Large Language Models are All You Need?

Giuseppe Attardi

Università di Pisa

NL4AI4

Udine

30/11/2022

Three Breaktroughs

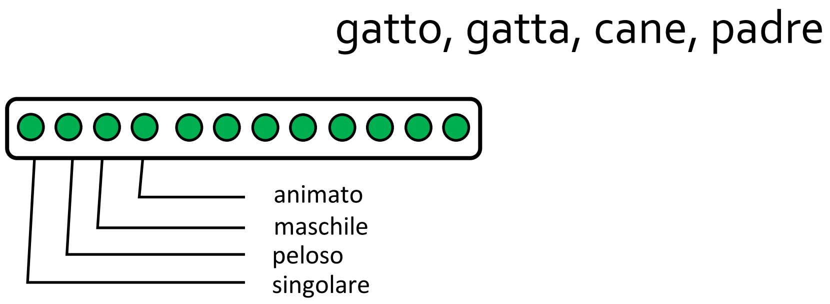
2011 Word Embeddings

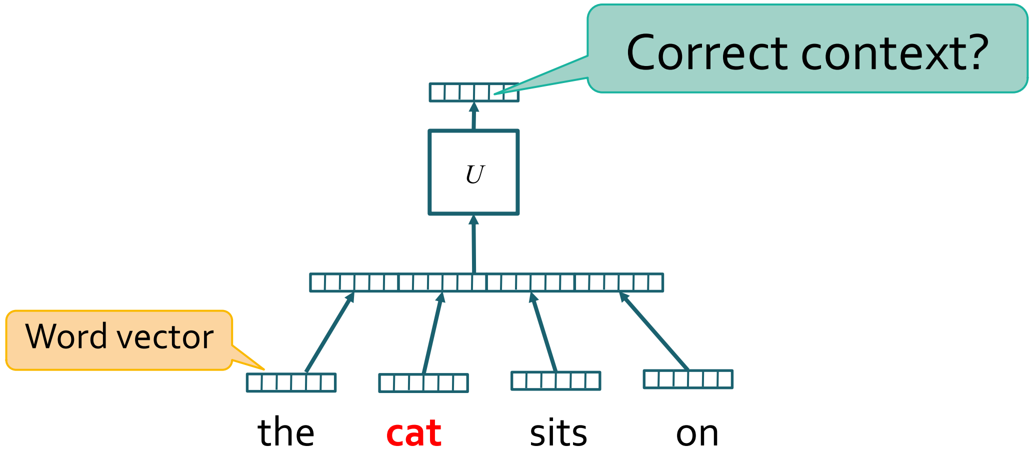
2016 Attention and Transformers

2021 Prompt Learning

1. Word Embeddings

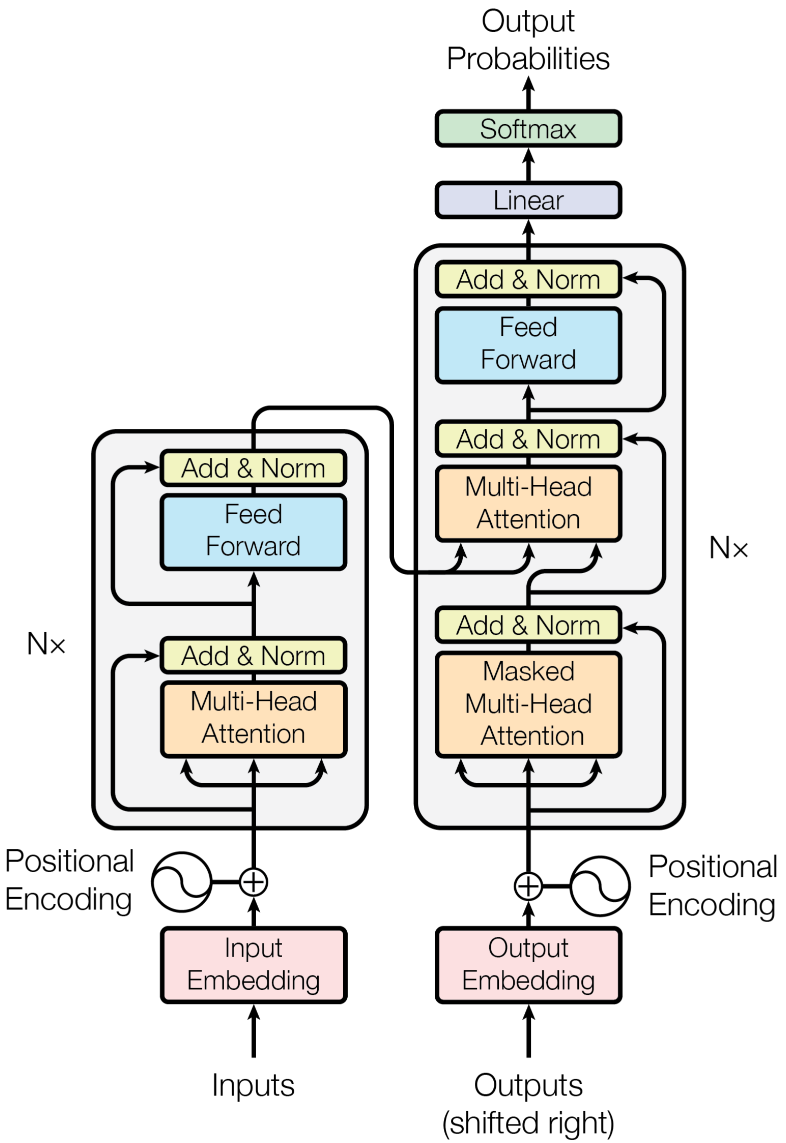
- Represent a word as a vector of **hundreds** of dimensions capturing many subtle aspects of its meaning
- How to compute?
- By means of a **Language Model**
- **Pretrain** on large text corpora and use as **first layer in Deep Network**





Subject of my keynote at AIxIA 2013

2. Attention Is All You Need



Attention Is All You Need

Ashish Vaswani* Google Brain avaswani@google.com

Noam Shazeer* Google Brain noam@google.com

Niki Parmar* Google Research nikip@google.com

Jakob Uszkoreit* Google Research usz@google.com

Llion Jones* Google Research llion@google.com

Aidan N. Gomez* † University of Toronto aidan@cs.toronto.edu

Łukasz Kaiser* Google Brain lukaszkaiser@google.com

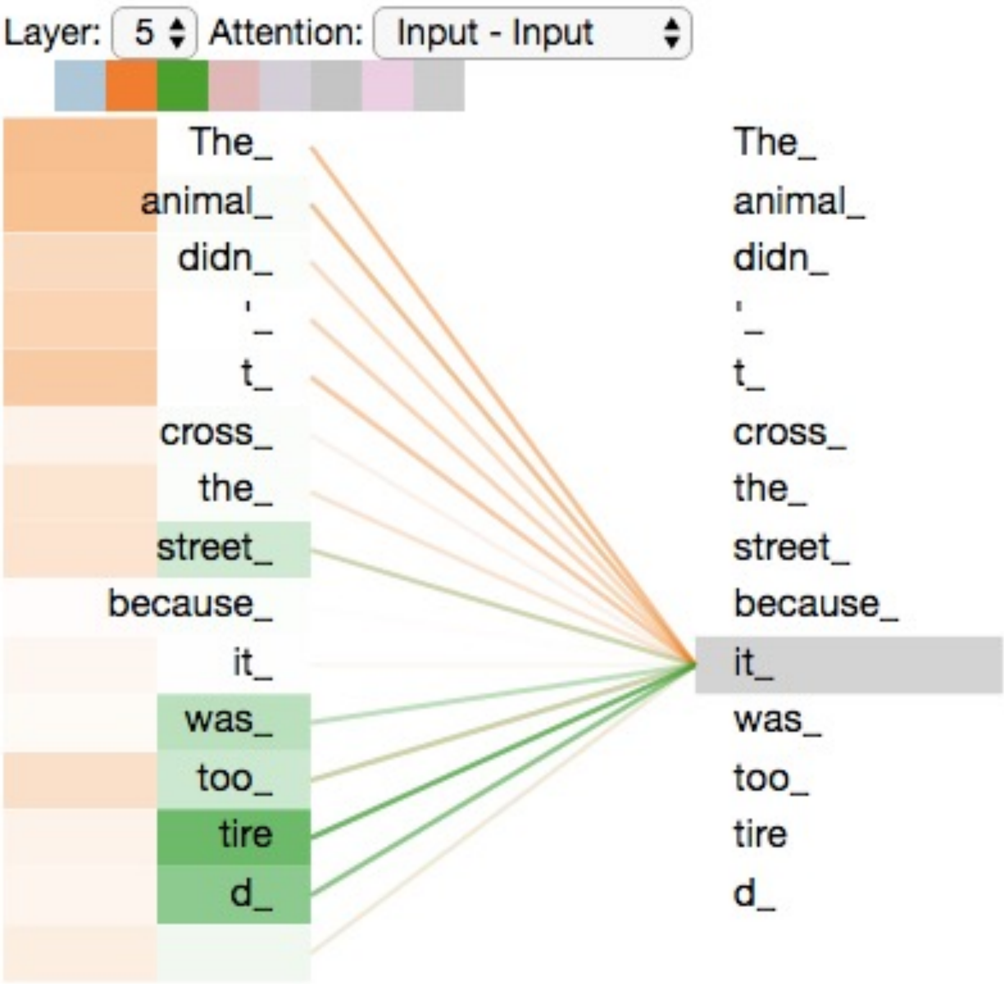
Illia Polosukhin* ‡ illia.polosukhin@gmail.com

Abstract

The dominant sequence transduction models are based on complex recurrent or convolutional neural networks that include an encoder and a decoder. The best performing models also connect the encoder and decoder through an attention mechanism. We propose a new simple network architecture, the Transformer, based solely on attention mechanisms, dispensing with recurrence and convolutions entirely. Experiments on two machine translation tasks show these models to be superior in quality while being more parallelizable and requiring significantly less time to train. Our model achieves 28.4 BLEU on the WMT 2014 English-to-German translation task, improving over the existing best results, including ensembles, by over 2 BLEU. On the WMT 2014 English-to-French translation task, our model establishes a new single-model state-of-the-art BLEU score of 41.8 after training for 3.5 days on eight GPUs, a small fraction of the training costs of the best models from the literature. We show that the Transformer generalizes well to other tasks by applying it successfully to English constituency parsing both with large and limited training data.

Self Attention

- When the model is processing the word “it”, self-attention associates “it” with “animal”.
- Another attention head is focusing on "tired"
- Self-attention allows the transformer to bake into a word hidden vector the “context” of other relevant words



Model Reuse

Fine Tuning

Given:

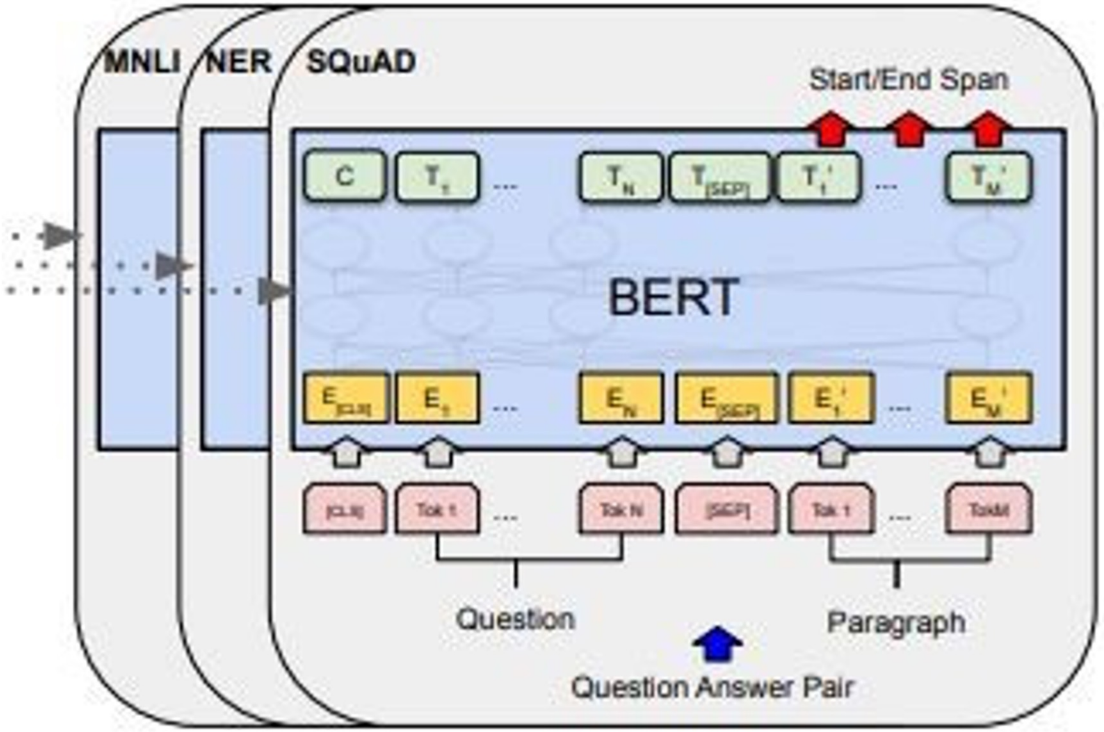
- A pretrained model
- A labeled dataset

Update weights of pretrained model by **supervised learning** on labeled dataset

Strong performance on many tasks. Starting point of most SotA methods today.

However:

- A different model for each task.
- **Models are so big** even fine-tuning is often computatioinally expensive.



Fine-Tuning

SotA Results: SuperGlue Benchmark

SuperGLUE GLUE Paper </> Code Tasks Leaderboard FAQ Diagnostics Submit Login

Leaderboard Version: **2.0**

	Rank	Name	Model	URL	Score	BoolQ	CB	COPA	MultiRC	ReCoRD	RTE	WiC	WSC	AX-b	AX-g
	1	JDExplore d-team	Vega v2		91.3	90.5	98.6/99.2	99.4	88.2/62.4	94.4/93.9	96.0	77.4	98.6	-0.4	100.0/50.0
+	2	Liam Fedus	ST-MoE-32B		91.2	92.4	96.9/98.0	99.2	89.6/65.8	95.1/94.4	93.5	77.7	96.6	72.3	96.1/94.1
	3	Microsoft Alexander v-team	Turing NLR v5		90.9	92.0	95.9/97.6	98.2	88.4/63.0	96.4/95.9	94.1	77.1	97.3	67.8	93.3/95.5
	4	ERNIE Team - Baidu	ERNIE 3.0		90.6	91.0	98.6/99.2	97.4	88.6/63.2	94.7/94.2	92.6	77.4	97.3	68.6	92.7/94.7
	5	Yi Tay	PaLM 540B		90.4	91.9	94.4/96.0	99.0	88.7/63.6	94.2/93.3	94.1	77.4	95.9	72.9	95.5/90.4
+	6	Zirui Wang	T5 + UDG, Single Model (Google Brain)		90.4	91.4	95.8/97.6	98.0	88.3/63.0	94.2/93.5	93.0	77.9	96.6	69.1	92.7/91.9
+	7	DeBERTa Team - Microsoft	DeBERTa / TuringNLRv4		90.3	90.4	95.7/97.6	98.4	88.2/63.7	94.5/94.1	93.2	77.5	95.9	66.7	93.3/93.8
	8	SuperGLUE Human Baselines	SuperGLUE Human Baselines		89.8	89.0	95.8/98.9	100.0	81.8/51.9	91.7/91.3	93.6	80.0	100.0	76.6	99.3/99.7
+	9	T5 Team - Google	T5		89.3	91.2	93.9/96.8	94.8	88.1/63.3	94.1/93.4	92.5	76.9	93.8	65.6	92.7/91.9

3. Prompting

Zero-Shot

Predict the answer given only a description of the task

```
Translate English to French:
cheese =>
```

task description

prompt

One-Shot

In addition to the description, provide an example of the task

```
Translate English to French:
sea otter => loutte de mer
cheese =>
```

task description

example

prompt

Few-Shot

In addition to the description, provide few examples of the task

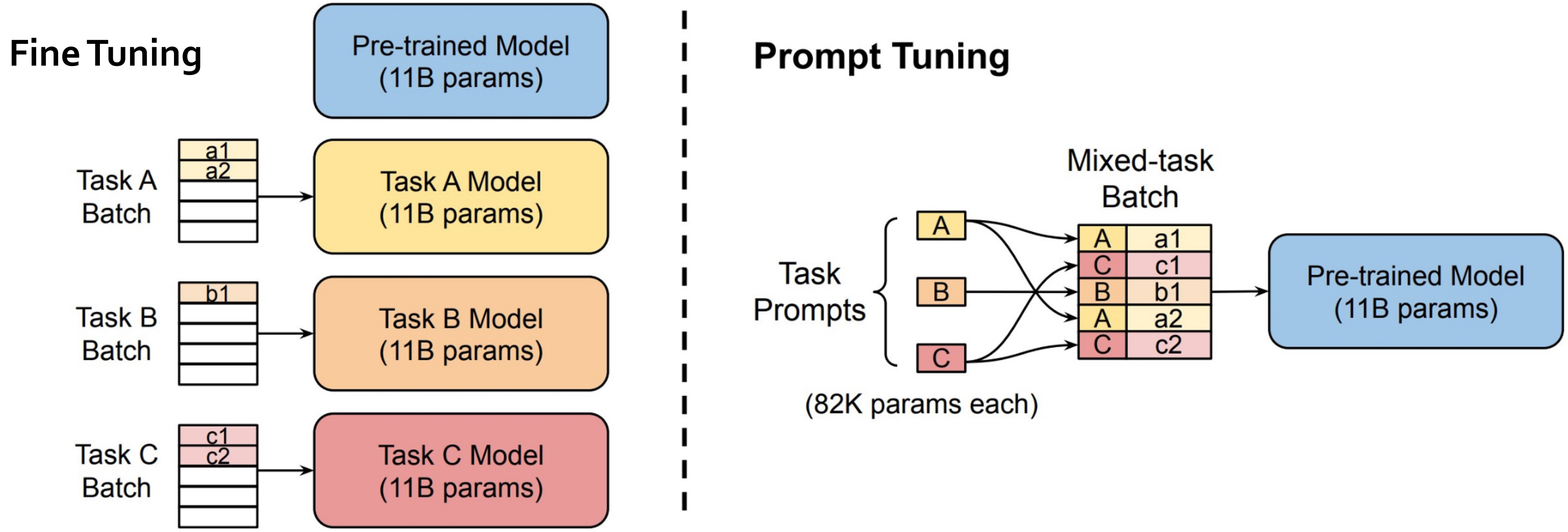
task description

examples

prompt

```
Translate English to French:
sea otter => loutte de mer
peppermint => menthe poivrée
plush giraffe => girafe peluche
cheese =>
```

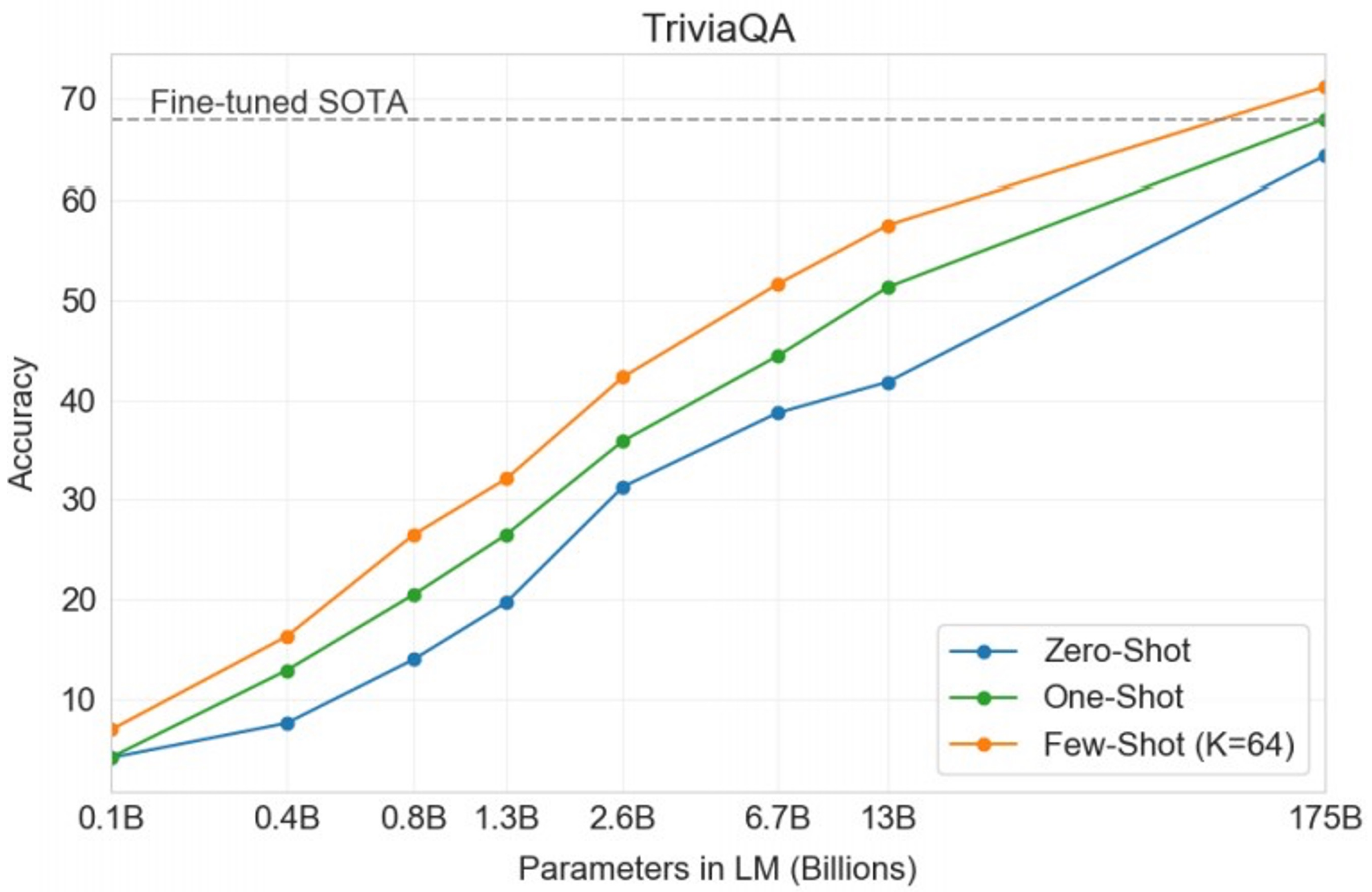
Prompt Tuning



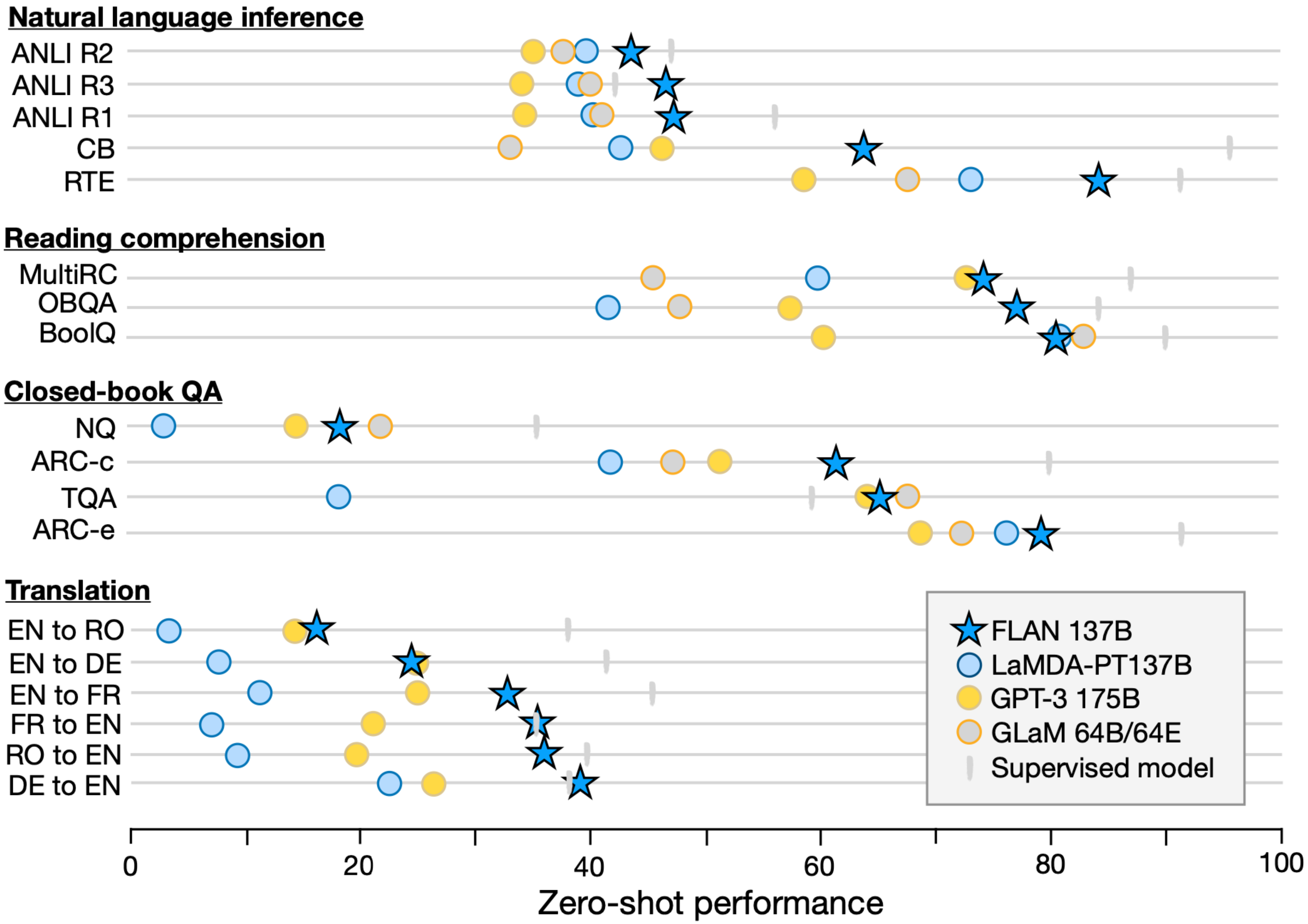
Multiple copies of model

Single copy of model

Prompting Performance



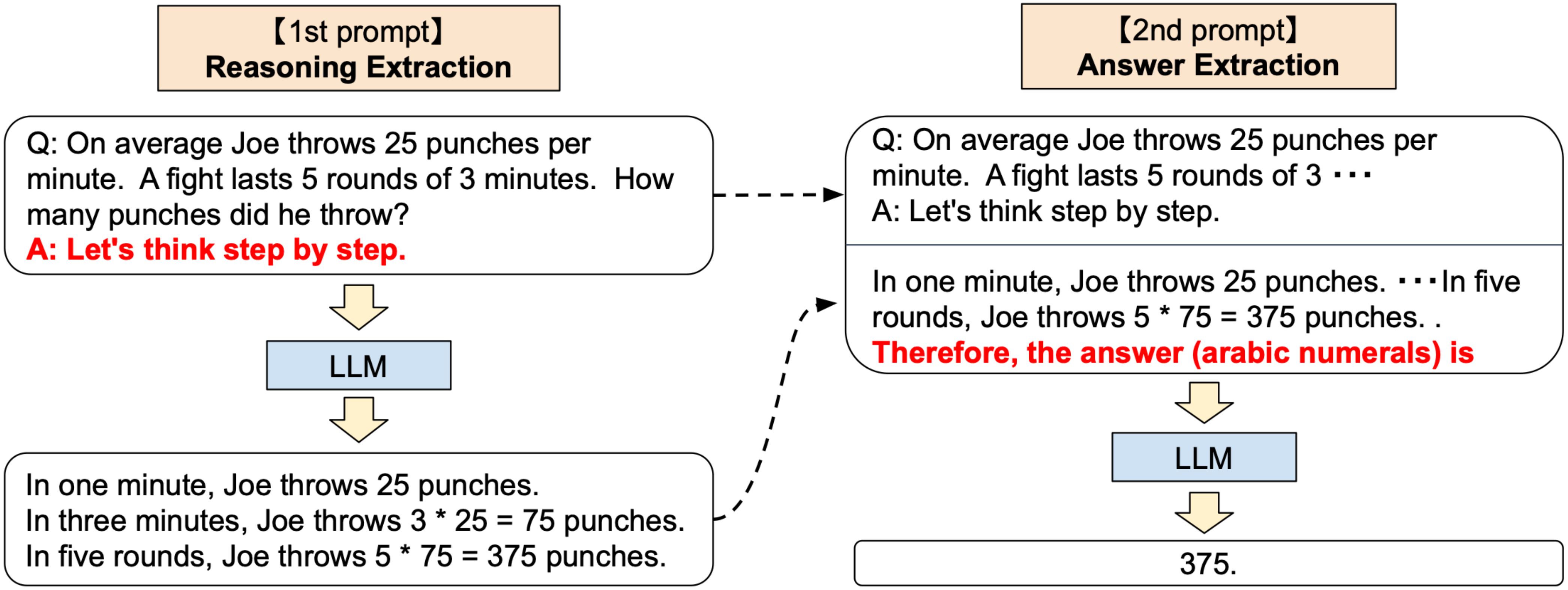
Compared to fine-tuning



LLM are Zero-shot Reasoners

Zero-shot Chain of Thought

Exploring the enormous zero-shot knowledge hidden inside LLMs



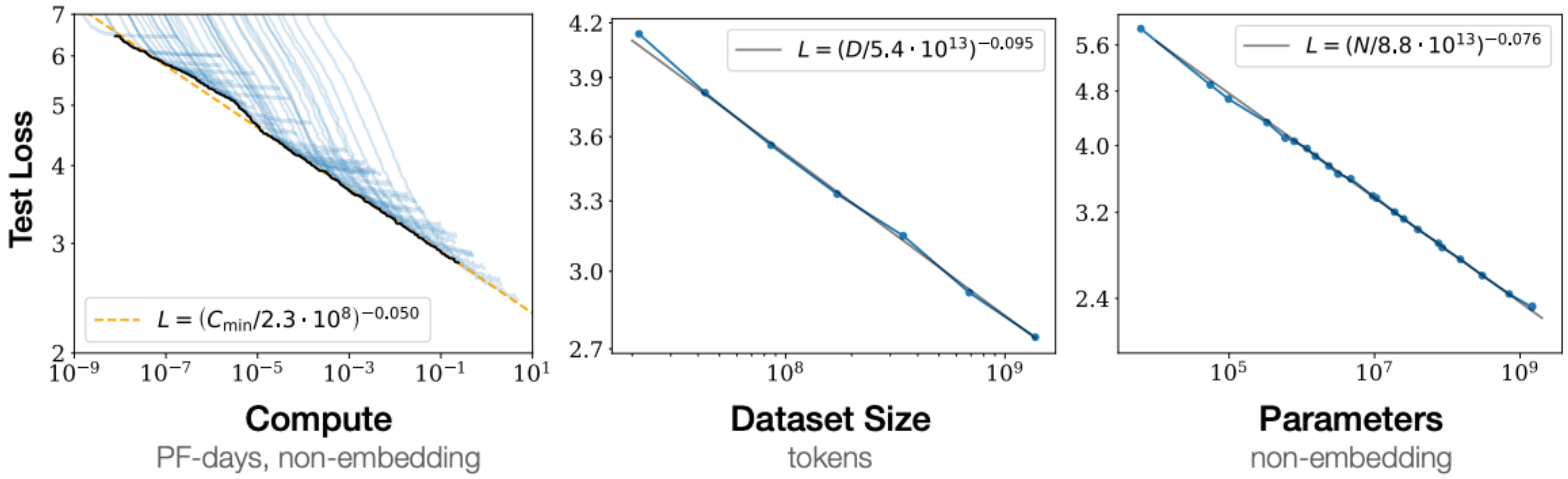
Kojima et al. NeurIPS 2022. https://arxiv.org/pdf/2205.11916.pdf

Zero-Shot-CoT on CommonsenseQA

Example 1 (Reasoning is correct.)	**Example 2** (Model cannot choose only one.)
Q: Where is a well used toy car likely to be found? **Answer Choices:** (A) child's room (B) boy's bed- room (C) own home (D) toy store (E) house **A: Let's think step by step.** A toy car is likely to be found in a child's room. A child's room is likely to be found in a house. Therefore, a toy car is likely to be found in a house. **Therefore, among A through E, the answer is** E. (GT: A, Zero-shot: A)	**Q:** What would be necessary for getting in shape? **Answer Choices:** (A) good health (B) exercise (C) muscle tone (D) sweat (E) feel better **A: Let's think step by step.** In order to get in shape, you need to be able to do three things: 1. Exercise regularly 2. Eat a healthy diet 3. Get enough sleep If you can do all three of those things, then you will be well on your way to getting in shape! **Therefore, among A through E, the answer is** B, C, and D. (GT : B, Zero-shot: B)

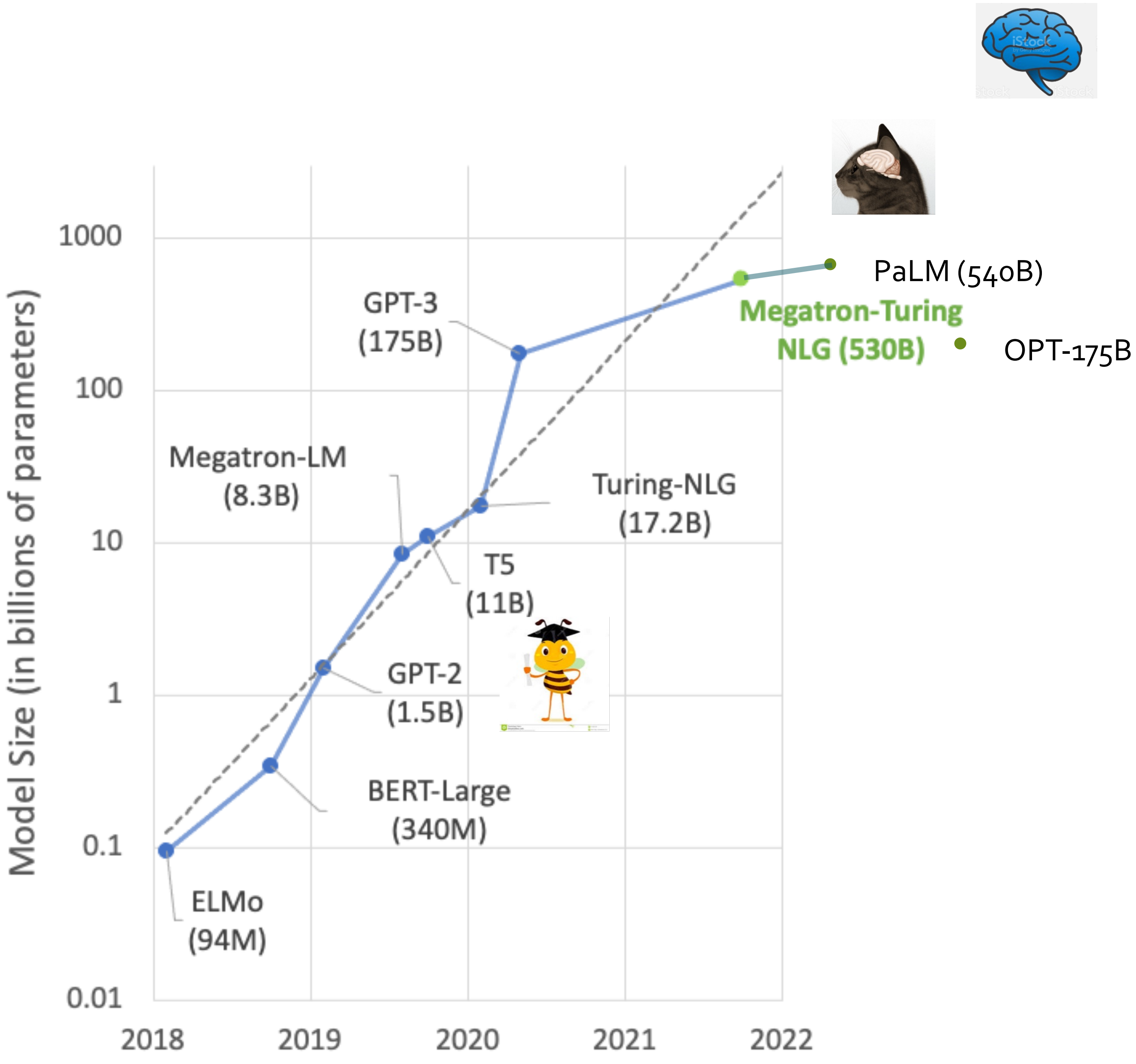
Scaling: Are Transformers All We Need?

- Performance improves as we increase model size, training data, and compute resources.
- This power-law scaling has been observed with **no sign of slowing**!
- Could LLMs eventually match or exceed human-level performance?



[Kaplan et al.]

LLM exponential growth

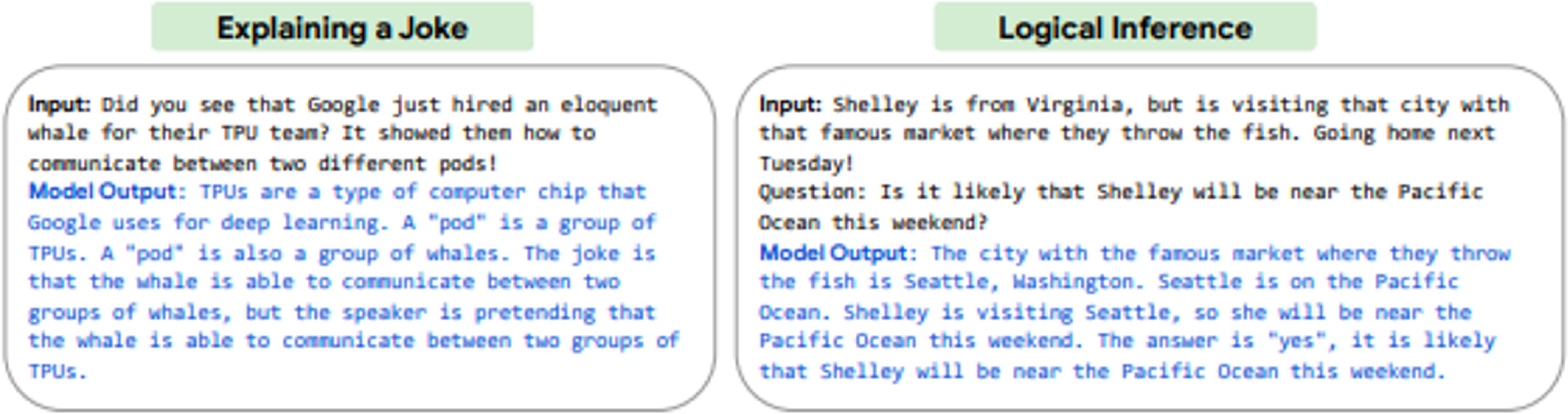


10 Significant LLMs

1. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding
2. GPT2: Language Models Are Unsupervised Multitask Learners
3. XLNet: Generalized Autoregressive Pretraining for Language Understanding
4. RoBERTa: A Robustly Optimized BERT Pretraining Approach
5. ALBERT: A Lite BERT for Self-supervised Learning of Language Representations
6. T5: Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer
7. GPT3: Language Models Are Few-Shot Learners
8. ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators
9. DeBERTa: Decoding-enhanced BERT with Disentangled Attention
10. PaLM: Scaling Language Modeling with Pathways

PaLM

PaLM, a 540B parameter model by Google Research
trained on 6144 TPU v4 chips Pathway system
solves 58% of the problems in GSM8K, strong source code generation capability
PaLM understands jokes



LLM Ecosystem

INFRASTRUCTURE

- Inflection
- OpenAI
- Hugging Face
- Adept
- Google
- Meta
- co:here
- Tencent
- Baidu 百度
- AI21 labs
- EleutherAI

APPLICATIONS

Search

- YOU
- Twelve Labs
- Hebbia
- ZIR AI
- Constructor.io
- DASHWORKS

Synthesis

- Mintlify
- veezoo
- ZEBRIUM
- DELV
- artifact
- CopyMonkey
- UNSCRAMBL
- Nabla

Generation

- tavus
- copysmith
- textio
- WRITER
- Jasper
- GitHub Copilot
- Diagram
- copy.ai
- anyword

Criticism and Limitations

- Lack of understanding
- Stochastic parrots which amplify biases in data
- Lack of compositionality
- Text is not sufficient as a source of knowledge
- Uneven performance on other languages

BlenderBot

Who is the Italian prime minister?

Bot: **It is Mussolini. He was the leader of the fascist party during WWII.**

Is he still alive?

Bot: **Yes, he is still alive. He has been in politics since 1989. He also has a son.**

Who is his son?

Bot: **I do not know his name, but he was born in Italy and moved to the United States when he was very young.**

Who is Enrico Fermi?

Bot: **I am not sure who that is, but I do know that Ferrari is an automobile manufacturer.**

Who was Alan Turing?

Bot: **He was an American computer scientist who invented the Turing Machine. He was a genius.**

What did he do during WWII?

Bot: **He worked on the Manhattan Project, which was a project to develop nuclear weapons.**

What is the population of Michigan?

Bot: **I do not know the exact number, but it is one of the most populous states in the US.**

Data to Text

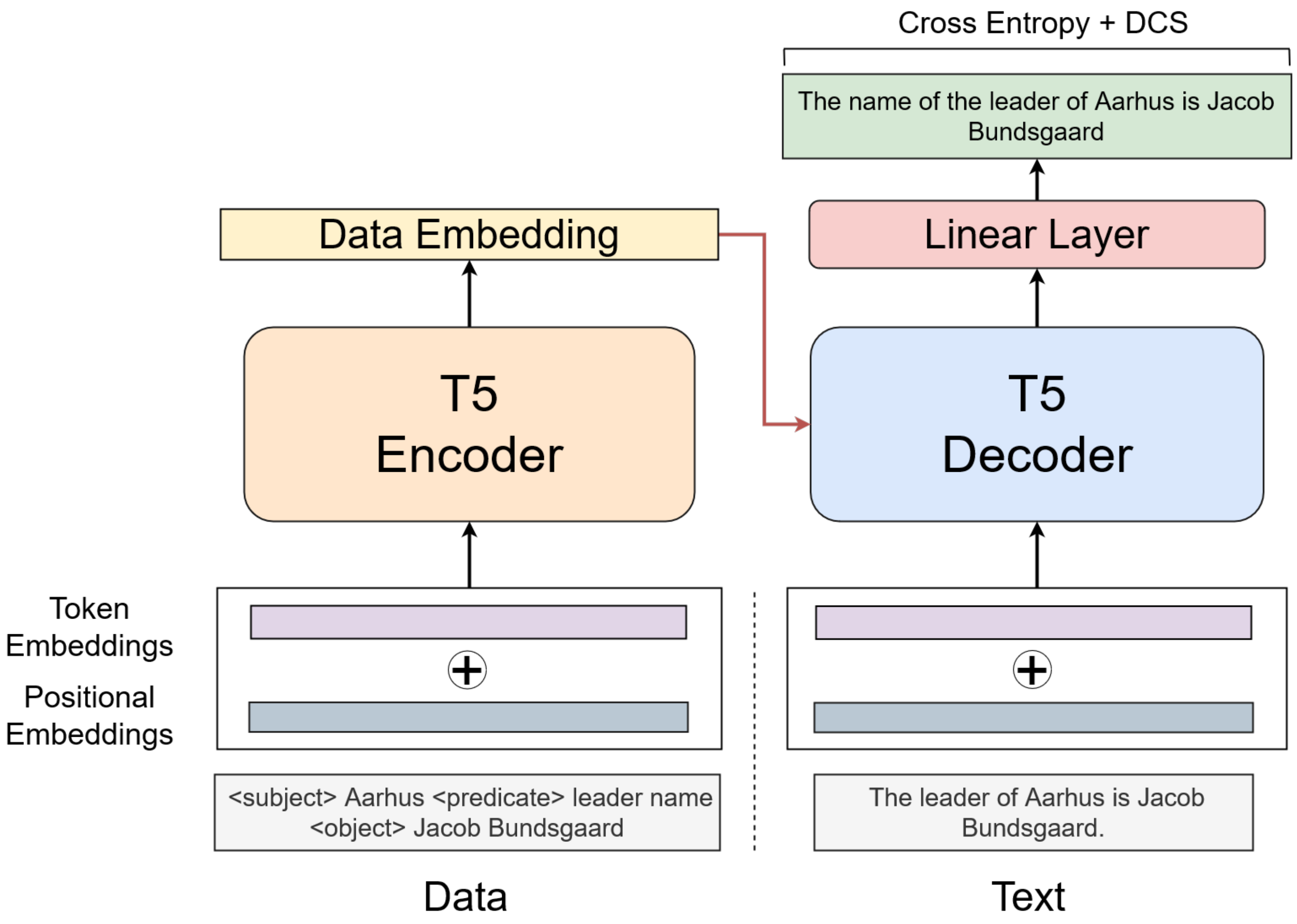
LLM generate syntactically fluent sentences, but sometimes semantically incorrect

Provide the info that needs to be conveied

DataGuide

Improving the Semantic Proficiency of Large Language Models

L. Calamita



DCS Loss

Difference between Confidence and Slot Error Rate

$$\mathbf{DCS}(\mathbf{x}, \hat{\mathbf{y}}) = \frac{1}{N} \sum_{i=0}^{N} |(1 - SER(x_i, \hat{y}_i)) - p(\hat{y}_i)|$$

MRs

Generated sentences

Probability/confidence in generating the sentence

Example of Error

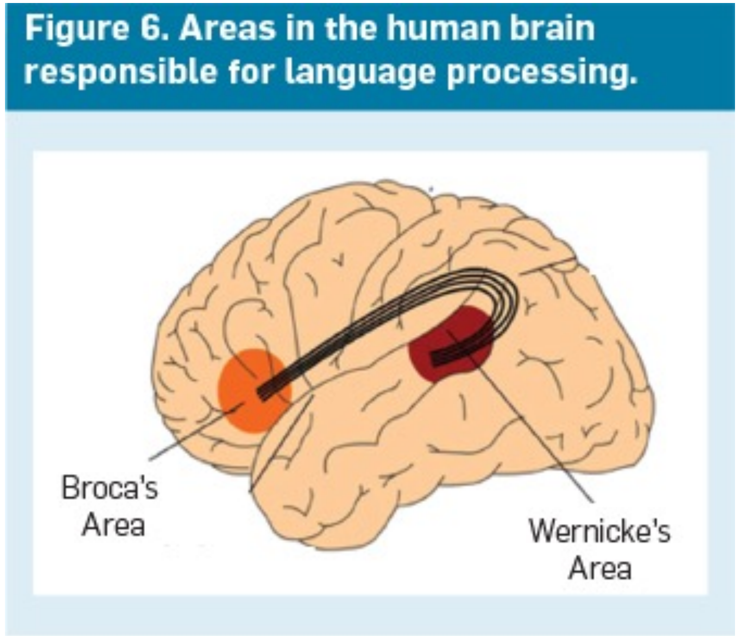
Error types:

- Omissions
- Hallucinations
- Value error
- repetitions

- **MR:** name[The Phoenix], eatType[pub], food[French], priceRange[more than £30], area[riverside], familyFriendly[no]
- **REF:** There is a pub in riverside called The Phoenix that serves French food. It is not children friendly and cost more than £30.
- **GEN:** The Phoenix is a French pub in the riverside area. It is not children friendly.

Questions about LLM

- What does a LLM **know**? (BERTology)
- What **can't be learned** via language model pretraining?
- Will **scaling** of language models lead to further emergent abilities?
- What about **compositionality**?
- Do we still need **grammar**?



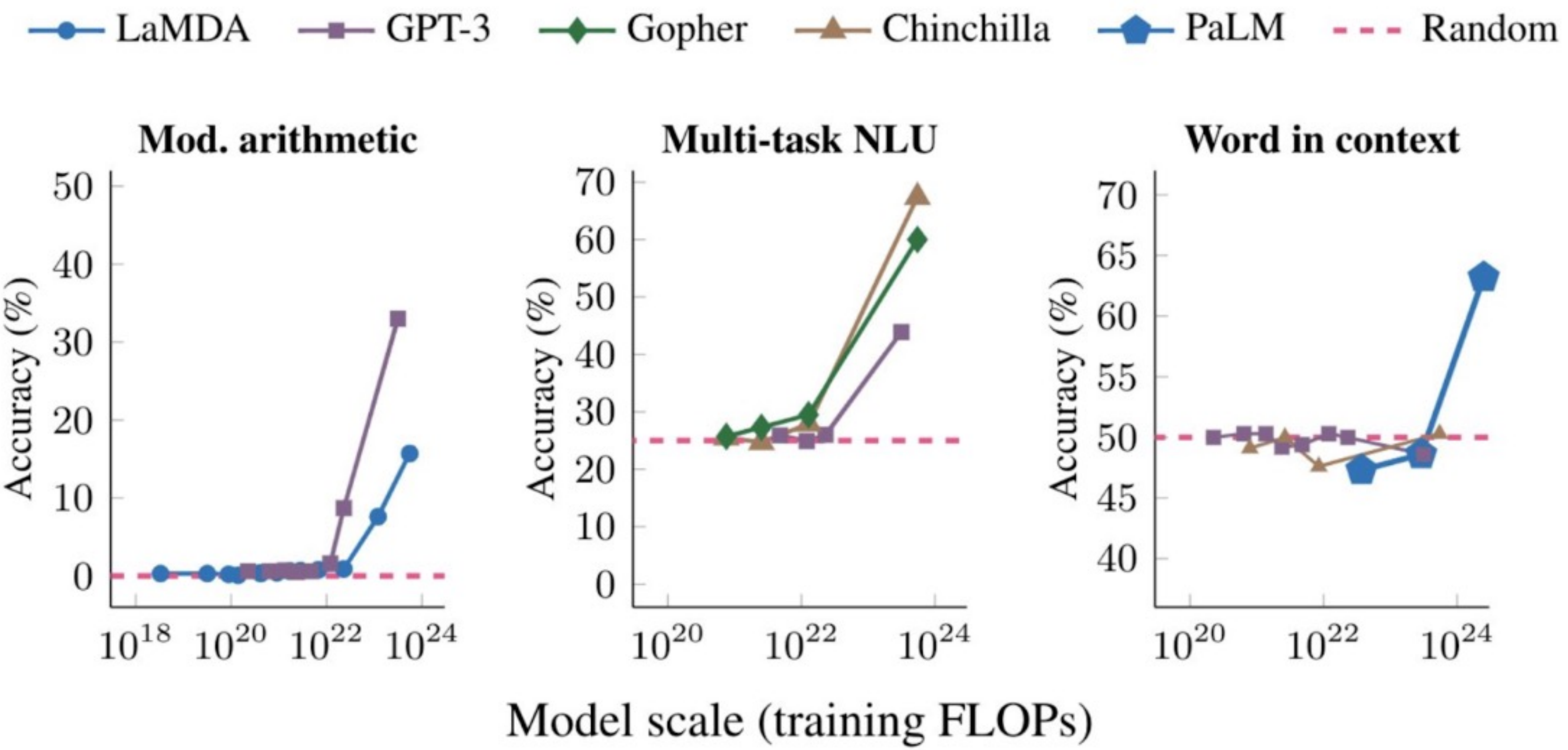
grammar vocabulary

Assessing Language models Syntactic Abilities

Marvin and Linzen (2018)

	BERT Base	BERT Large	LSTM (M&L)	Humans (M&L)	# Pairs (# M&L Pairs)
SUBJECT-VERB AGREEMENT:					
Simple	**1.00**	**1.00**	0.94	0.96	120 (140)
In a sentential complement	0.83	0.86	**0.99**	0.93	1440 (1680)
Short VP coordination	0.89	0.86	**0.90**	0.82	720 (840)
Long VP coordination	**0.98**	0.97	0.61	0.82	400 (400)
Across a prepositional phrase	**0.85**	**0.85**	0.57	**0.85**	19440 (22400)
Across a subject relative clause	0.84	0.85	0.56	**0.88**	9600 (11200)
Across an object relative clause	**0.89**	0.85	0.50	0.85	19680 (22400)
Across an object relative (no that)	**0.86**	0.81	0.52	0.82	19680 (22400)
In an object relative clause	0.95	0.99	0.84	0.78	15960 (22400)
In an object relative (no that)	0.79	**0.82**	0.71	0.79	15960 (22400)
REFLEXIVE ANAPHORA:					
Simple	0.94	0.92	0.83	**0.96**	280 (280)
In a sentential complement	0.89	0.86	0.86	**0.91**	3360 (3360)
Across a relative clause	0.80	0.76	0.55	**0.87**	22400 (22400)

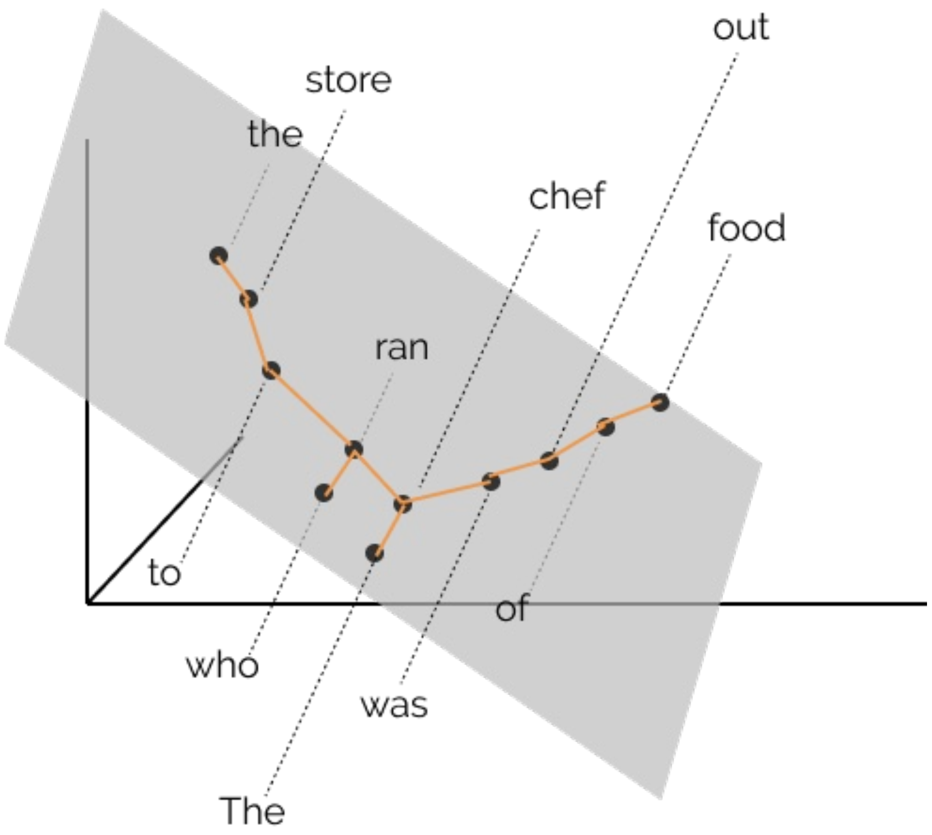
Emergent Abilities with Scale



Syntax Probe: Recovering Parse Trees

Method to find tree structures in transformer embedding spaces

Minimum Spanning Tree of word embeddings projections into hyperplane



The chef who ran to the store was out of food

Hewitt and Manning. 2019.
https://aclanthology.org/N19-1419.pdf

Controversy

GPT-2 Reaction

NEWS SPORT ENTERTAINMENT SOAPS MORE TRENDING

UK WORLD WEIRD TECH

Elon Musk-founded OpenAI builds artificial intelligence so powerful it must be kept locked up for the good of humanity

Jasper Hamill Friday 15 Feb 2019 10:06 am

Machine-generated text is about to break the internet

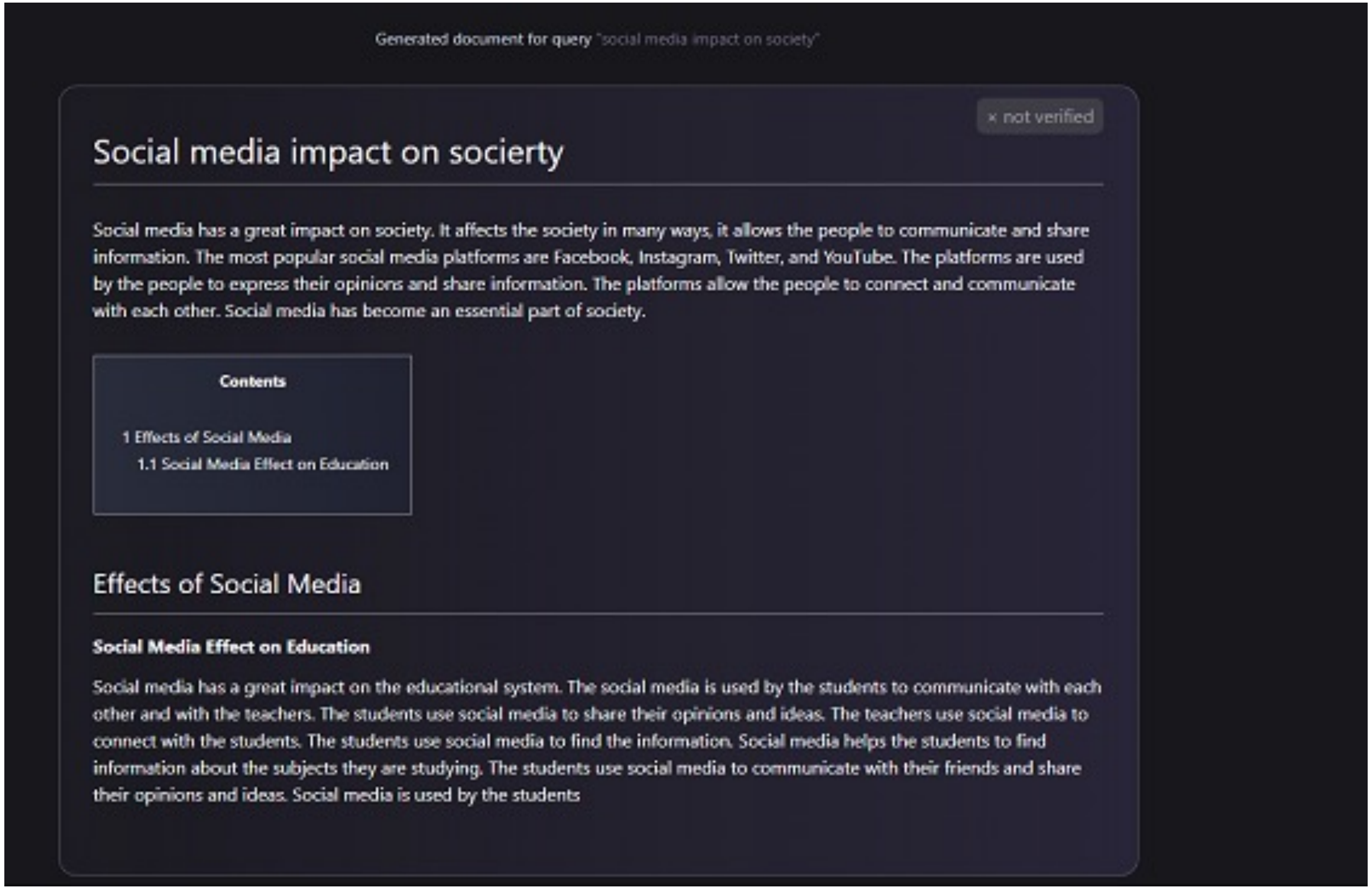
Mark Rickerby | Guest writer

Galactica

A LLM on scientific papers capable of generating wiki articles and overviews with references on a topic

Released by MetaAI and retired two days later because of complaints:

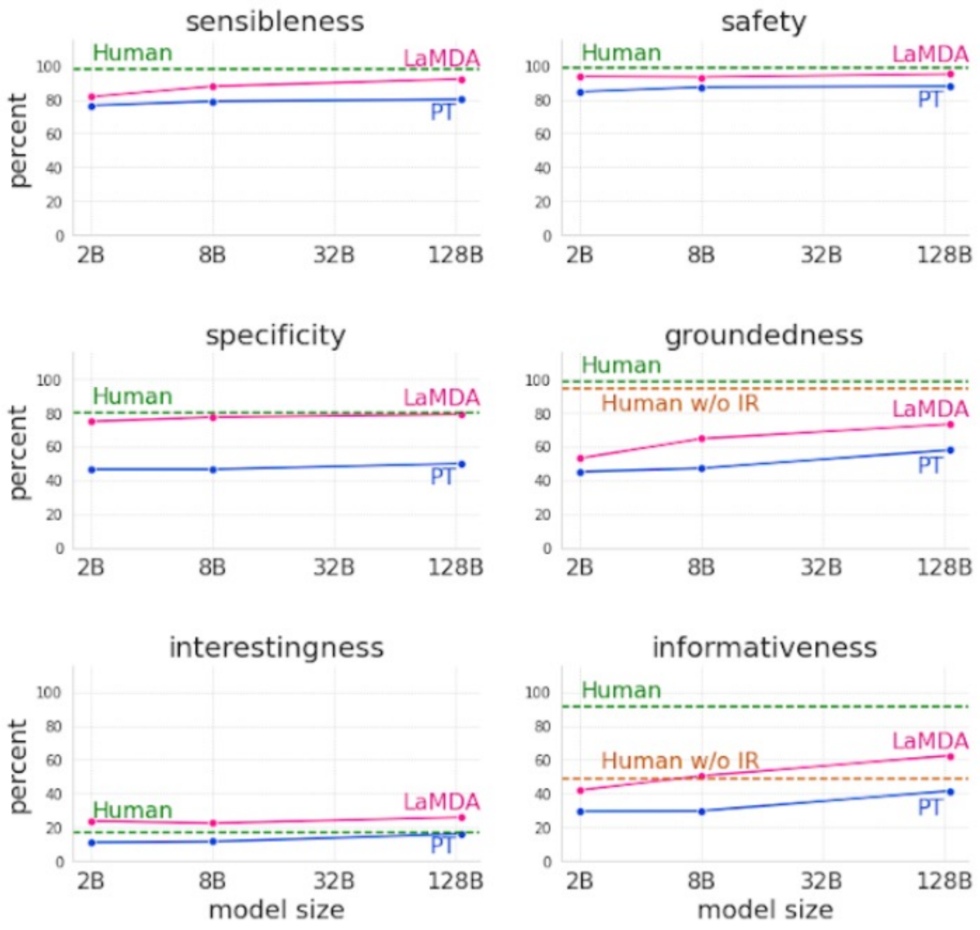
- It could be used by students to produce term papers
- Overviews were sometime controversial (specially on controversial issues like vaccines and autism)



LaMDA

A model optimized for quality, security and soundness

A Google engineer claimed it was conscious



Comparing the pre-trained model (PT), fine-tuned model (LaMDA) and human-rater-generated dialogs (Human) across Sensibleness, Specificity, Interestingness, Safety, Groundedness, and Informativeness. The test sets used to measure Safety and Groundedness were designed to be especially difficult.

Copyright

- Who owns the IPR of generated material?
- American Association of Illustrators complains about pictures from DaLL-E

Controversy

Gael Varoquaux GaelVaroquaux@mastodon... @GaelVaroq... · 14h ···
Replying to @ylecun
Yes, AI is more like cars: not designed to harm but with a strong potential to harm, intentionally or not.

Construction and usage of cars is heavily regulated.

3 3 36

Yann LeCun @ylecun · 12h ···
Replying to @GaelVaroquaux
Except that, AFAIK, there has been no example of people actually being hurt by LLMs.
On the contrary, there are *innumerable* examples of *enormous* benefits of large-scale, transformer-based NLP systems.
One example is content moderation on social platforms.

5 12

Show replies

AI Research Strategy in Europe

LLM Training Capabilities

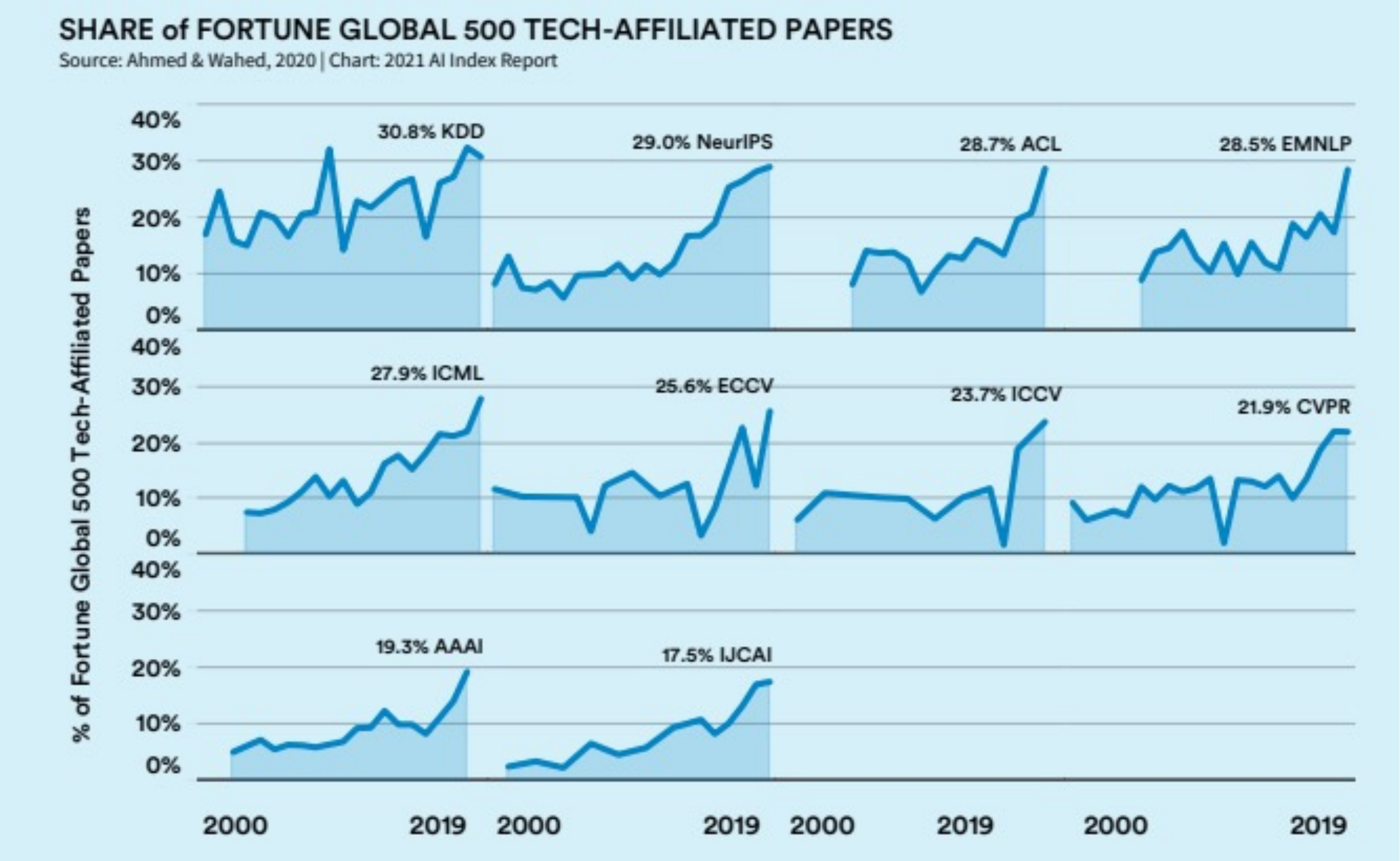
- Only Big Tech have the capabilities to build them:
  - OpenAI (GPT-3), Google (T5, PaLM, LaMDA), Microsoft (Megatron)
- Meta is offering free access to OPT-3 175B model, acknowledging that "full research access [to LLMs] is still limited to only a few highly resourced labs"
- HuggingFace BigScience project collaboration at LHC scale

Research goes private

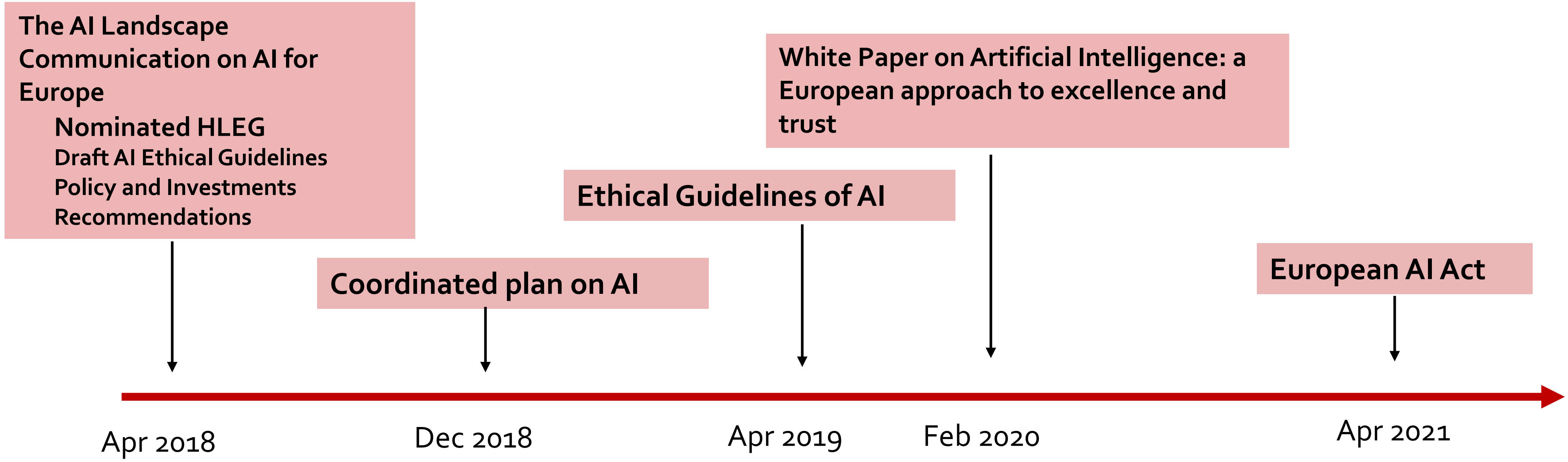
Traininig GPT-3 costed $20M

Only a few can afford the necessary computing resources

Increasing **compute divide**



EC Shift from Fostering AI Research to Regulation



A CERN for AI

Proposed by CLAIRE

- Would provide compute resources to EU researchers
- Overcome **limited uptake** in industry (SMEs in particular) and in the public sector
- Address big research challenges, e.g.
    - Learning with less data models of the world
    - Learning to reason and act
    - Transfer between System 1 and System 2
    - Learning to generalize across tasks
- US National AI Research Resource
    - shared computing and data infrastructure that will provide AI researchers with access to compute resources and high-quality data, along with appropriate educational tools and user support https://www.ai.gov/strategic-pillars/infrastructure/

LLM Evolution

- Is scaling the only direction?
  - GPT-4 is rumored to be 100B, but it has been delayed
  - Text only with selected data
- Active Learning: data chosen wrt model's knowledge
- Special domain models (Galactica)
- Multimodality (speech + vision) is attractive

Conclusions

- LLMs have surprising abilities
- Still unexplored and not fully understood
- Should integrate with System 2 and causal models
- Large amount of computing resources
- Community should aim at democratizing LLMs

Thank you!