Google Research

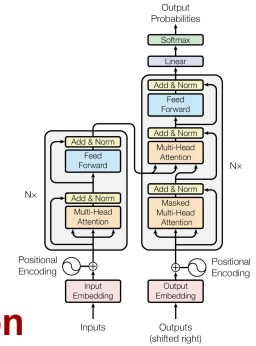
Is scale all we need?

Slav Petrov
on behalf of many wonderful
colleagues at Google Research

slav@google.com

Google Research

Number of Synapses in Biological & Artificial Systems



NMT w/ Attention
Resnet50
[25-50M]

Transformer
[400M]

Facebook ResNeXt101
Open AI GPT2
MSR ZeRO
NVidia Megatron-LM
Google-T5
[1-10B]

GPT3
M4
[80-600B]

#synapses

10^{6}> 10^{9} 10^{12} 10^{13} 10^{14} 10^{15}

Fruit fly **Honey bee** **Mouse** **Cat** **Macaque** **Human**

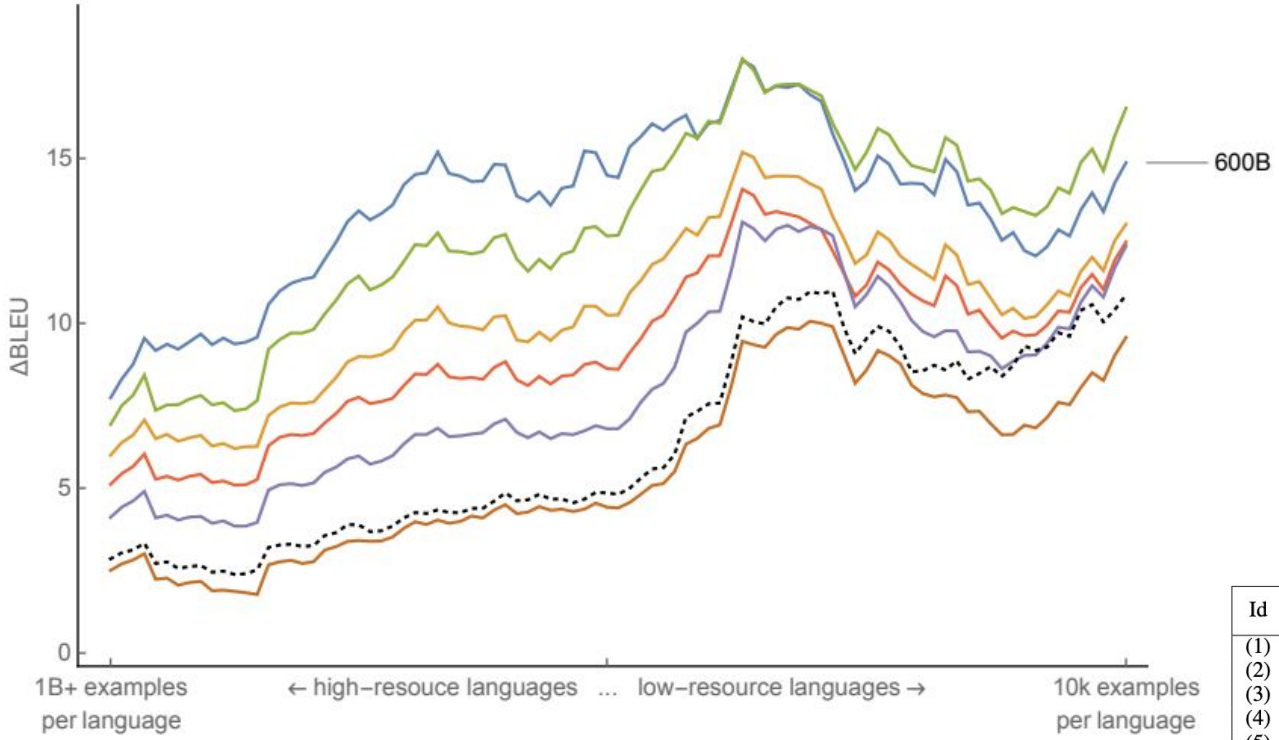




List of animals by number of neurons (Wikipedia)

Google Research

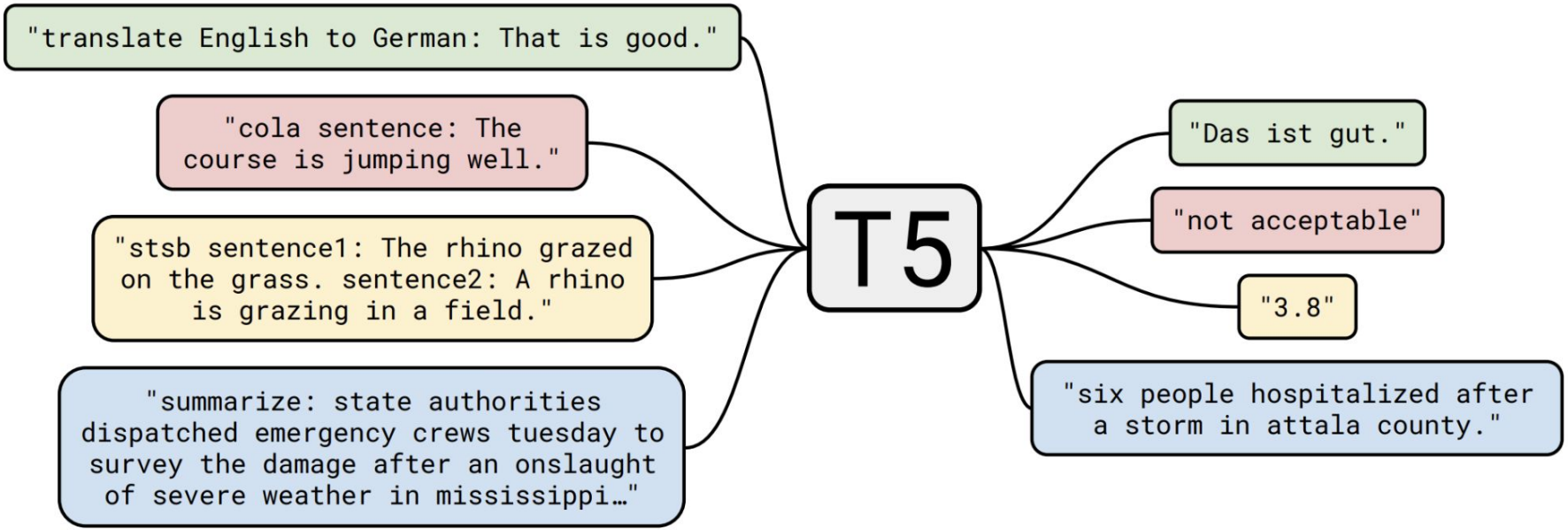
Massive Neural Machine Translation



Id	Model	BLEU avg.	ΔBLEU avg.	Weights
(1)	MoE(2048E, 36L)	**44.3**	13.5	600B
(2)	MoE(2048E, 12L)	41.3	10.5	200B
(3)	MoE(512E, 36L)	43.7	12.9	150B
(4)	MoE(512E, 12L)	40.0	9.2	50B
(5)	MoE(128E, 36L)	39.0	8.2	37B
(6)	MoE(128E, 12L)	36.7	5.9	12.5B
*	T(96L)	36.9	6.1	2.3B
*	Baselines	30.8	-	100×0.4B

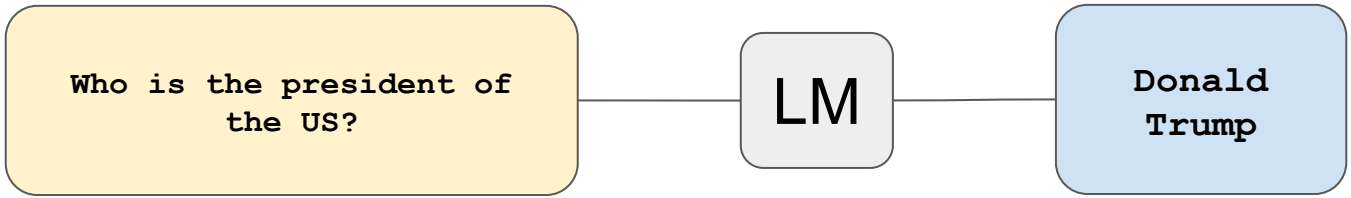
"GShard: Scaling Giant Models with Conditional Computation and Automatic Sharding" ICLR '21
D. Lepikhin, H. Lee, Y. Xu, D. Chen, O. Firat, Y. Huang, M. Krikun, N. Shazeer, Z. Chen

Are large language models enough?



"Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer," Arxiv '19
Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, Peter J. Liu

But what do we do when something goes wrong?



This talk is about aspects that won't be solved by scale alone:

- **Memory**
- **Attribution**
- **Generation**
- **Time**
- **Tools**
- **Trust**

From "implicit" to "explicit" memory

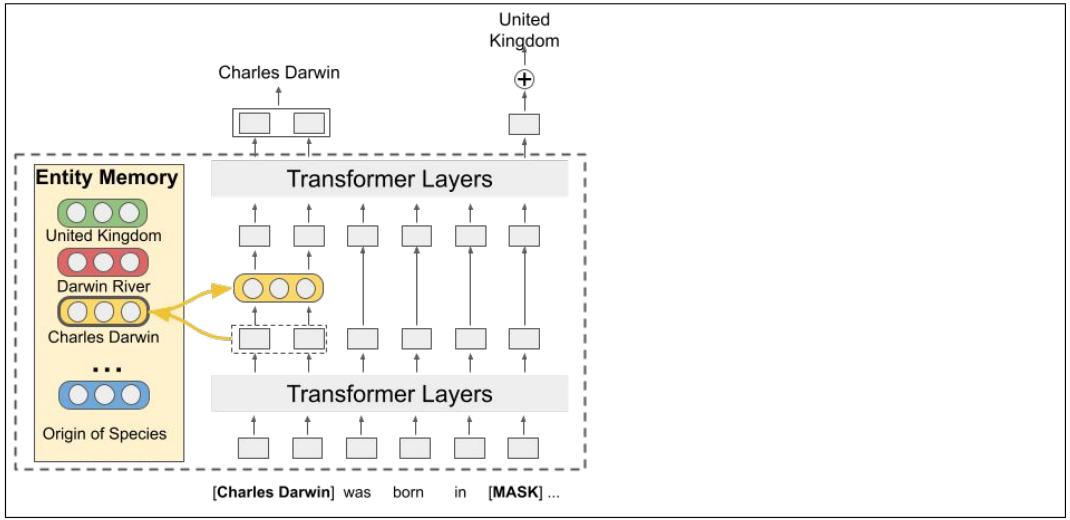
Most current LMs are not grounded / do not attribute their output.
The knowledge is "implicitly" somewhere in the parameters.

- **Attribution to sources**
  - **Structured knowledge:** knowledge graph entities and relations
  - **Unstructured knowledge:** documents, images, video
  - **Training examples:** these can also be treated as "memories"

- **Benefits**
  - Provenance provides interpretability & trustworthiness
  - Greater memorization capacity
  - Greater efficiency, thanks to sparse memory access
  - Generalization by controlling memories

Entities as Experts - Learn Entity Memories

Matches T5 performance on Trivia QA with 1% of parameters used per example.



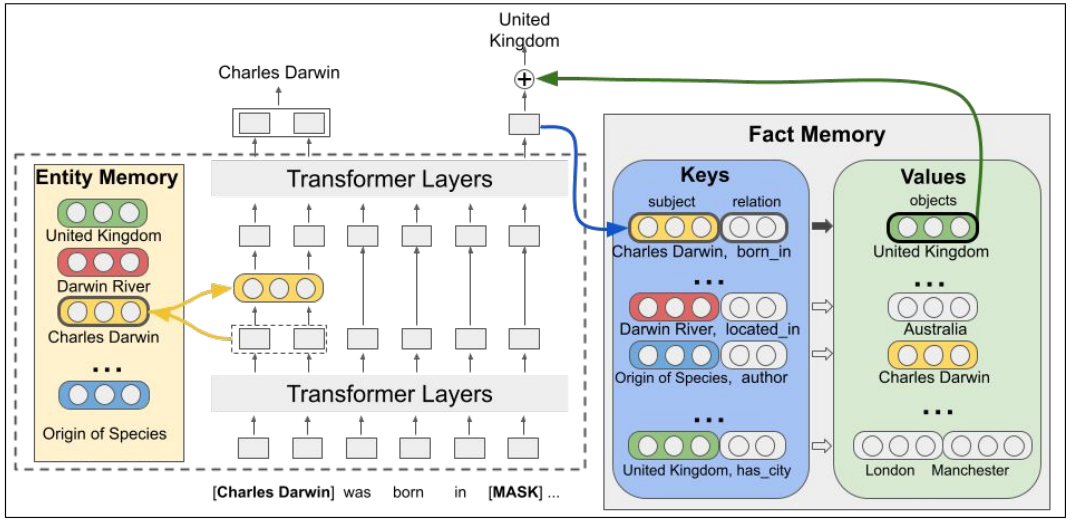
	Activated Params	TriviaQA Accuracy
T5	11B	42.3%
EaE	95M	43.2%

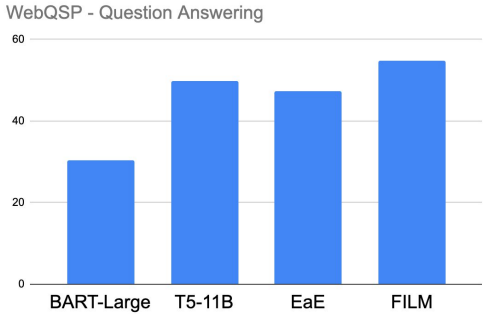
"Entities as Experts: Sparse Memory Access with Entity Supervision" EMNLP '20
Thibault Févry, Livio Baldini Soares, Nicholas FitzGerald, Eunsol Choi, Tom Kwiatkowski

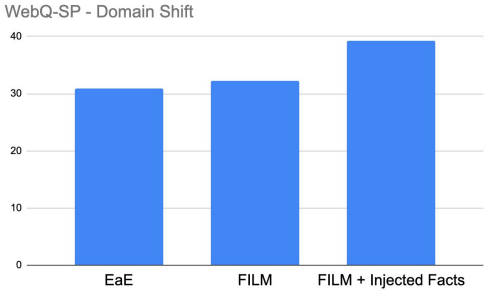
FILM (Fact Injected LM) - Adding Facts

Adding facts from the KG increases performance over very large LMs.

And, lets us update the model's knowledge about the world without retraining.

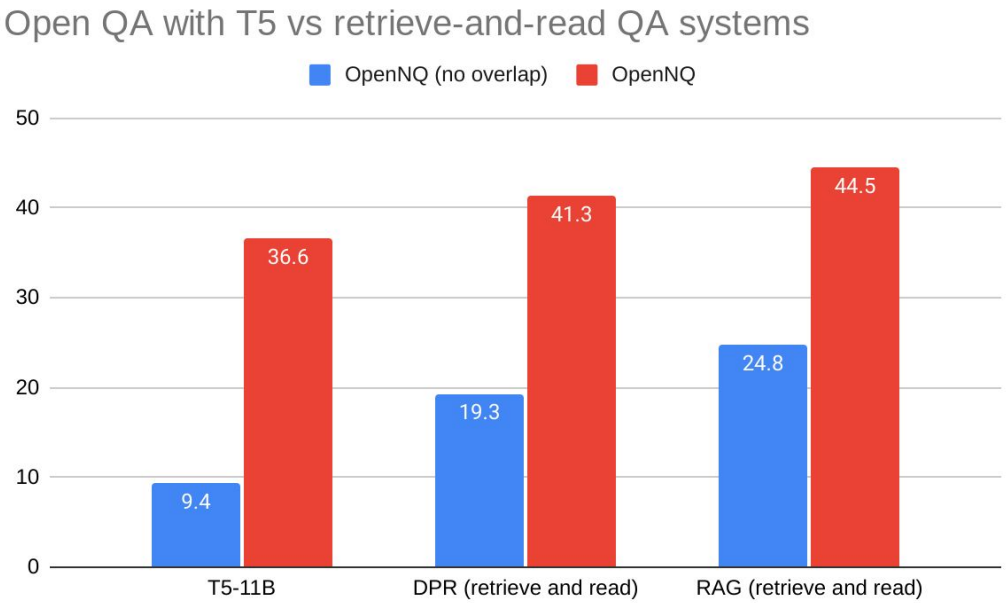






"Adaptable and Interpretable Neural Memory Over Symbolic Knowledge" NAACL '21
Pat Verga, Haitian Sun, Livio Baldini Soares, William Cohen

Retrieval-Augmented Language Models



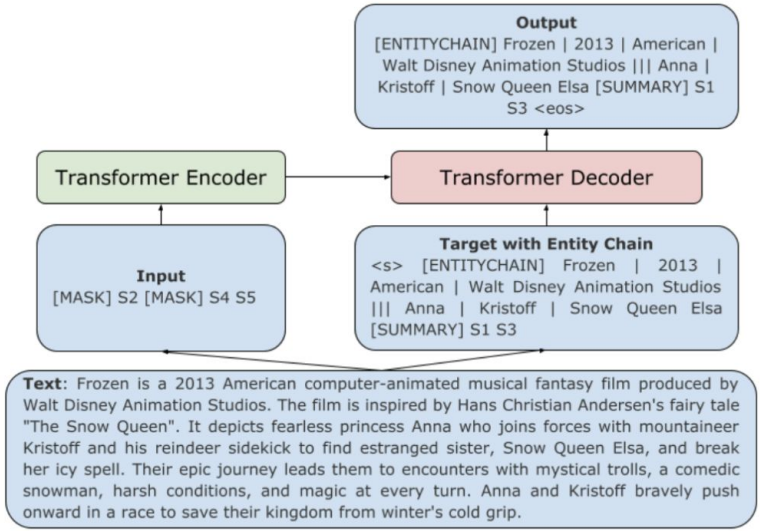
T5-11B model is *competitive* with contemporaneous QA systems on NQ Open benchmark...

...but performs much less well on test questions that are "novel" - i.e., no similar question or answer in the training set - with performance less than ½ the state-of-the-art*

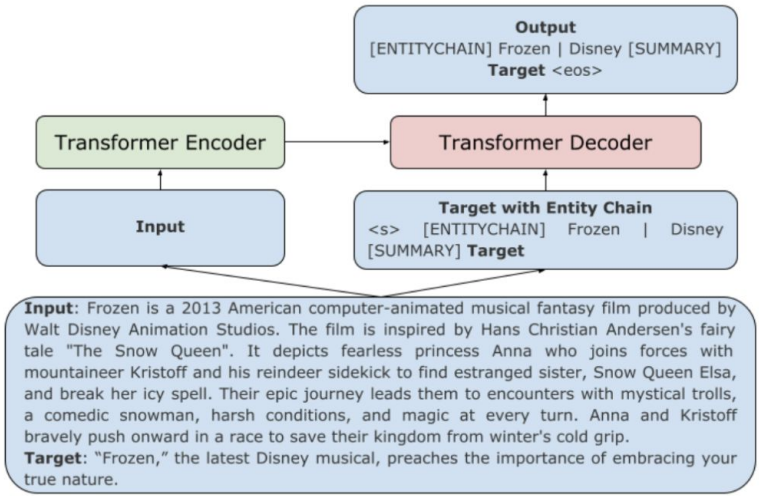
"Question and Answer Test-Train Overlap in Open-Domain Question Answering Datasets" EACL '21
Patrick Lewis, Pontus Stenetorp, Sebastian Riedel

Google Research

Step-Wise Long-Form Generation with Pre-Training



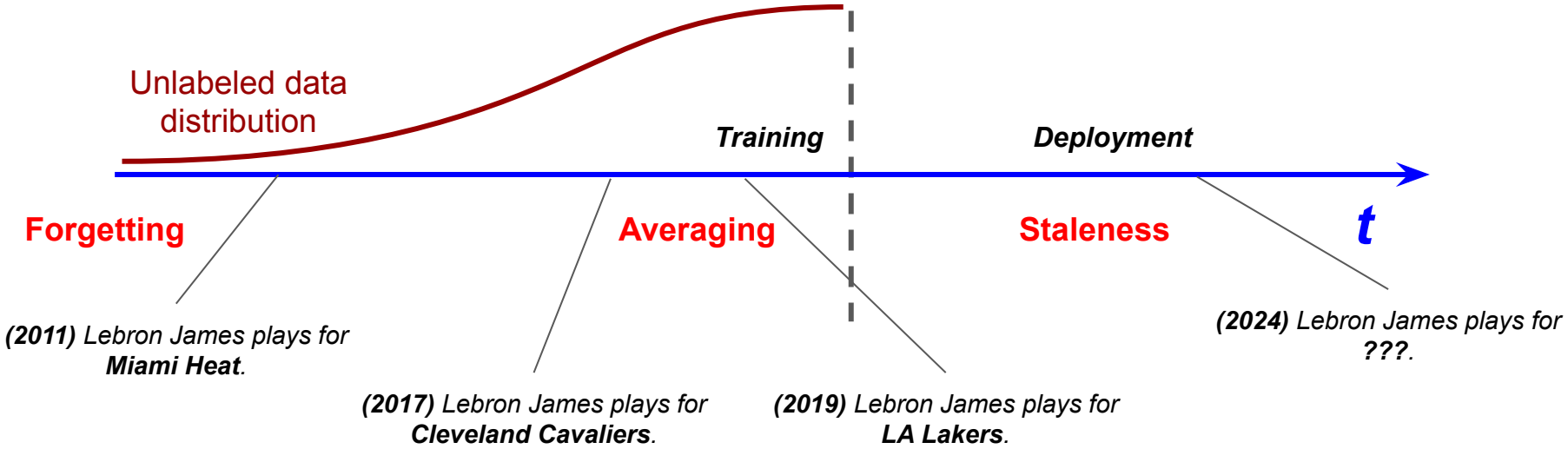
Pre-Training



Fine-Tuning

"Planning with Entity Chains for Abstractive Summarization" Arxiv '21
Shashi Narayan, Yao Zhao, Joshua Maynez, Gonçalo Simoes, Ryan McDonald

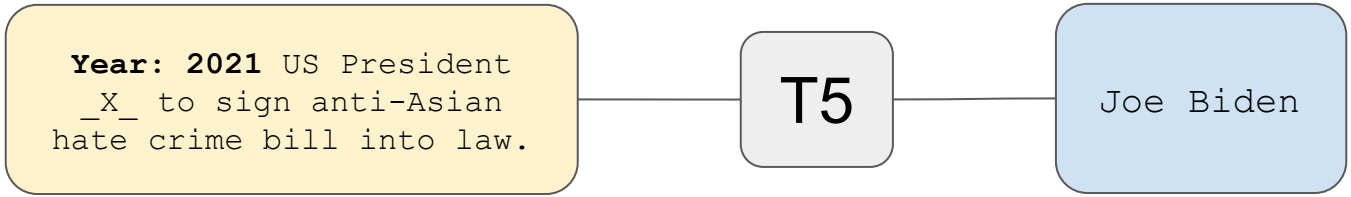
Training Data Timeline



"Time-Aware Language Models as Temporal Knowledge Bases" Arxiv '21
Bhuwan Dhingra, Jeremy R. Cole, Julian Martin Eisenschlos, Daniel Gillick, Jacob Eisenstein, William W. Cohen

Joint Models of Text and Time

- **Idea**: Instead of **Pr(x)** model **Pr(x, t)**
  - Many corpora come with timestamps (e.g. News)



- Preliminary results show that such models can improve memorization of the past, improve calibration of future events and allow for **30x** faster adaptation to new data as it becomes available.

"Time-Aware Language Models as Temporal Knowledge Bases" Arxiv '21
Bhuwan Dhingra, Jeremy R. Cole, Julian Martin Eisenschlos, Daniel Gillick, Jacob Eisenstein, William W. Cohen

Numerical reasoning?

Question:

How many yards do the first two field goals converted add up to?

Passage:

Jay Feely getting a **53-yard field goal**. In the second quarter, Miami drew closer as Feely kicked a **44-yard field goal**, yet New York replied with

Answer: **97** (53 + 44)

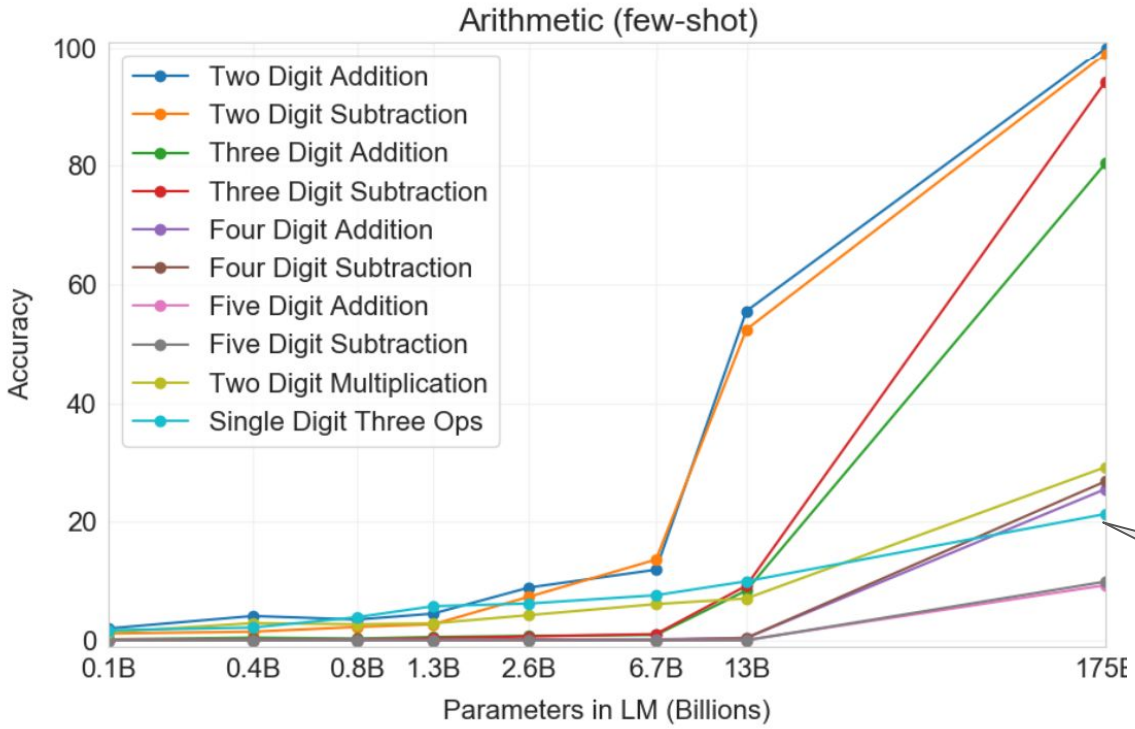
Google Assistant Scenario

User: Wake me up 30 minutes before my earliest meeting tomorrow.

Assistant: Ok, setting an alarm for **8:30 A.M.**.

QA example from:
DROP: A Reading Comprehension Benchmark Requiring Discrete Reasoning Over Paragraphs, Dua et al., 2019

Can we get numerical capabilities from scale?



But:

Q: What is 1000 + 4000?
A: 5000

Q: What is 4000 + 1000?
A: 2000

Addition/Subtraction of 3+ digits, and single digit three ops

Language Models are Few-Shot Learners, Brown et al., 2020

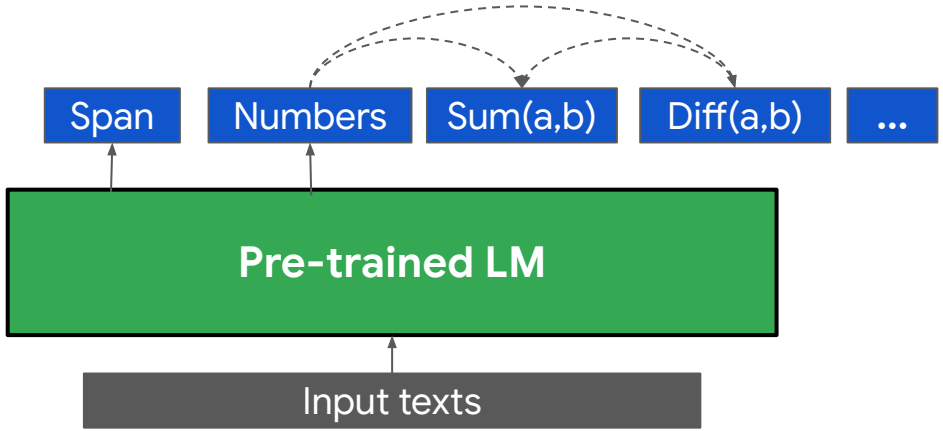
Adding numerical reasoning abilities to LMs

Combine the LM with a lightweight graph-based model to perform computation

Pros: Interpretable via the graph.
Cons: Can't generalize to unseen computations.

Add numerical data to pre-training or multi-task fine-tuning

Pros: Can be more general purpose.
Cons: Need to collect/synthesize data.



"Giving BERT a Calculator: Finding Operations and Arguments with Reading Comprehension" EMNLP '19
Daniel Andor, Luheng He, Kenton Lee, Emily Pitler

Trustworthy NLP

No machine learned model will be perfect.
How can we build trustworthy systems out of untrustworthy components?

Three pillars for trustworthiness:

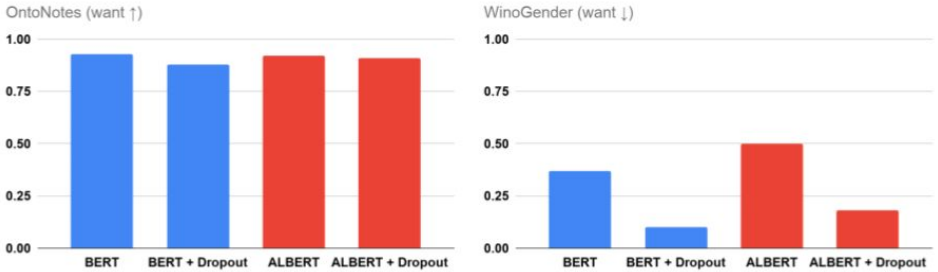
- Managing expectations / robustness / fairness
- Proof of work / Interpretability
- Controllable policies

Managing expectations

- Trust implies expectations about behavior
- Trust builds over time when expectations are fulfilled
- Well-defined problems make it easy for the expectations to match the abilities

Desired and undesired correlations

1. Carefully evaluate unintended correlations.
2. Be mindful of seemingly innocuous configuration differences.
3. Focus on general mitigations.



Mitigation →		Name (A-M)		
Evaluation ↓	Baseline	Same Gender	Flip Gender	Random Gender
DisCo (Names A-M)	3.9	2.5	2.6	**1.2**
DisCo (Names N-Z)	2.5	2.6	2.3	**2.0**
DisCo (Terms)	1.1	1.3	1.3	**0.8**

"Measuring and Reducing Gendered Correlations in Pre-trained Models" Arxiv '20
Kellie Webster, Xuezhi Wang, Ian Tenney, Alex Beutel, Emily Pitler, Ellie Pavlick, Jilin Chen, Ed Chi, Slav Petrov

Proof of work

In school we aren't happy if students just give the right answers:

- we want the right answers for the right reasons

Proof of work for a model is not necessarily mathematical proof though

- many ways of convincing that a reasoning pattern is valid

Proof of work example: QED

Decompose passage-based QA into:

- retrieval
- coreference
- entailment

Question: who wrote the film howl's moving castle?
Passage: Howl's Moving Castle is a 2004 Japanese animated fantasy film written and directed by Hayao Miyazaki. It is based on the novel of the same name, which was written by Diana Wynne Jones. The film was produced by Toshio Suzuki.
Answer: Hayao Miyazaki

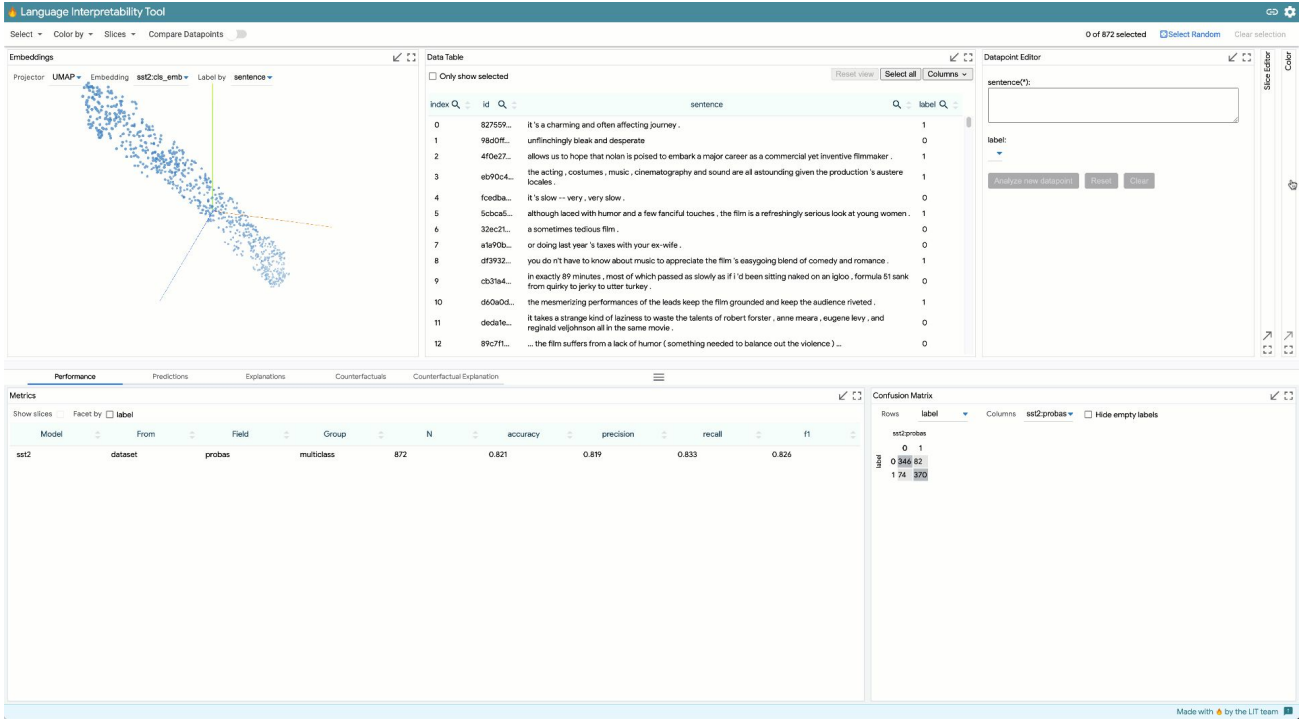
---

(1) Sentence Selection
Howl's Moving Castle is a 2004 Japanese animated fantasy film written and directed by Hayao Miyazaki.
(2) Referential Equality
the film howl's moving castle = Howl's Moving Castle
(3) Entailment
X is a 2004 Japanese animated fantasy film written and directed by ANSWER.⊢ ANSWER wrote X.

"QED: A Framework and Dataset for Explanations in Question Answering" TACL '21
Matthew Lamm, Jennimaria Palomaki, Chris Alberti, Daniel Andor, Eunsol Choi, Livio Baldini Soares, Michael Collins

Google Research

Language Interpretability Tool



"The Language Interpretability Tool: Extensible, Interactive Visualizations and Analysis for NLP Models" EMNLP '20

I. Tenney, J. Wexler, J. Bastings, T. Bolukbasi, A. Coenen, S. Gehrmann, E. Jiang, M. Pushkarna, C. Radebaugh, E. Reif, A. Yuan

Controllable policies

Pretrained models are a description of the world

- the world has good and bad things in it

Most problems we want to solve are inherently open-ended and ambiguous

- we need to be able to tip the scales to force the right outcome

How can we design systems with the right control affordances?

Controllable policies

Example: gender in translation

- different languages require that different things are gendered
- so translation models often have to infer gender
- by default the most common gender in the data is inferred
- this is bad

Alternative:

- align the training data using a simple word alignment mechanism
- use this to annotate source genderless text with target gender
- train a model that respects these annotations
- now any learned or programmable policy can be used to set gender

Conclusions

- Scale is important

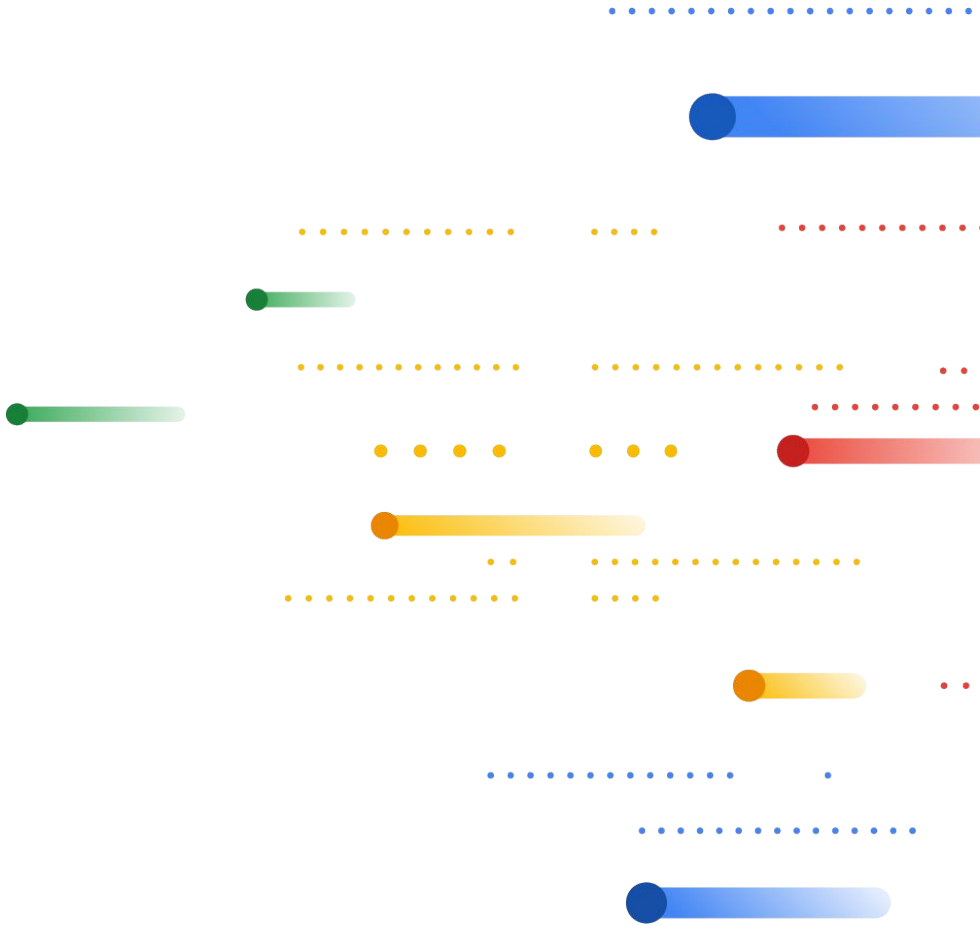
- But there are many important problems that will likely not be solved with scale alone:
  - Memory
  - Attribution
  - Planning for Long-Fortm Generation
  - Modeling Time
  - Using Tools
  - Building Trust

- Other important topics that we didn't talk about:
  - Datasets
  - Evaluation Metrics

Google Research

Thank you!

Slav Petrov
on behalf of many wonderful
colleagues at Google Research



slav@google.com