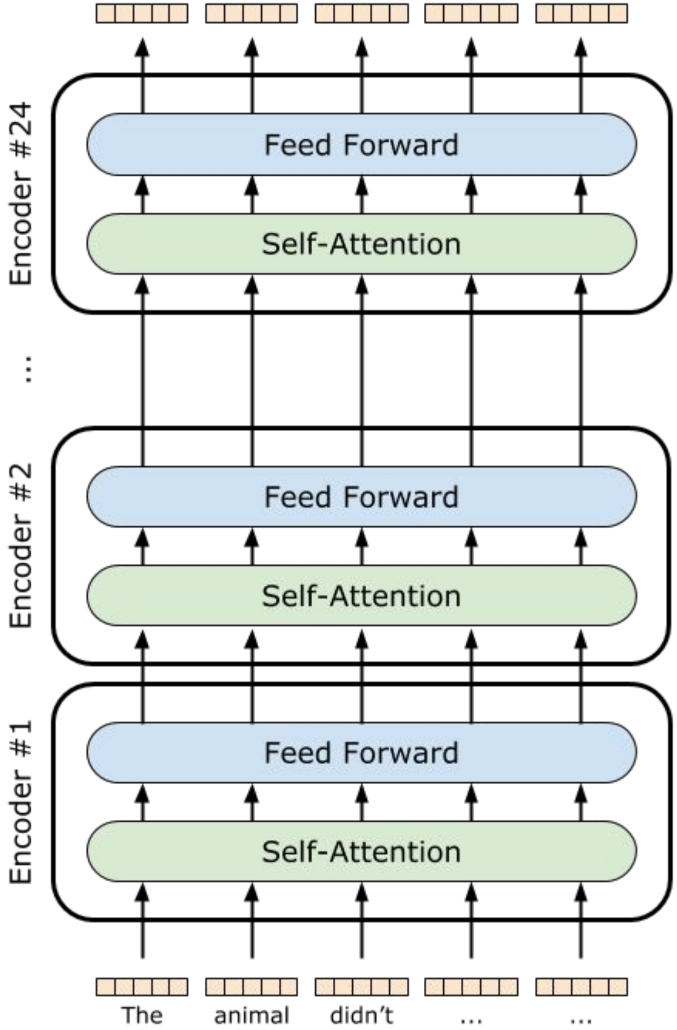
Outline

- Transformers
  - Architecture and attention
  - Models
- oLMpics
  - Overview
  - Evaluation methods
  - Task results
- Attention
  - Attention norms
  - Patterns
- Conclusion

Transformers

- Architecture
- Attention
  - Tokens interact directly
  - Query, Key, Value
  - Multiple heads

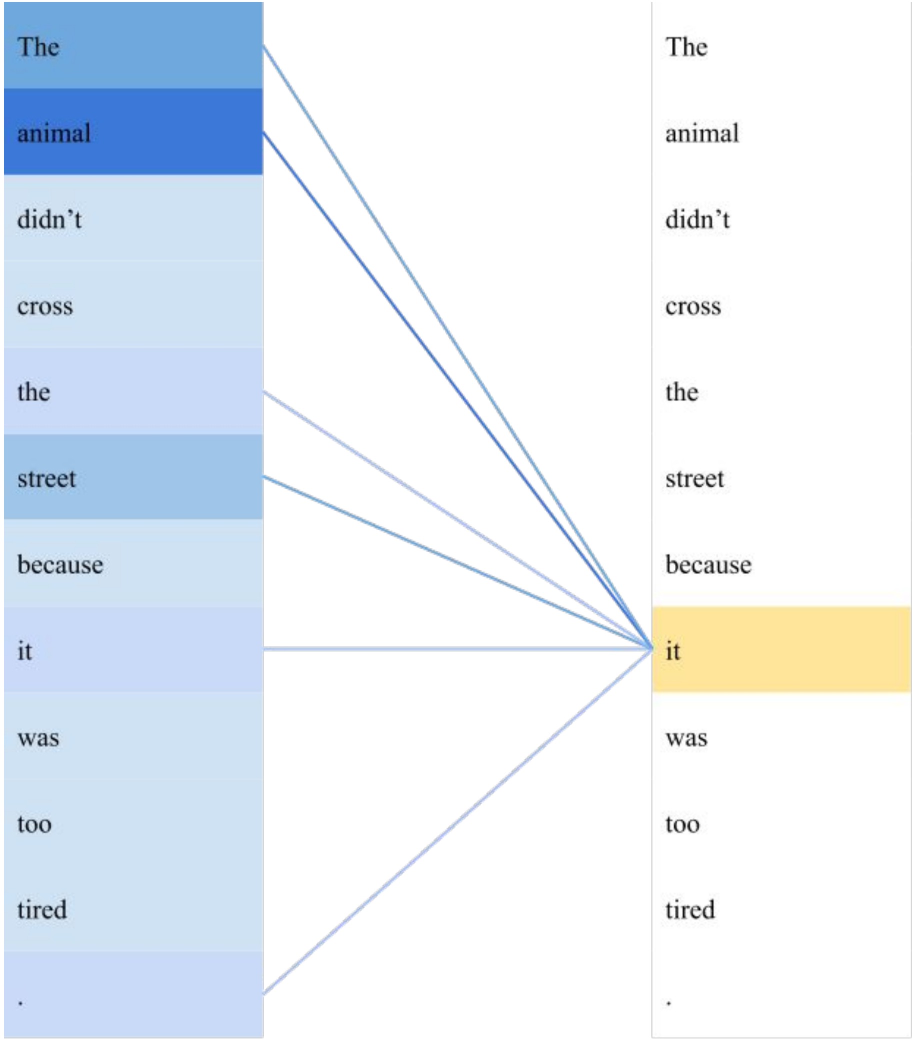
$$\mathrm{Attention}(Q,K,V)=\mathrm{softmax}(\frac{QK^T}{\sqrt{d_k}})V$$



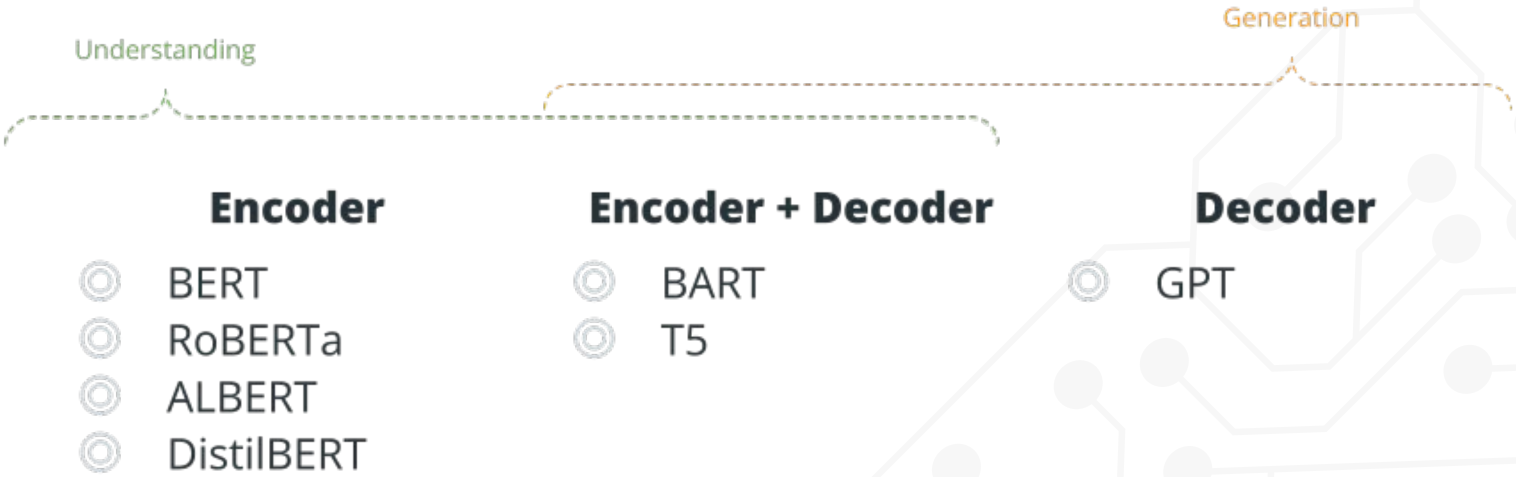
Transformers

- Architecture
- Attention
  - Tokens interact directly
  - Query, Key, Value
  - Multiple heads

$$\mathrm{Attention}(Q,K,V) = \mathrm{softmax}(\frac{QK^T}{\sqrt{d_k}})V$$



Transformer Models



Understanding

Generation

Encoder

- BERT
- RoBERTa
- ALBERT
- DistilBERT

Encoder + Decoder

- BART
- T5

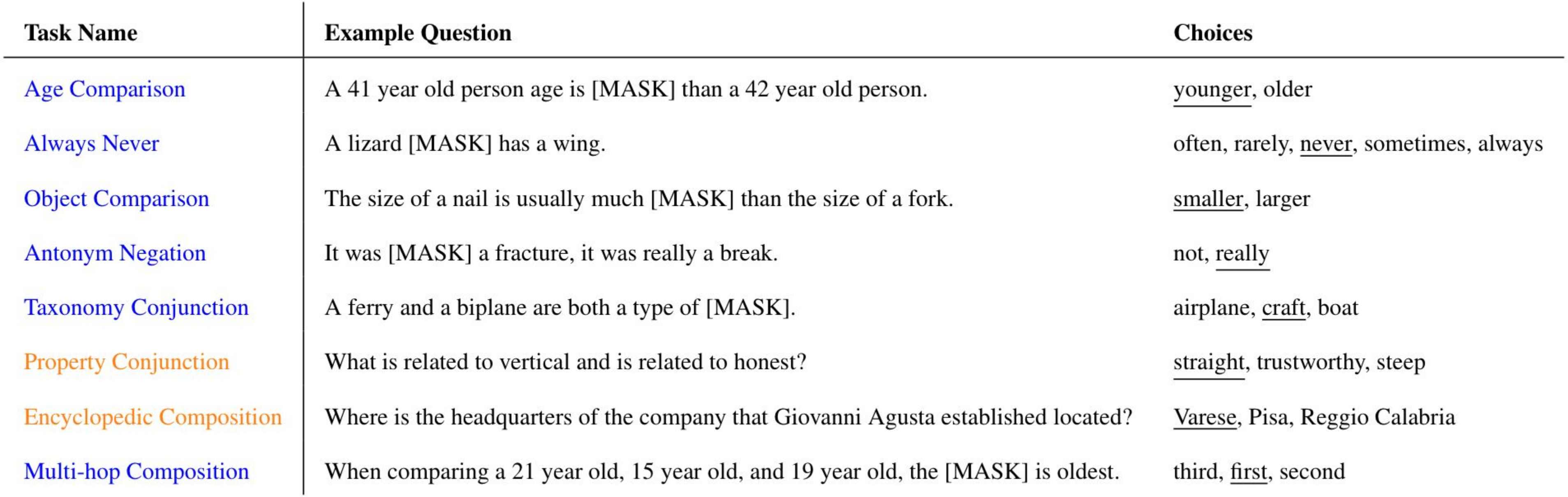
Decoder

- GPT

Differences:

- Architecture
- Size
- Pre-training objective
- Pre-training data

oLMpics Overview

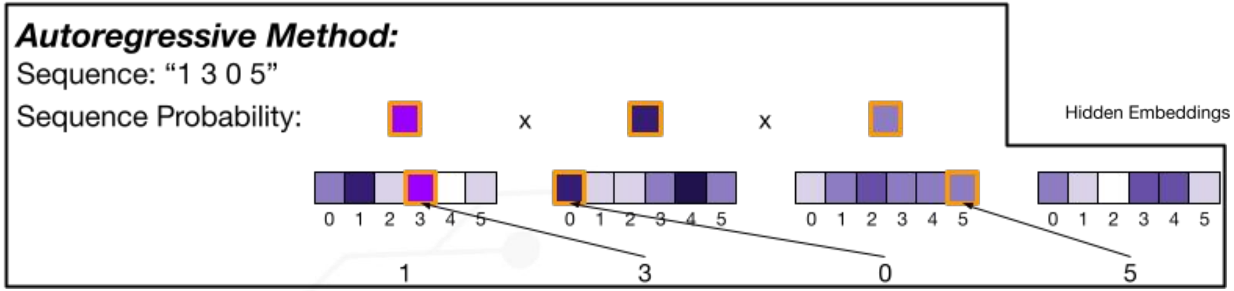


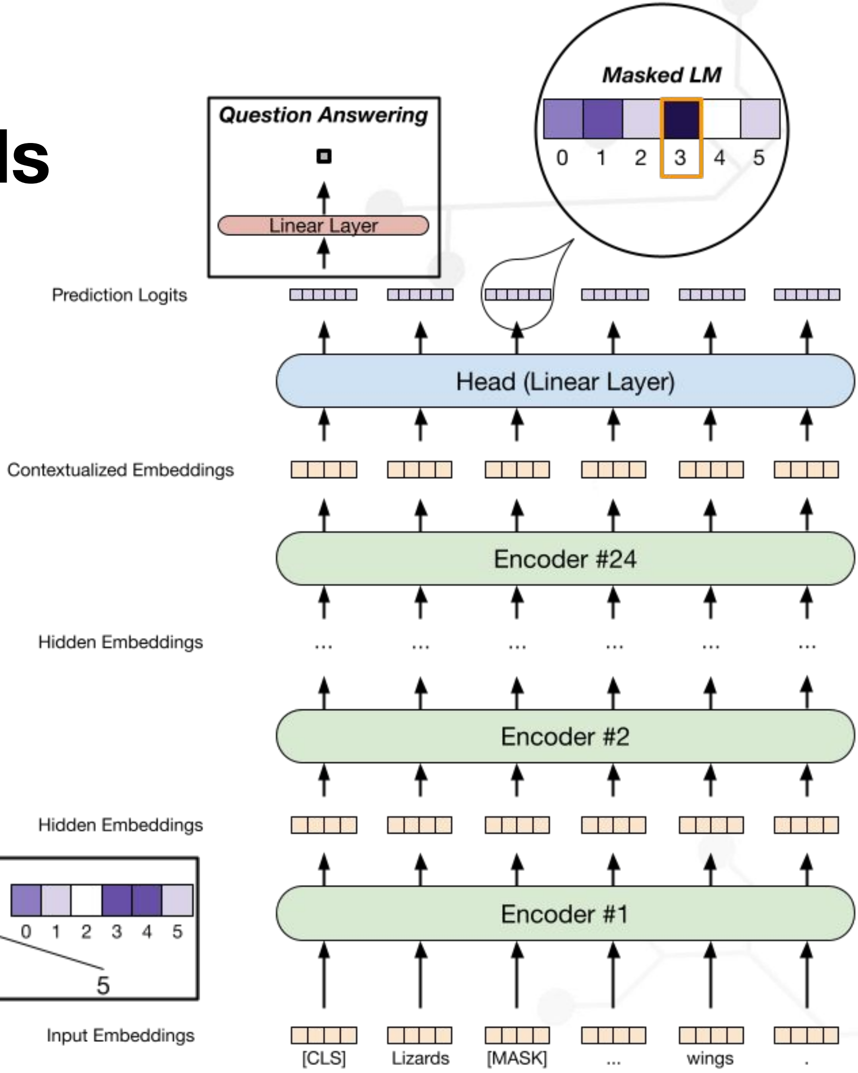
Task Name	Example Question	Choices
Age Comparison	A 41 year old person age is [MASK] than a 42 year old person.	younger, older
Always Never	A lizard [MASK] has a wing.	often, rarely, never, sometimes, always
Object Comparison	The size of a nail is usually much [MASK] than the size of a fork.	smaller, larger
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
Property Conjunction	What is related to vertical and is related to honest?	straight, trustworthy, steep
Encyclopedic Composition	Where is the headquarters of the company that Giovanni Agusta established located?	Varese, Pisa, Reggio Calabria
Multi-hop Composition	When comparing a 21 year old, 15 year old, and 19 year old, the [MASK] is oldest.	third, first, second

Blue: MLM
Orange: QA

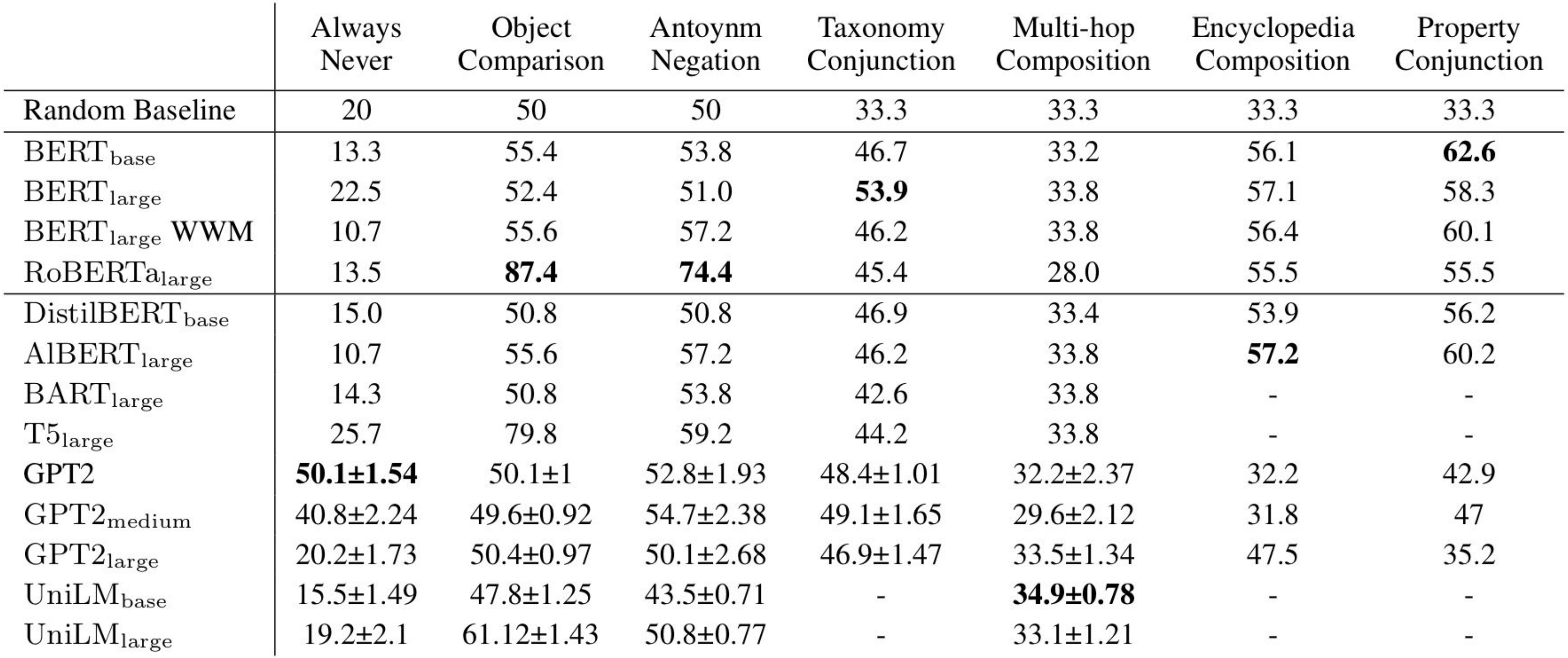
oLMpics Evaluation Methods

- MLM
- QA
- Modification for GPT2





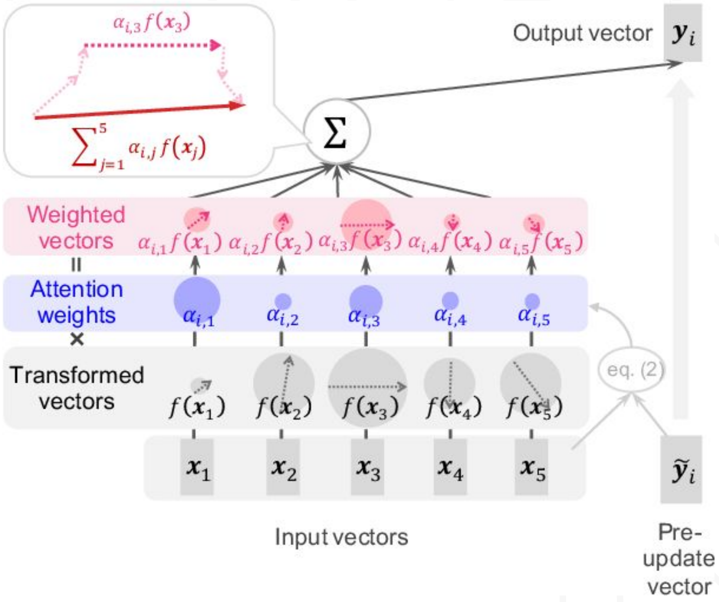
oLMpics Results



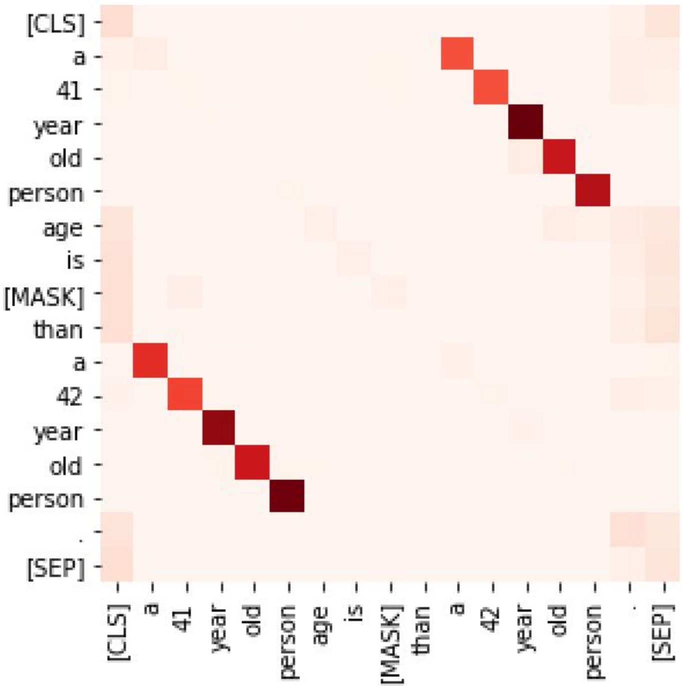
	Always Never	Object Comparison	Antoynm Negation	Taxonomy Conjunction	Multi-hop Composition	Encyclopedia Composition	Property Conjunction
Random Baseline	20	50	50	33.3	33.3	33.3	33.3
BERT$_{\text{base}}$	13.3	55.4	53.8	46.7	33.2	56.1	**62.6**
BERT$_{\text{large}}$	22.5	52.4	51.0	**53.9**	33.8	57.1	58.3
BERT$_{\text{large}}$ WWM	10.7	55.6	57.2	46.2	33.8	56.4	60.1
RoBERTa$_{\text{large}}$	13.5	**87.4**	**74.4**	45.4	28.0	55.5	55.5
DistilBERT$_{\text{base}}$	15.0	50.8	50.8	46.9	33.4	53.9	56.2
AlBERT$_{\text{large}}$	10.7	55.6	57.2	46.2	33.8	**57.2**	60.2
BART$_{\text{large}}$	14.3	50.8	53.8	42.6	33.8	-	-
T5$_{\text{large}}$	25.7	79.8	59.2	44.2	33.8	-	-
GPT2	**50.1±1.54**	50.1±1	52.8±1.93	48.4±1.01	32.2±2.37	32.2	42.9
GPT2$_{\text{medium}}$	40.8±2.24	49.6±0.92	54.7±2.38	49.1±1.65	29.6±2.12	31.8	47
GPT2$_{\text{large}}$	20.2±1.73	50.4±0.97	50.1±2.68	46.9±1.47	33.5±1.34	47.5	35.2
UniLM$_{\text{base}}$	15.5±1.49	47.8±1.25	43.5±0.71	-	**34.9±0.78**	-	-
UniLM$_{\text{large}}$	19.2±2.1	61.12±1.43	50.8±0.77	-	33.1±1.21	-	-

Attention Norms

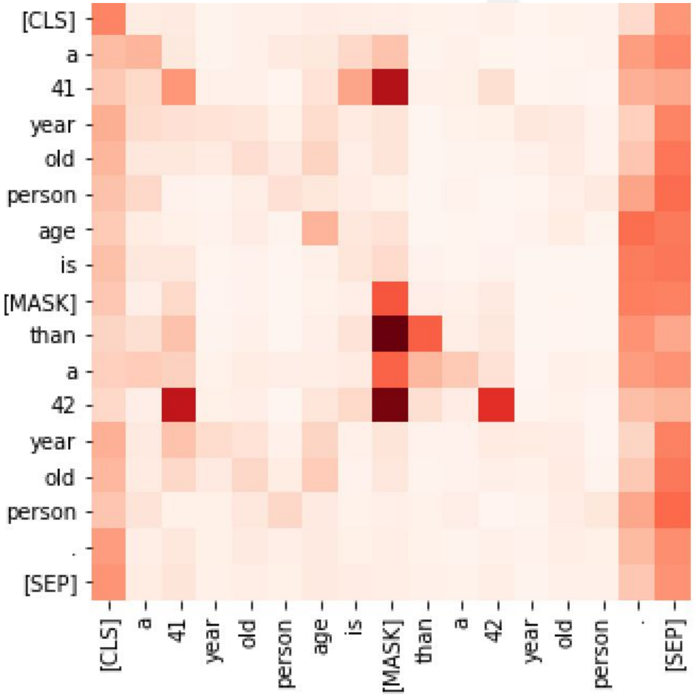
- Attention weights can be useful in understanding what a model looks at
- However, more recently attention norms have been shown to be more accurate
  - Attention formula can be rearranged
  - The norm of this product between the attention weights and transformed value vectors is the “attention norm”



Attention Norm Patterns



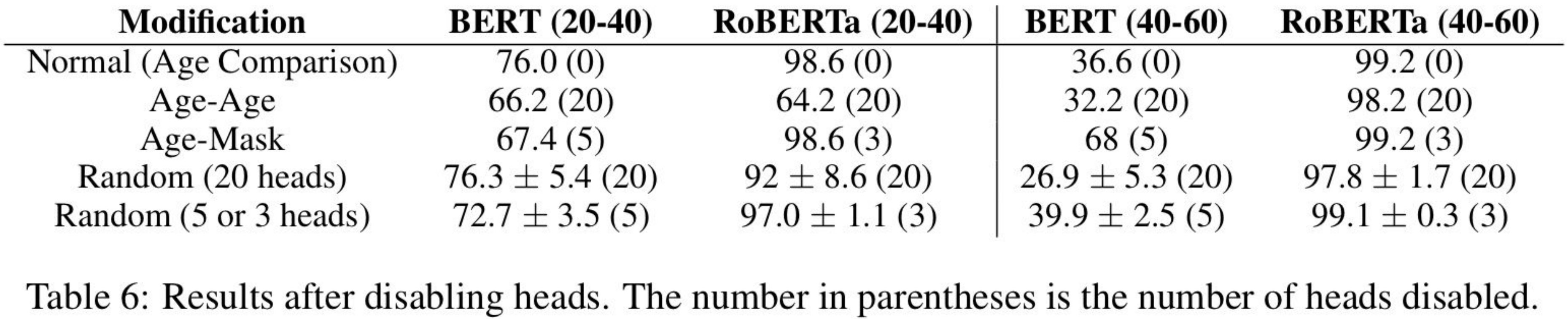
Age-Age Pattern



Age-MASK Pattern

Age-Age and Age-MASK Importance

To determine whether heads are important, we compare the effect of disabling the heads to disabling the same amount of random heads



Modification	BERT (20-40)	RoBERTa (20-40)	BERT (40-60)	RoBERTa (40-60)
Normal (Age Comparison)	76.0 (0)	98.6 (0)	36.6 (0)	99.2 (0)
Age-Age	66.2 (20)	64.2 (20)	32.2 (20)	98.2 (20)
Age-Mask	67.4 (5)	98.6 (3)	68 (5)	99.2 (3)
Random (20 heads)	76.3 ± 5.4 (20)	92 ± 8.6 (20)	26.9 ± 5.3 (20)	97.8 ± 1.7 (20)
Random (5 or 3 heads)	72.7 ± 3.5 (5)	97.0 ± 1.1 (3)	39.9 ± 2.5 (5)	99.1 ± 0.3 (3)

Table 6: Results after disabling heads. The number in parentheses is the number of heads disabled.

Conclusion

- We analyzed the differences between pre-trained models
    - Zero-shot evaluation on oLMpics tasks
        - Different models perform well on different tasks, there’s no clear leader
        - None of the models can solve composition task
    - Hidden representation analysis - attention norms
        - Intuitive features like Age-MASK do not contribute to performance
- Adapted oLMpics zero-shot setup for autoregressive models

Acknowledgements

Thanks to:

- My mentors Vladislav Lialin and Dr. Rumshisky for their guidance and support
- Namratar Shivagunde for doing the GPT and UniLM experiments, as well as helping develop oLMpics for autoregressive models
- Dr. Gerovitch and the PRIMES program for giving me this research opportunity
- My parents

Questions?