COS 484

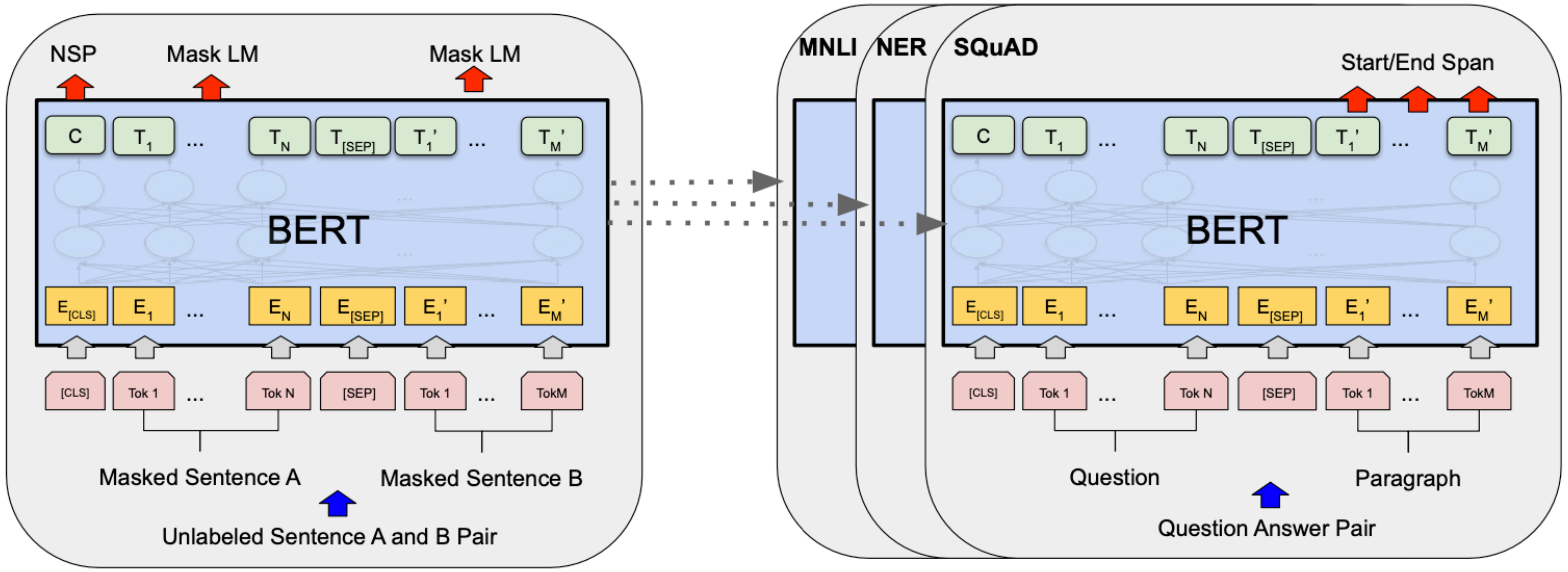
Natural Language Processing

L16: Pre-training and large language models (LLMs)

Spring 2024

Recap: Pretraining / fine-tuning

"Pre-train" a model on a large dataset for task X, then "fine-tune" it on a dataset for task Y



Pre-training

Fine-Tuning

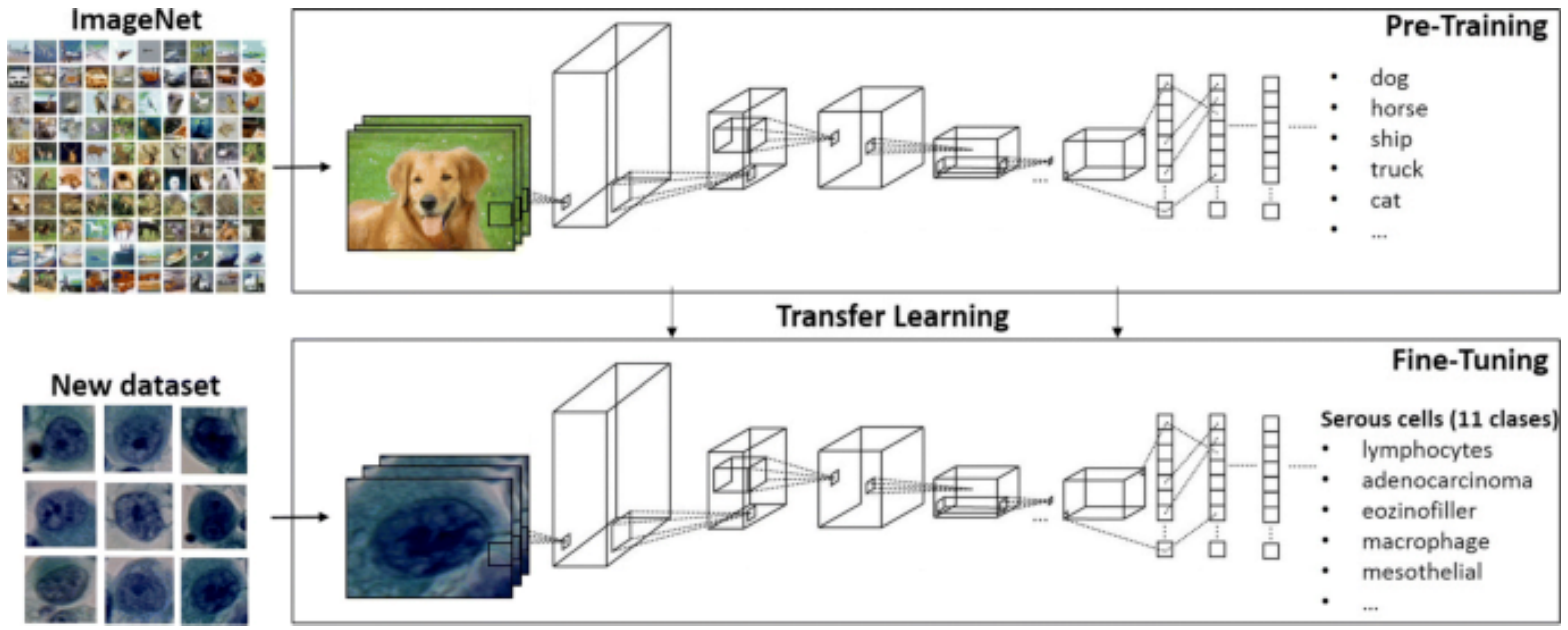
"**Fine-tuning** is the process of **taking the network learned by these pre-trained models**, and **further training the model**, often via an added neural net classifier that takes the top layer of the network as input, to perform some downstream task."

Fine-tuning is a training process and takes **gradient descent steps**!

Recap: Pretraining / fine-tuning

Pre-training

Fine-tuning



1.28M images, 1000 classes

3652 images, 11 classes

Pre-training

Natural language [MASK] (NLP) is an [MASK] subfield of linguistics, computer science, and artificial [MASK] concerned with the interactions [MASK] computers and human [MASK] …

→ processing, interdisciplinary, Intelligence, between, language

3.3B tokens
(512 tokens per segment)

Fine-tuning

contains no wit , only labored gags → negative

the greatest musicians → positive

very good viewing alternative → positive

67k examples, 2 classes

Recap: Pretraining / fine-tuning

Experiments on GLUE (Wang et al., 2019)

# of examples range between 2.5k and 392k examples

System	MNLI-(m/mm)	QQP	QNLI	SST-2	CoLA	STS-B	MRPC	RTE	**Average**
	392k	363k	108k	67k	8.5k	5.7k	3.5k	2.5k	-
Pre-OpenAI SOTA	80.6/80.1	66.1	82.3	93.2	35.0	81.0	86.0	61.7	74.0
BiLSTM+ELMo+Attn	76.4/76.1	64.8	79.8	90.4	36.0	73.3	84.9	56.8	71.0
OpenAI GPT	82.1/81.4	70.3	87.4	91.3	45.4	80.0	82.3	56.0	75.1
$\text{BERT}_{\text{BASE}}$	84.6/83.4	71.2	90.5	93.5	52.1	85.8	88.9	66.4	79.6
$\text{BERT}_{\text{LARGE}}$	**86.7/85.9**	**72.1**	**92.7**	**94.9**	**60.5**	**86.5**	**89.3**	**70.1**	**82.1**

Today we are going to see other uses of pre-trained models:

1) few-shot examples (e.g., 32)
2) No fine-tuning (= no gradient updates)

This lecture

- Post-BERT models of pre-training / fine-tuning
- GPT-3: prompting and in-context learning
- Instruction tuning, RLHF, ChatGPT, GPT-4, …
- Limitations of LLMs

Post-BERT models for pre-training/fine-tuning

RoBERTa

- BERT is still under-trained
- Removed the next sentence prediction pre-training — it adds more noise than benefits!
- Trained longer with 10x data & bigger batch sizes
- Pre-trained on 1,024 V100 GPUs for one day in 2019

Model	**data**	**bsz**	**steps**	**SQuAD** (v1.1/2.0)	**MNLI-m**	**SST-2**
RoBERTa						
with BOOKS + WIKI	16GB	8K	100K	93.6/87.3	89.0	95.3
+ additional data (§3.2)	160GB	8K	100K	94.0/87.7	89.3	95.6
+ pretrain longer	160GB	8K	300K	94.4/88.7	90.0	96.1
+ pretrain even longer	160GB	8K	500K	**94.6/89.4**	**90.2**	**96.4**
$\text{BERT}_{\text{LARGE}}$						
with BOOKS + WIKI	13GB	256	1M	90.9/81.8	86.6	93.7

(Liu et al., 2019): RoBERTa: A Robustly Optimized BERT Pretraining Approach

ALBERT

Key idea: **parameter sharing** across different layers + smaller embedding sizes

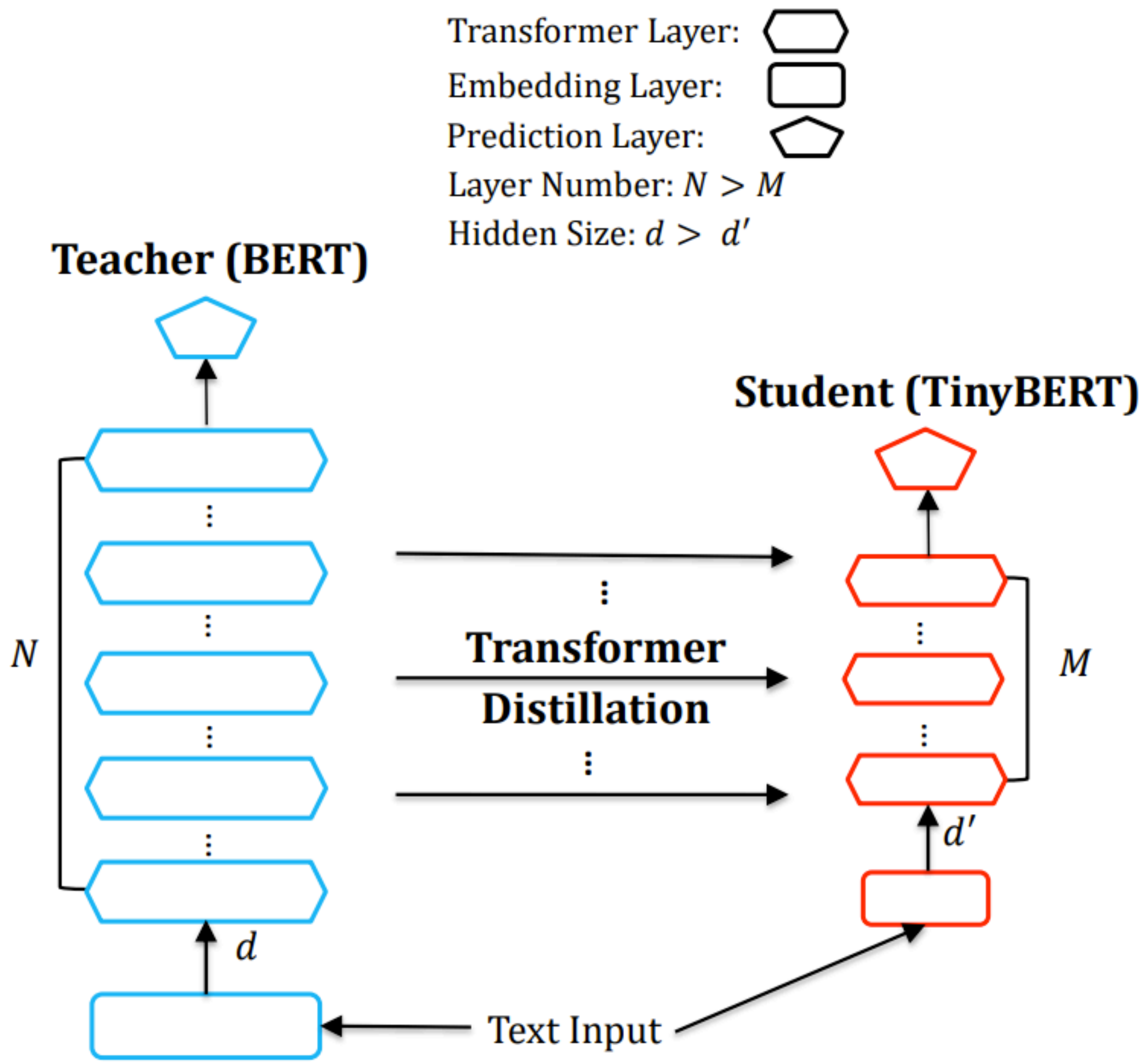
Model		Parameters	Layers	Hidden	Embedding	Parameter-sharing
BERT	base	108M	12	768	768	False
	large	334M	24	1024	1024	False
ALBERT	base	12M	12	768	128	True
	large	18M	24	1024	128	True
	xlarge	60M	24	2048	128	True
	xxlarge	235M	12	4096	128	True

Model		Parameters	SQuAD1.1	SQuAD2.0	MNLI	SST-2	RACE	Avg	Speedup
BERT	base	108M	90.4/83.2	80.4/77.6	84.5	92.8	68.2	82.3	4.7x
	large	334M	92.2/85.5	85.0/82.2	86.6	93.0	73.9	85.2	1.0
ALBERT	base	12M	89.3/82.3	80.0/77.1	81.6	90.3	64.0	80.1	5.6x
	large	18M	90.6/83.9	82.3/79.4	83.5	91.7	68.5	82.4	1.7x
	xlarge	60M	92.5/86.1	86.1/83.1	86.4	92.4	74.8	85.5	0.6x
	xxlarge	235M	**94.1/88.3**	**88.1/85.1**	**88.0**	**95.2**	**82.3**	**88.7**	0.3x

AlBERT models have less # of parameters (less storage), but they can be slower because the model architectures are larger

(Lan et al., 2020): ALBERT: A Lite BERT for Self-supervised Learning of Language Representations

DistillBERT / TinyBERT / MobileBERT



https://github.com/abhilash1910/DistilBERT--SQuAD-v1-Notebook

Key idea: produce a smaller model (student) that distill information from the BERT models (teacher)

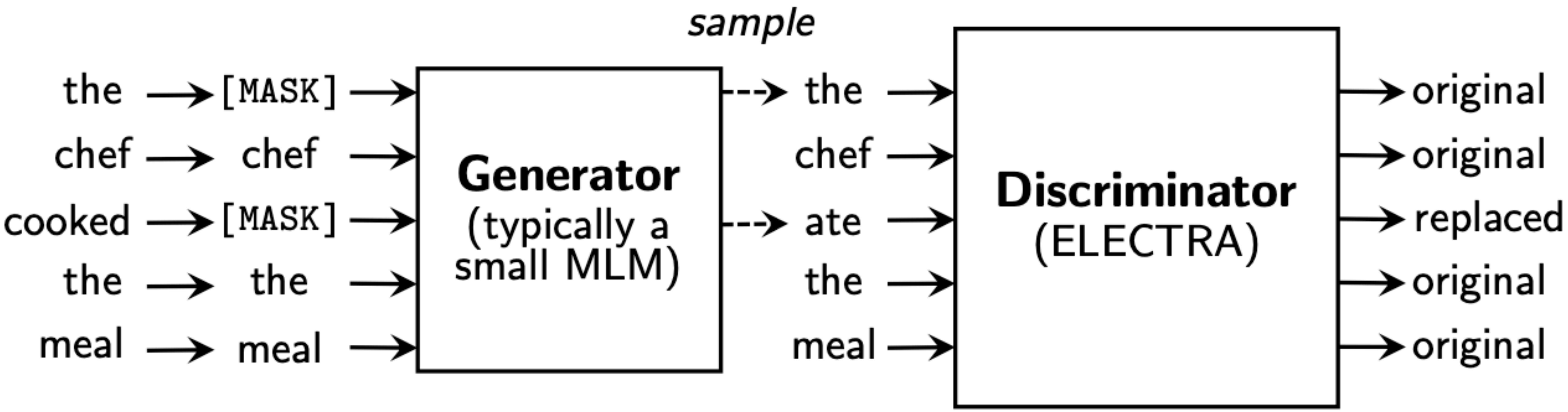
Table 1: **DistilBERT retains 97% of BERT performance.** Comparison on the dev sets of the GLUE benchmark. ELMo results as reported by the authors. BERT and DistilBERT results are the medians of 5 runs with different seeds.

Model	**Score**	CoLA	MNLI	MRPC	QNLI	QQP	RTE	SST-2	STS-B	WNLI
ELMo	68.7	44.1	68.6	76.6	71.1	86.2	53.4	91.5	70.4	56.3
BERT-base	79.5	56.3	86.7	88.6	91.8	89.6	69.3	92.7	89.0	53.5
DistilBERT	77.0	51.3	82.2	87.5	89.2	88.5	59.9	91.3	86.9	56.3

(Sanh et al., 2019): DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter

ELECTRA

ELECTRA provides a more **efficient** training method,
because it predicts 100% of tokens (instead of 15%) every time

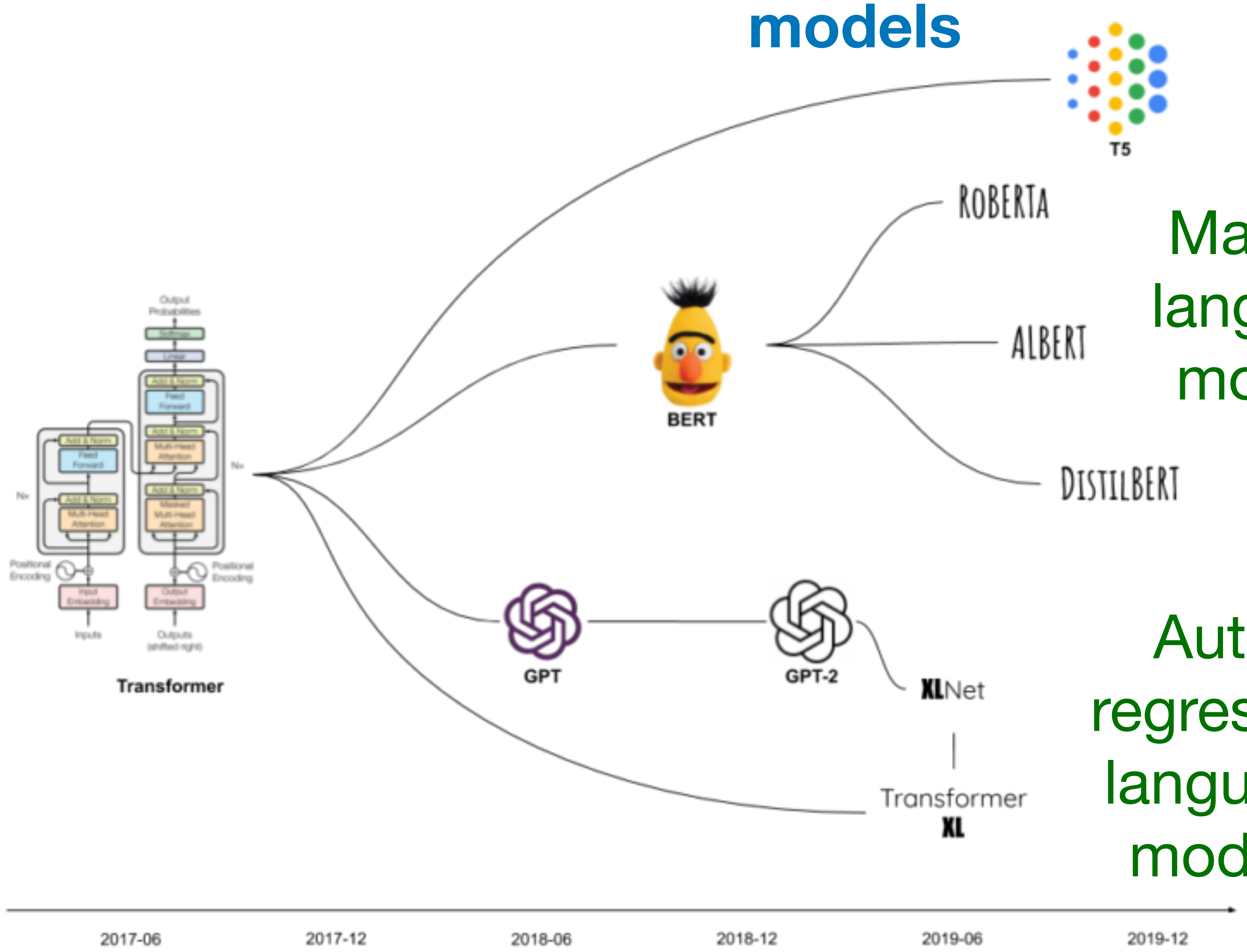


Only the discriminator will be used for downstream fine-tuning

(Clark et al., 2020): ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators

Three major forms of pre-training

Text-to-text models



Masked language models

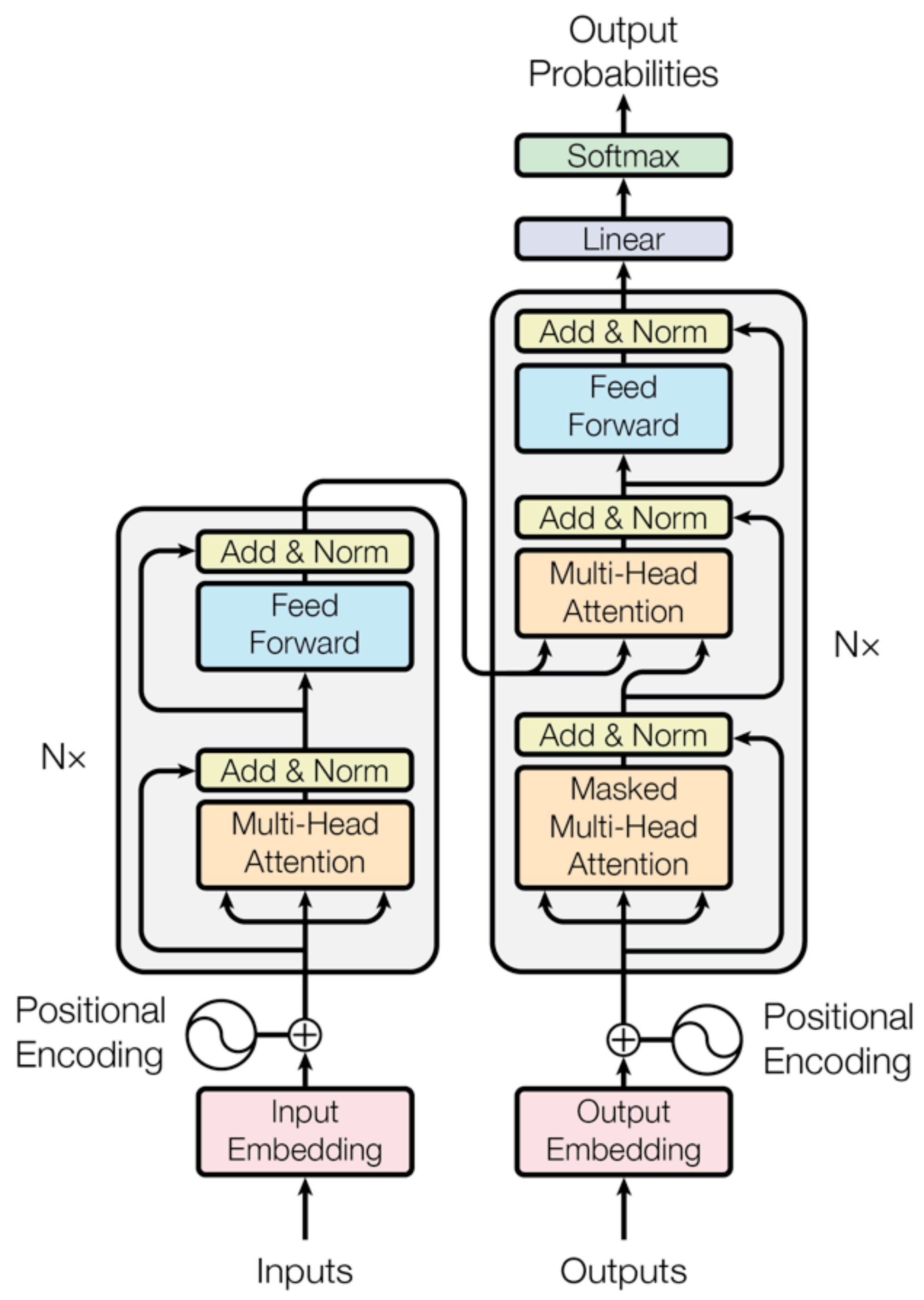
Auto-regressive language models

- Masked language models = Transformer encoder
- Autoregressive language models = Transformer decoder
- **Text-to-text models = Transformer encoder-decoder**

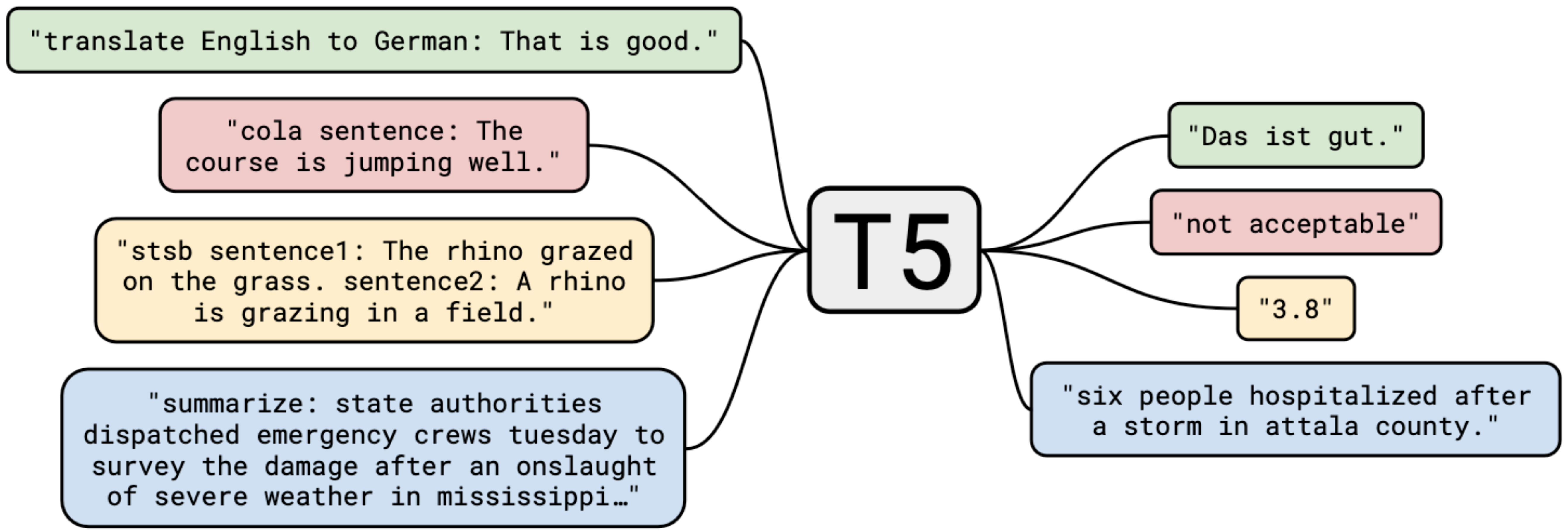
https://www.factored.ai/2021/09/21/an-intuitive-explanation-of-transformer-based-models/

Text-to-text models

- So far, **encoder-only models (e.g., BERT)** enjoy the benefits of **bidirectionality** but they can't be used to generate text
- **Decoder-only models (e.g., GPT)** can do generation but they are left-to-right LMs..
- Text-to-text models combine the best of both worlds!

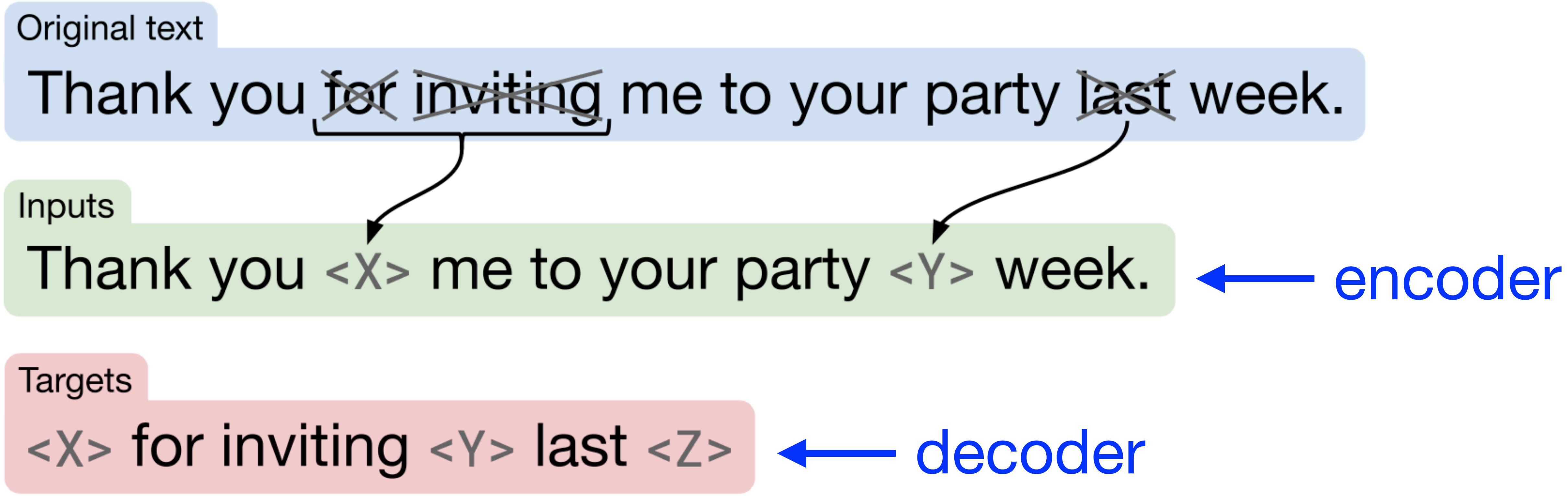


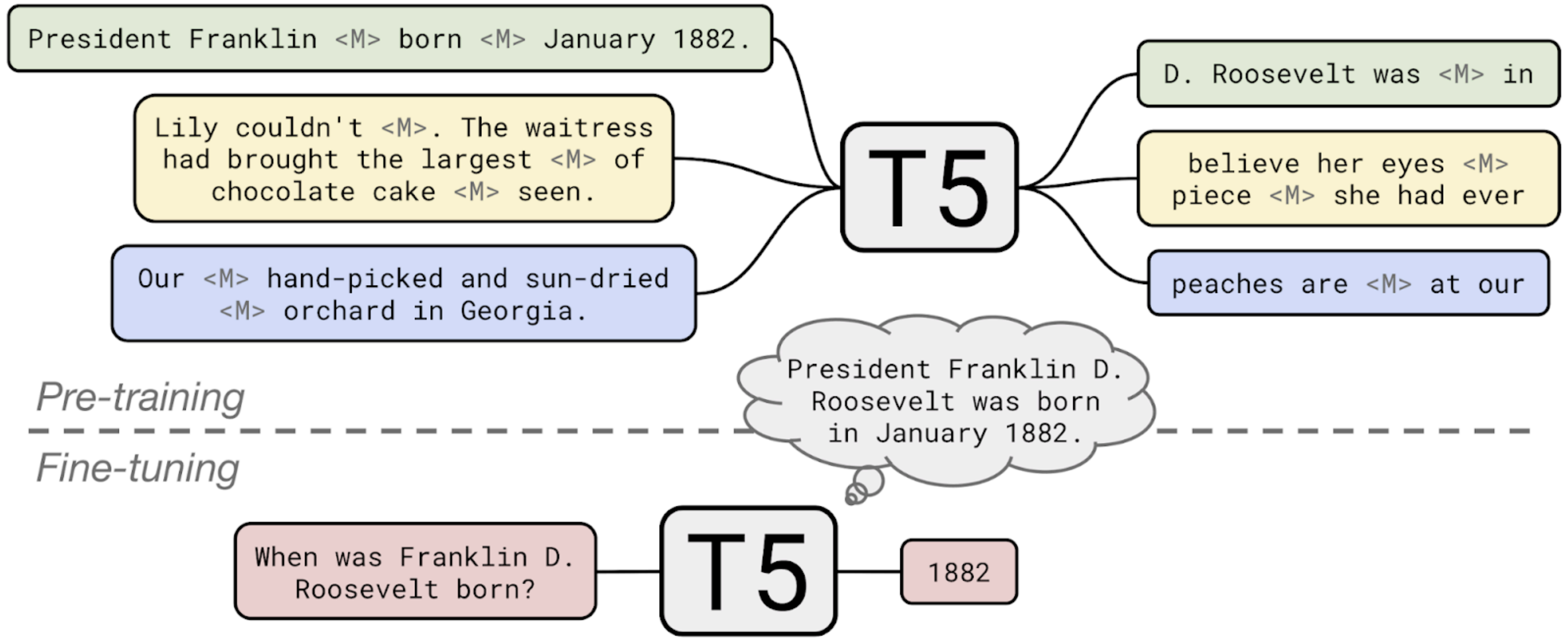
T5 = <u>T</u>ext-<u>t</u>o-<u>T</u>ext <u>T</u>ransfer <u>T</u>ransformer



(Raffel et al., 2020): Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer

T5 models





T5 comes in different sizes:

- t5-small.
- t5-base.
- t5-large.
- t5-3b.
- t5-11b.

(Raffel et al., 2020): Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer

How to use these pre-trained models?

Transformers

Transformers

Search documentation ⌘K

V4.27.2 EN 92,354

- CANINE
- CodeGen
- ConvBERT
- CPM
- CTRL
- DeBERTa
- DeBERTa-v2
- DialoGPT
- **DistilBERT**
- DPR
- ELECTRA

DistilBERT

All model pages distilbert | Hugging Face Spaces

Overview

The DistilBERT model was proposed in the blog post Smaller, faster, cheaper, lighter: Introducing DistilBERT, a distilled version of BERT, and the paper DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. DistilBERT is a small, fast, cheap and light Transformer model trained by distilling BERT base. It has 40% less parameters than *bert-base-uncased*, runs 60% faster while preserving over 95% of BERT's performances as measured on the GLUE language understanding benchmark.

```python
>>> from transformers import AutoTokenizer

>>> tokenizer = AutoTokenizer.from_pretrained("bert-base-cased")


>>> def tokenize_function(examples):
...     return tokenizer(examples["text"], padding="max_length", truncation=True)


>>> tokenized_datasets = dataset.map(tokenize_function, batched=True)
```

```python
>>> from transformers import AutoModelForSequenceClassification

>>> model = AutoModelForSequenceClassification.from_pretrained("bert-base-cased", num_labels=5)
```

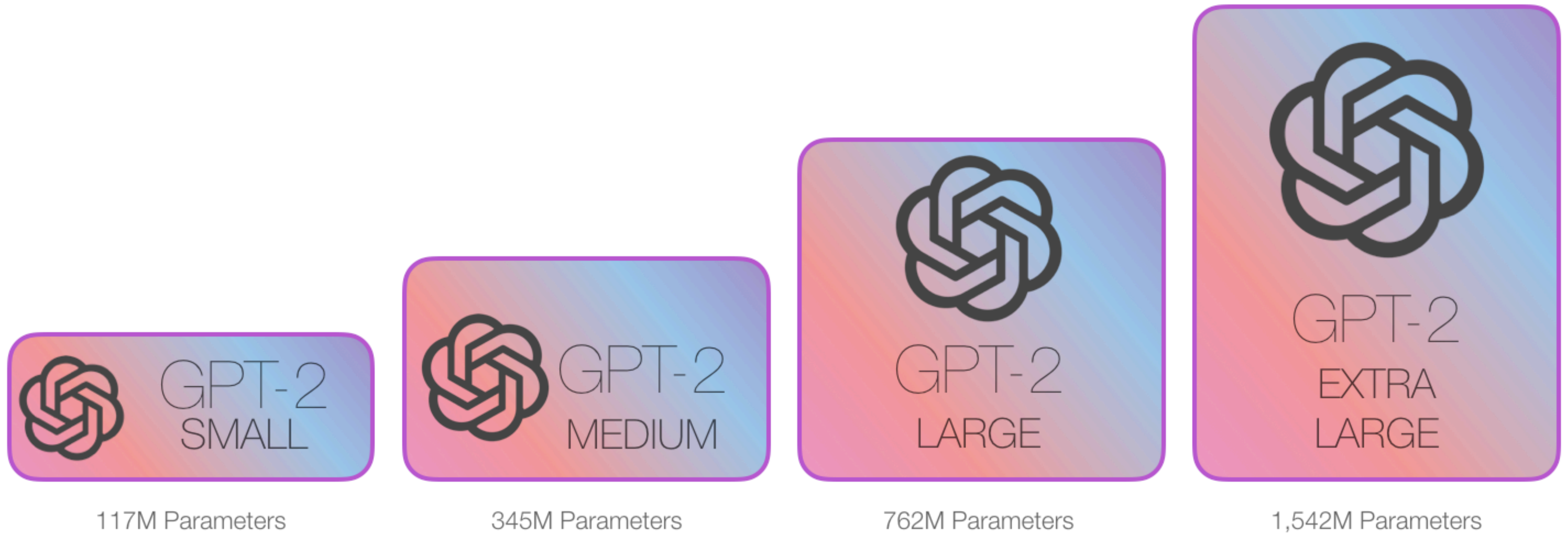
GPT-3: Prompting and In-context Learning

From GPT to GPT-2 to GPT-3

- All **decoder-only Transformer-based language models**
- Model size ↑, training corpora ↑

GPT-2

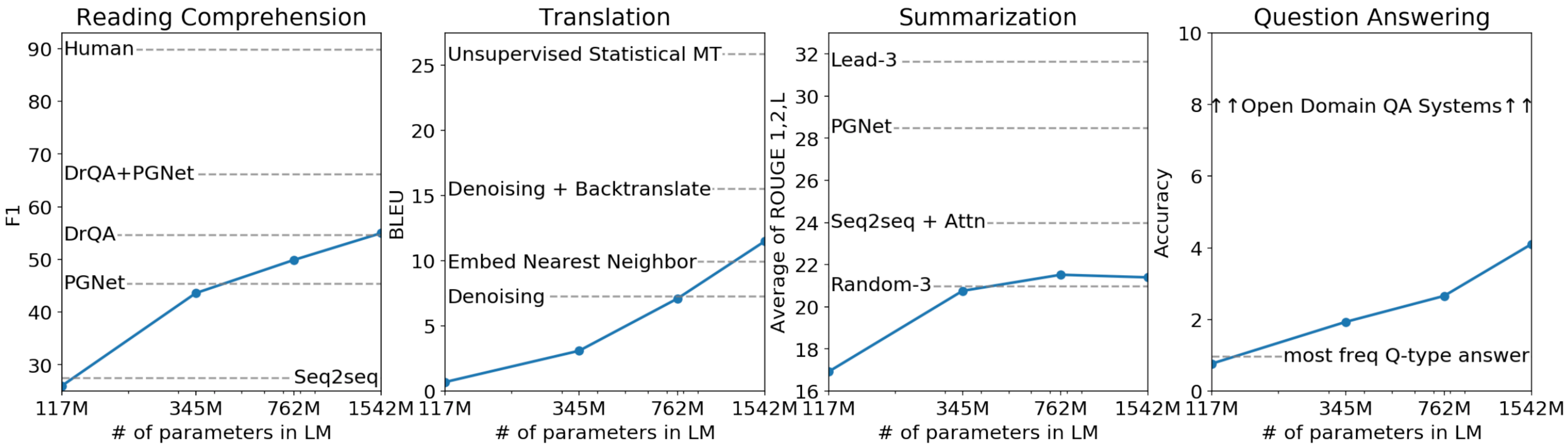
Context size = 1024

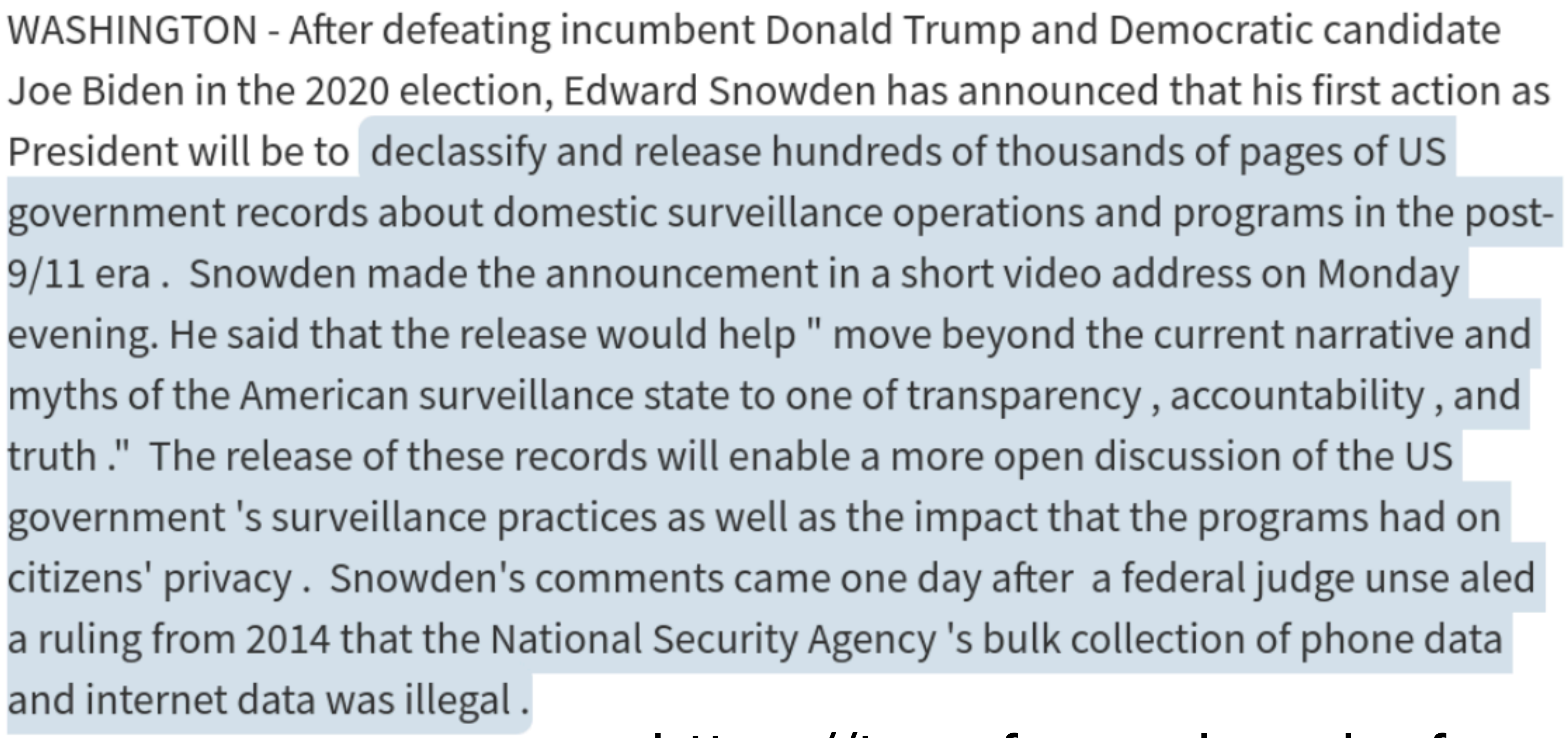


.. trained on 40Gb of Internet text ..

(Radford et al., 2019): Language Models are Unsupervised Multitask Learners

GPT-2 started to achieve strong zero-shot performance



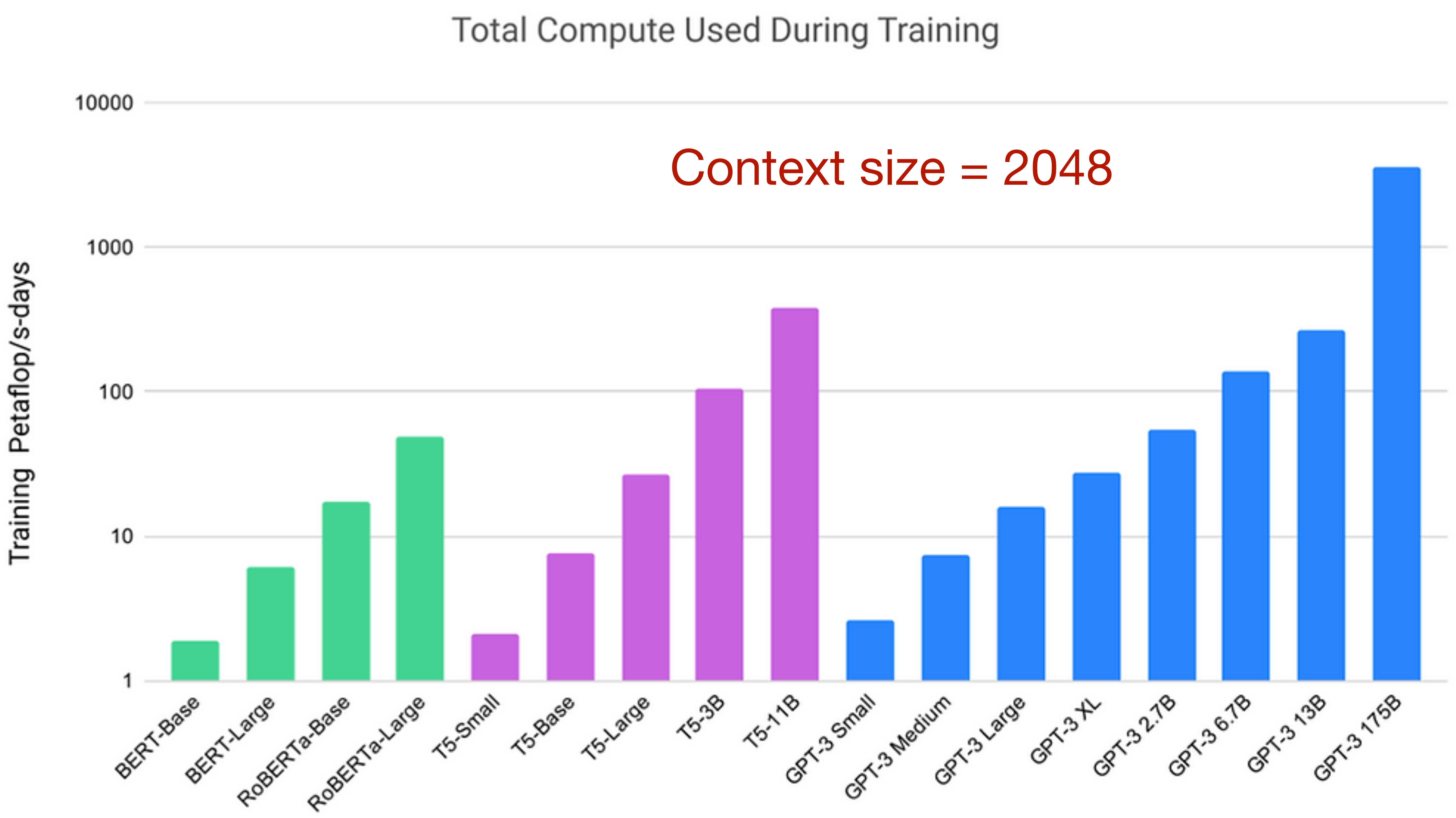


https://transformer.huggingface.co/doc/gpt2-large

(Radford et al., 2019): Language Models are Unsupervised Multitask Learners

GPT-3: language models are few-shot learners

- GPT-2 → GPT-3: 1.5B → 175B (# of parameters), ~14B → 300B (# of tokens)



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

Paradigm shift since GPT-3

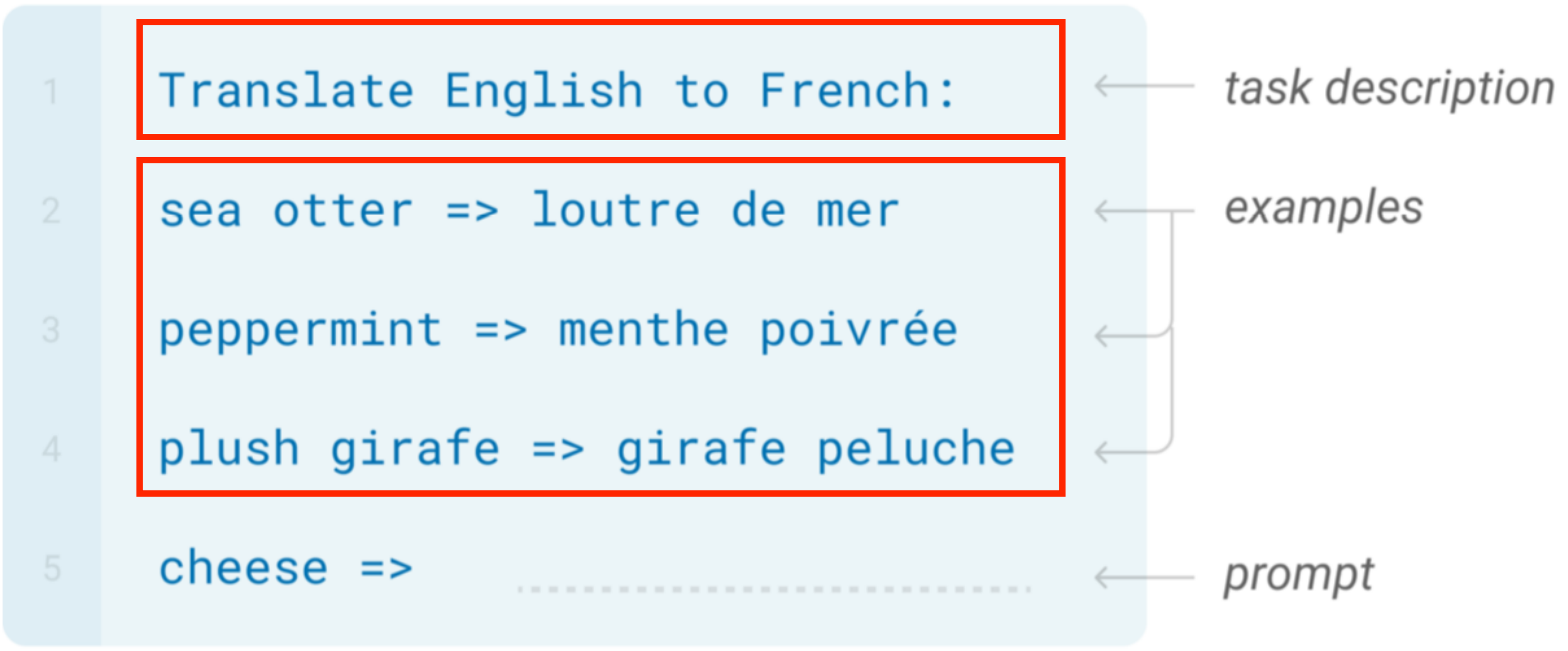
- Before GPT-3, **fine-tuning** is the default way of doing learning in models like BERT/T5/GPT-2
  - SST-2 has 67k examples, SQuAD has 88k (passage, answer, question) triples
- Fine-tuning requires computing the gradient and applying a parameter update on every example (or every K examples in a mini-batch)
- However, this is very expensive for the 175B GPT-3 model

GPT-3: Few-shot learning

- GPT-3 proposes an alternative: **in-context learning**

Few-shot

In addition to the task description, the model sees a few examples of the task. No gradient updates are performed.



- This is just a forward pass, **no gradient update at all**!
- You only need to feed a small number of examples (e.g., 32)

(On the other hand, you can't feed many examples at once too as it is bounded by context size)

GPT-3: task specifications

Context →	Passage: Saint Jean de Brébeuf was a French Jesuit missionary who travelled to New France in 1625. There he worked primarily with the Huron for the rest of his life, except for a few years in France from 1629 to 1633. He learned their language and culture, writing extensively about each to aid other missionaries. In 1649, Brébeuf and another missionary were captured when an Iroquois raid took over a Huron village . Together with Huron captives, the missionaries were ritually tortured and killed on March 16, 1649. Brébeuf was beatified in 1925 and among eight Jesuit missionaries canonized as saints in the Roman Catholic Church in 1930. Question: How many years did Saint Jean de Brébeuf stay in New France before he went back to France for a few years? Answer:
Target Completion →	4

DROP
(a reading comprehension task)

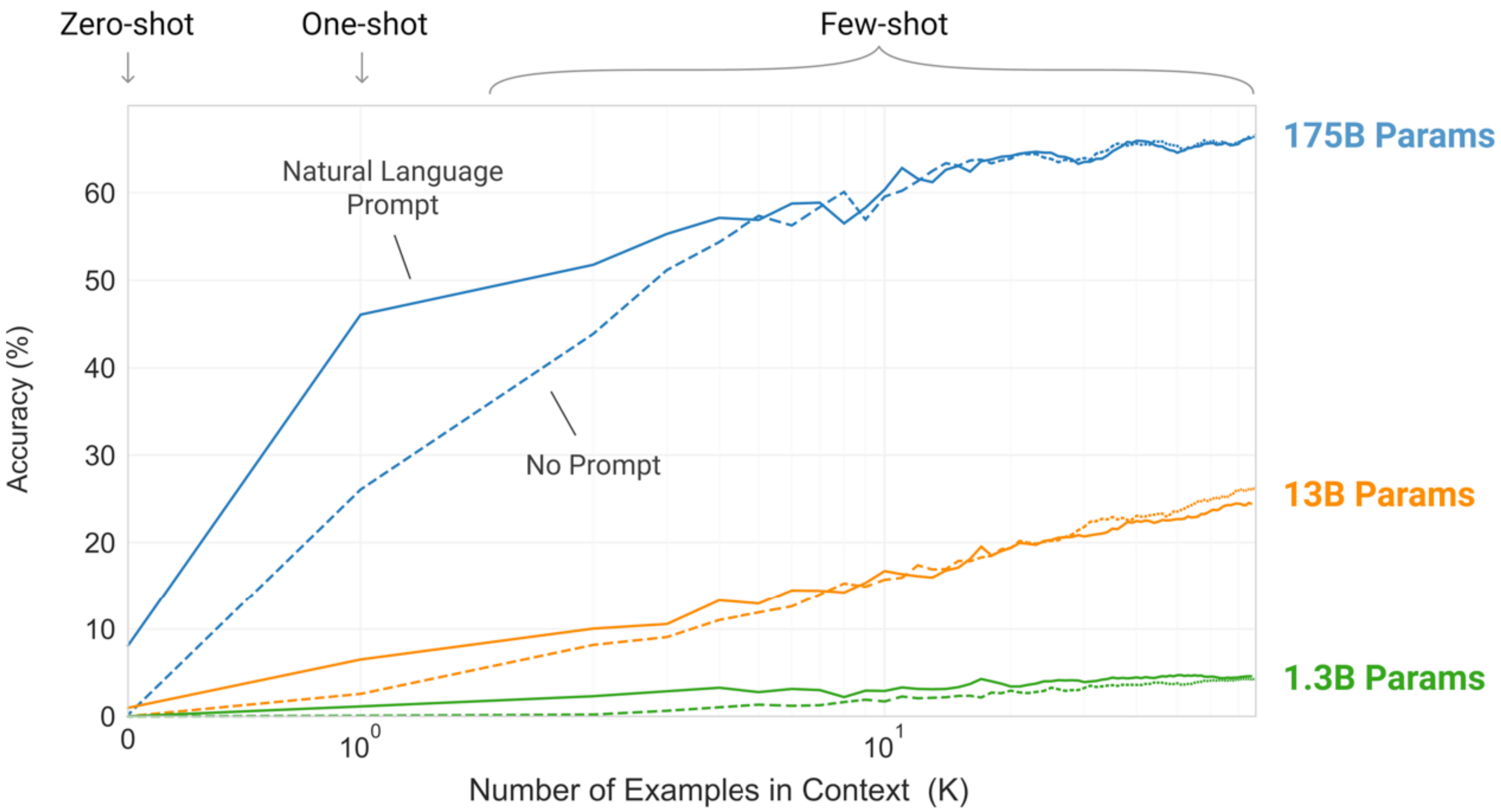
Context →	Please unscramble the letters into a word, and write that word: skicts =
Target Completion →	sticks

Unscrambling words

Context →	An outfitter provided everything needed for the safari. Before his first walking holiday, he went to a specialist outfitter to buy some boots. question: Is the word 'outfitter' used in the same way in the two sentences above? answer:
Target Completion →	no

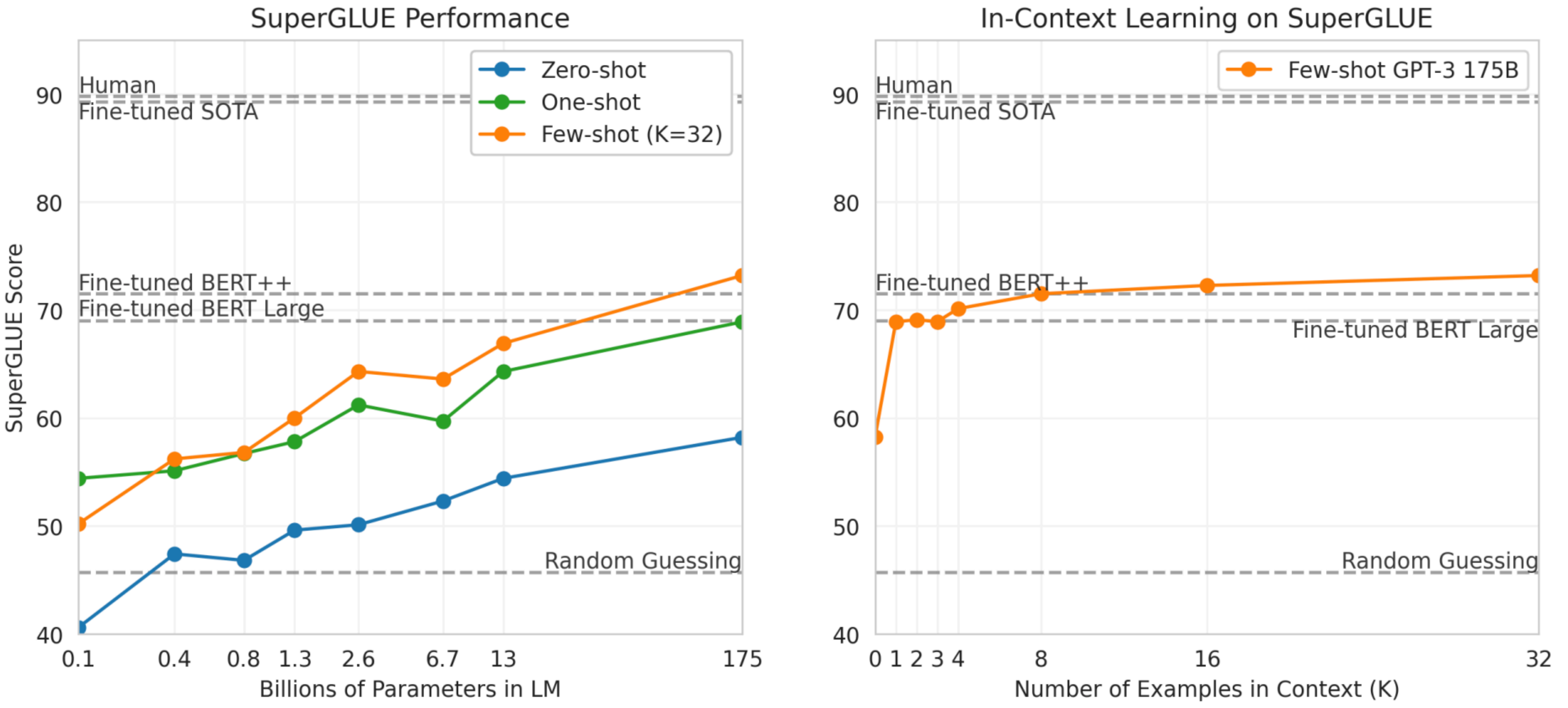
Word in context (WiC)

GPT-3's in-context learning



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

GPT-3 performance on SuperGLUE



(Wang et al., 2019) SuperGLUE: A Stickier Benchmark for General-Purpose Language Understanding Systems

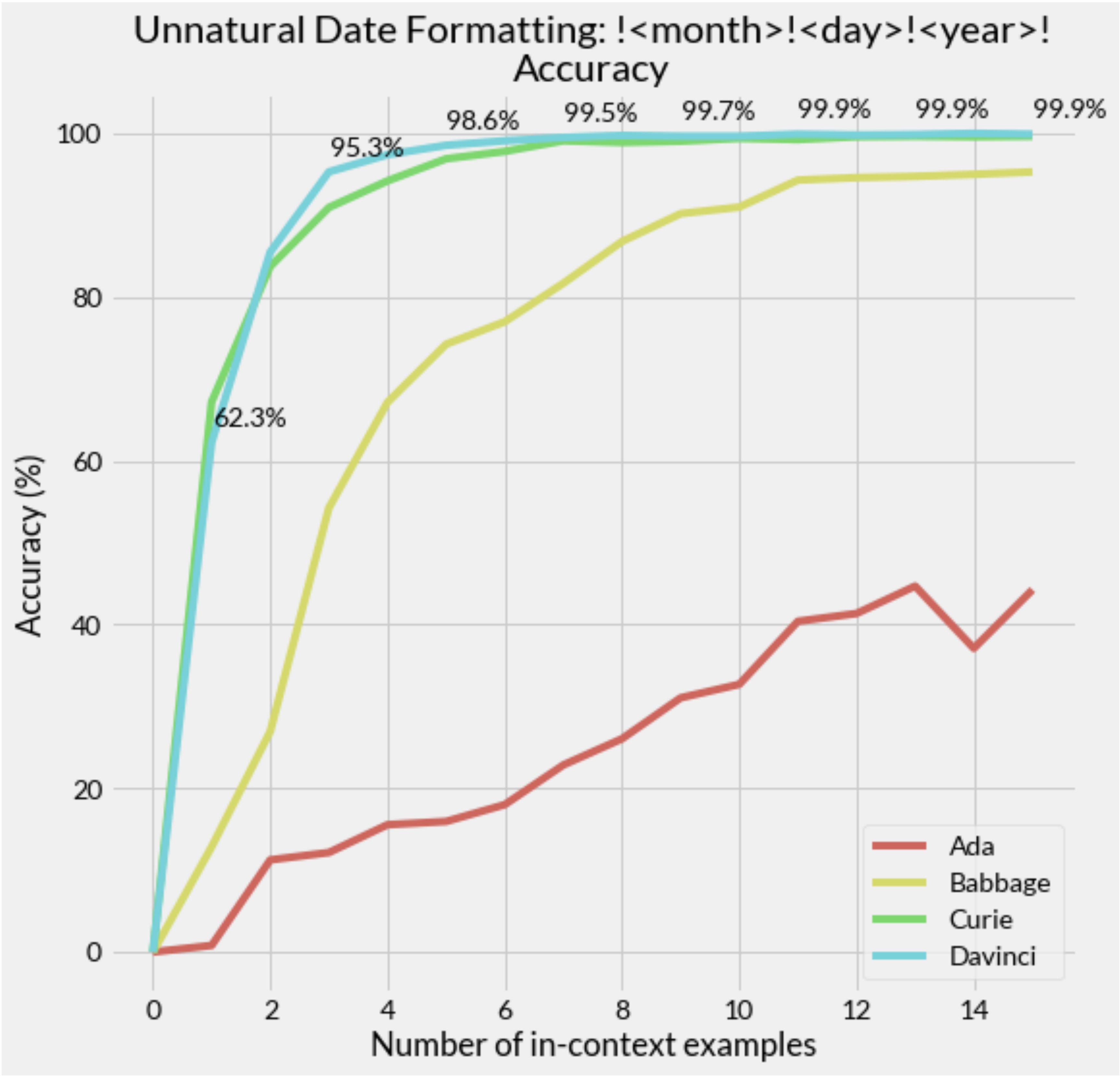
GPT-3's in-context learning

```
Input: 2014-06-01
Output: !06!01!2014!
Input: 2007-12-13
Output: !12!13!2007!
Input: 2010-09-23
Output: !09!23!2010!
Input: 2005-07-23
Output: !07!23!2005!
```

in-context examples

test example

model completion



http://ai.stanford.edu/blog/in-context-learning/

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

Chain-of-thought (CoT) prompting

Standard Prompting

Input

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

A: The answer is 11.

Q: The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?

Model Output

A: The answer is 27.

Chain of Thought Prompting

Input

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

A: Roger started with 5 balls. 2 cans of 3 tennis balls each is 6 tennis balls. 5 + 6 = 11. The answer is 11.

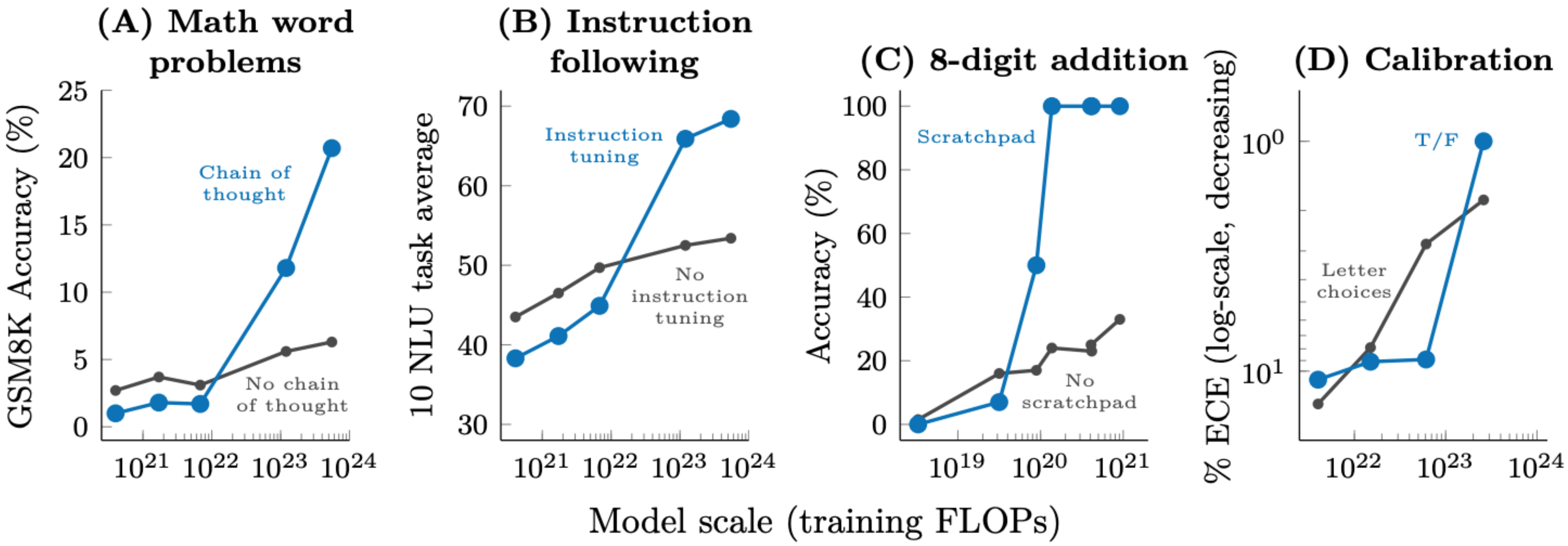
Q: The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?

Model Output

A: The cafeteria had 23 apples originally. They used 20 to make lunch. So they had 23 - 20 = 3. They bought 6 more apples, so they have 3 + 6 = 9. The answer is 9.

(Wei et al., 2022): Chain-of-Thought Prompting Elicits Reasoning in Large Language Models

Emergent properties of LLMs

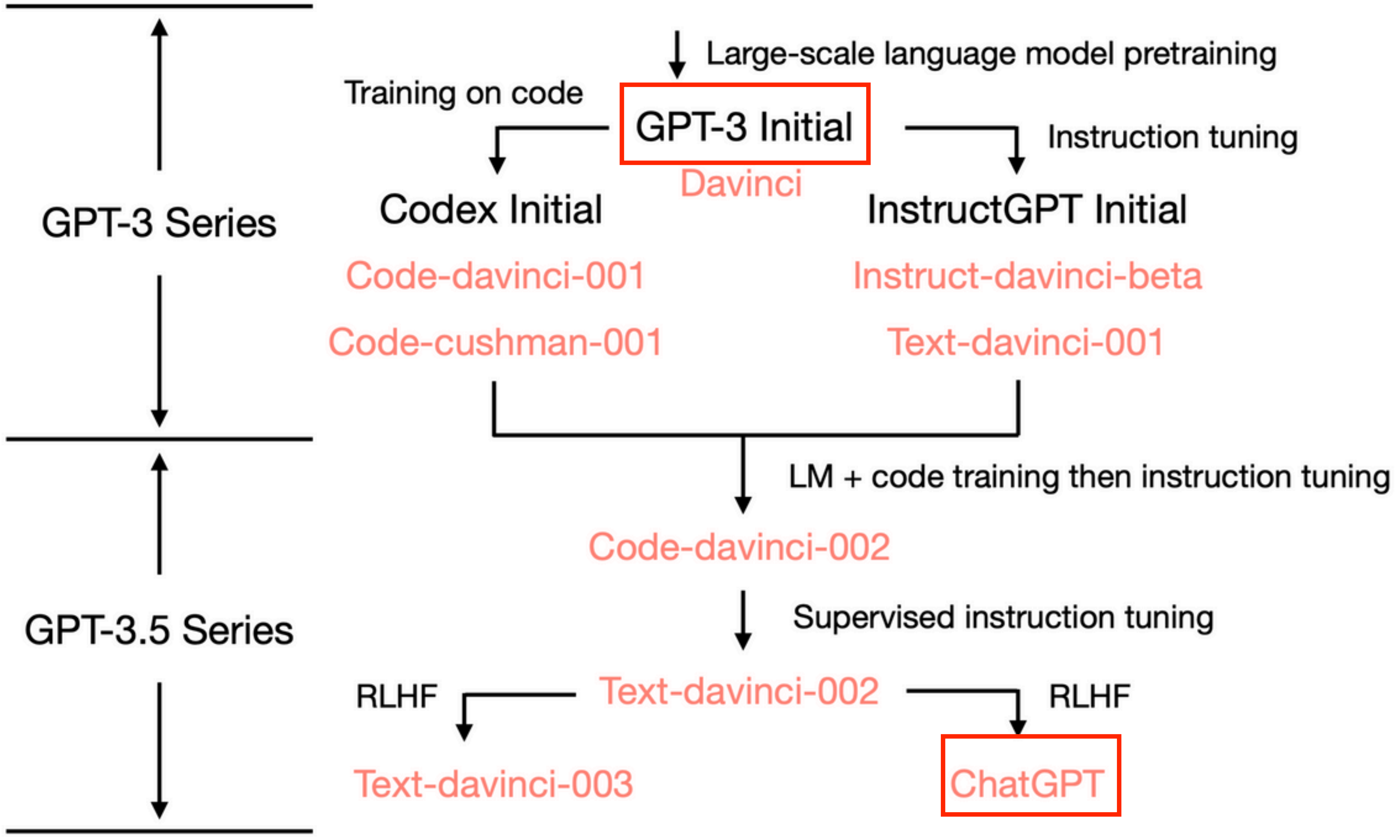


(Wei et al., 2022) Emergent Abilities of Large Language Models

What happened after GPT-3?

(Is model size $\uparrow$, training corpora $\uparrow$ the only way to go?)

How was ChatGPT developed?

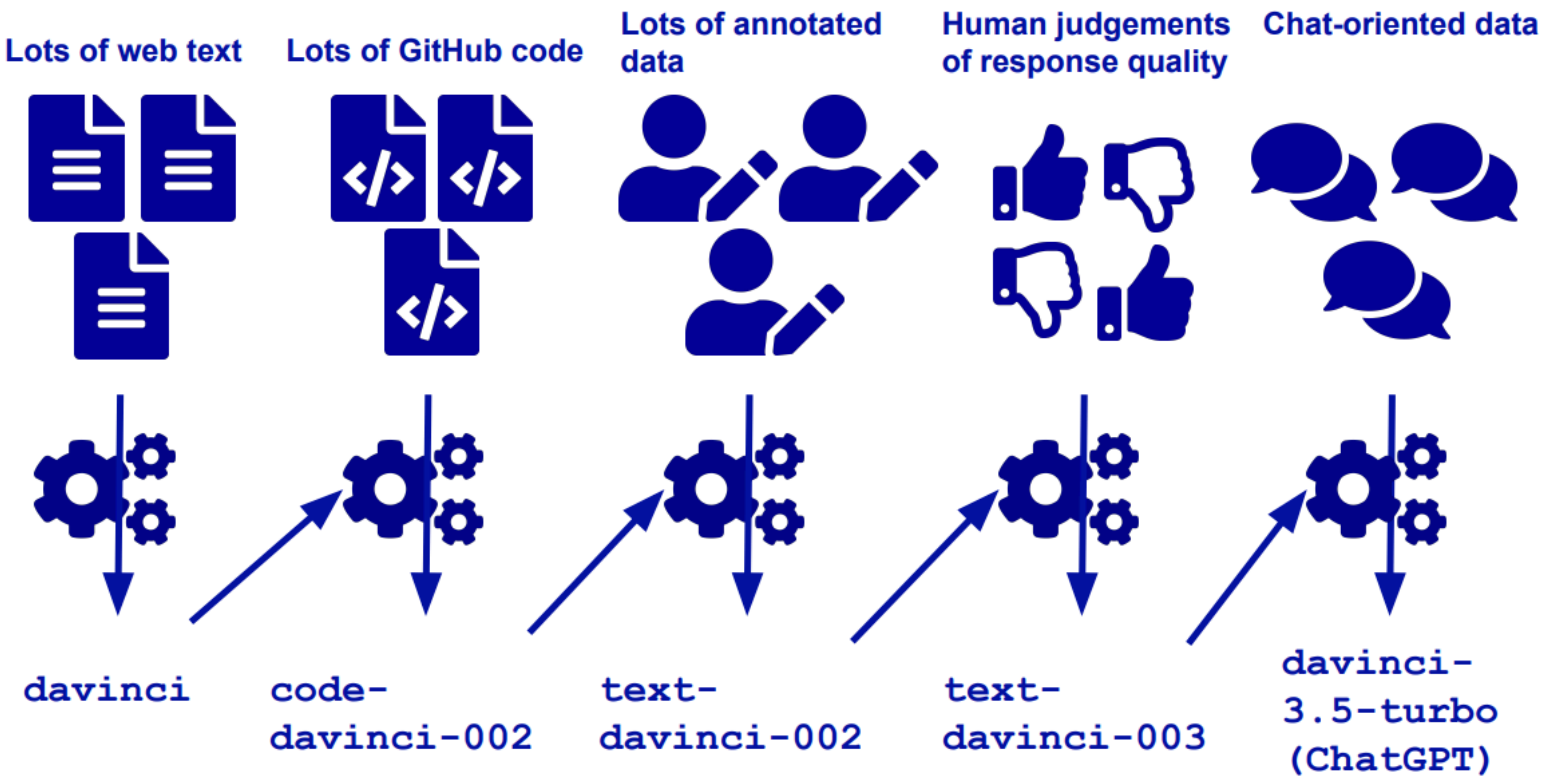


What's new?

- Training on code
- **Supervised instruction tuning**
- **RLHF = Reinforcement learning from human feedback**

https://yaofu.notion.site/How-does-GPT-Obtain-its-Ability-Tracing-Emergent-Abilities-of-Language-Models-to-their-Sources-b9a57ac0fcf74f30a1ab9e3e36fa1dc1

How was ChatGPT developed?

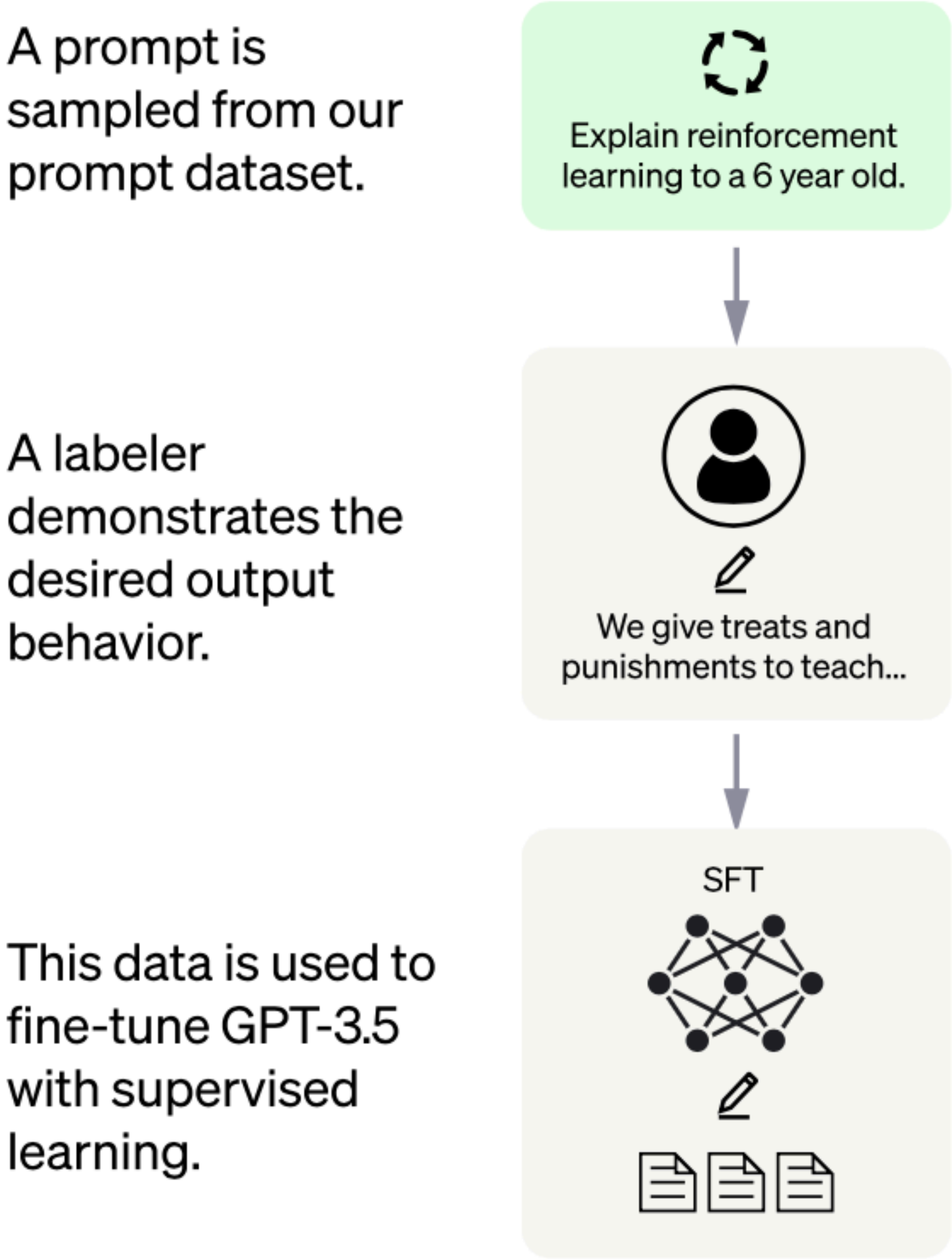


(Slide credit: Graham Neubig)

InstructGPT: Supervised instruction tuning + RLHF

Step 1

Collect demonstration data and train a supervised policy.



(Ouyang et al., 2022): Training language models to follow instructions with human feedback

Supervised instruction tuning

Use-case	Prompt
Brainstorming	List five ideas for how to regain enthusiasm for my career
Generation	Write a short story where a bear goes to the beach, makes friends with a seal, and then returns home.
Rewrite	This is the summary of a Broadway play: """ {summary} """ This is the outline of the commercial for that play: """

Use-case	(%)
Generation	45.6%
Open QA	12.4%
Brainstorming	11.2%
Chat	8.4%
Rewrite	6.6%
Summarization	4.2%
Classification	3.5%
Other	3.5%
Closed QA	2.6%
Extract	1.9%

Number of Prompts

SFT Data		
split	source	size
train	labeler	11,295
train	customer	1,430
valid	labeler	1,550
valid	customer	103

SFT data: only ~13k (not public)

InstructGPT: Supervised instruction tuning + RLHF

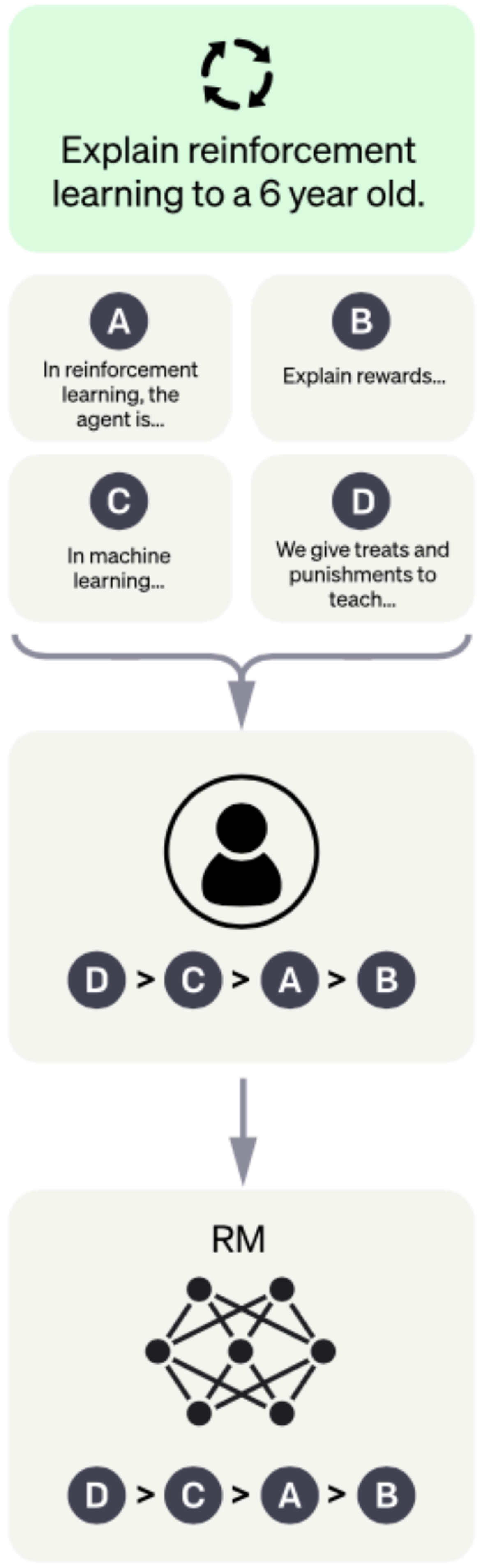
Step 2

Collect comparison data and train a reward model.

A prompt and several model outputs are sampled.

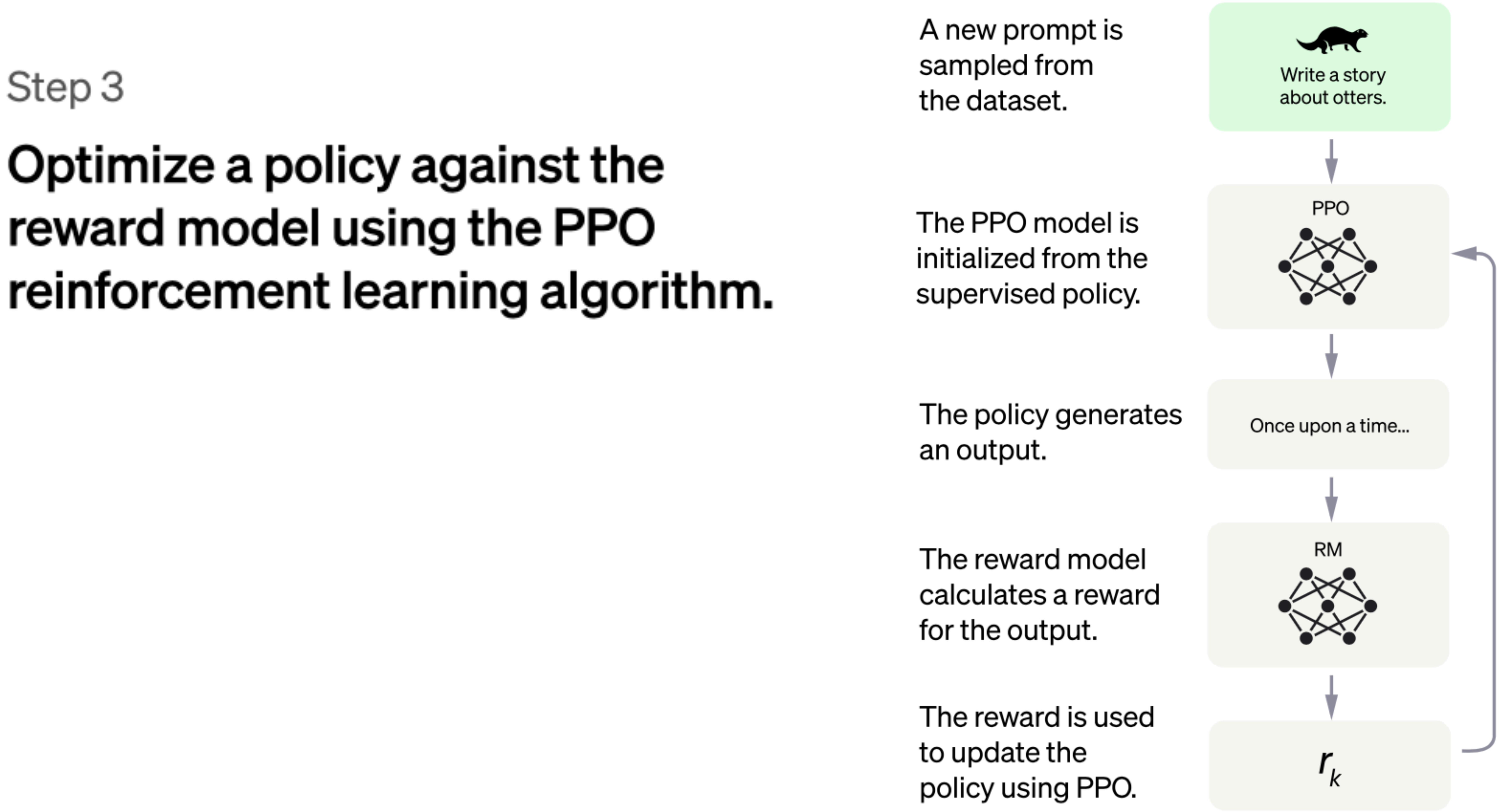
A labeler ranks the outputs from best to worst.

This data is used to train our reward model.



(Ouyang et al., 2022): Training language models to follow instructions with human feedback

InstructGPT: Supervised instruction tuning + RLHF



(Ouyang et al., 2022): Training language models to follow instructions with human feedback

ChatGPT = InstructGPT + dialogue data

Introducing ChatGPT

We've trained a model called ChatGPT which interacts in a conversational way. The dialogue format makes it possible for ChatGPT to answer followup questions, admit its mistakes, challenge incorrect premises, and reject inappropriate requests.

"We trained this model using Reinforcement Learning from Human Feedback (RLHF), **using the same methods as InstructGPT**, but with slight differences in the data collection setup. We trained an initial model using supervised fine-tuning: human AI trainers provided conversations in which they played both sides—the user and an AI assistant. We gave the trainers access to model-written suggestions to help them compose their responses. **We mixed this new dialogue dataset with the InstructGPT dataset**, which we transformed into a dialogue format."

Human feedback data is the key!

https://openai.com/blog/chatgpt

Recent models are getting smaller?

RESEARCH

Introducing LLaMA: A foundational, 65-billion-parameter large language model

February 24, 2023

- **Smaller models** trained on **1.4T**, high-quality & publicly available data
- The models are public
- "LLaMA-13B outperforms GPT-3 (175B) on most benchmarks, and LLaMA-65B is competitive with the best models, Chinchilla-70B and PaLM-540B"

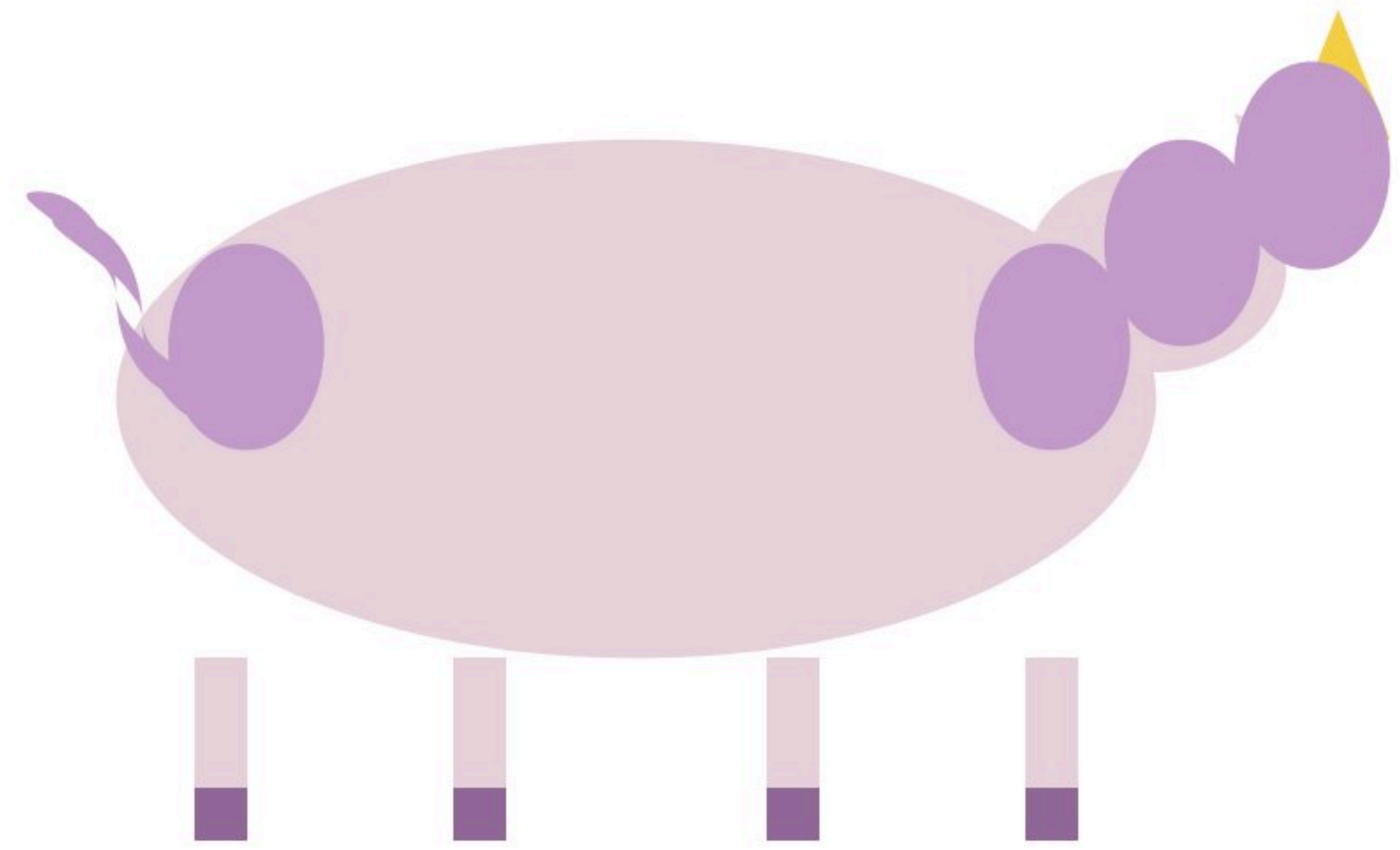
(Touvron et al., 2023): LLaMA: Open and Efficient Foundation Language Models

GPT-4

- GPT-4: a **multi-modal** model capable of processing image and text inputs and producing text outputs.
- Model size and training details unknown
- Can process up to 32k context size

Prompt: Draw a unicorn in TiKZ.

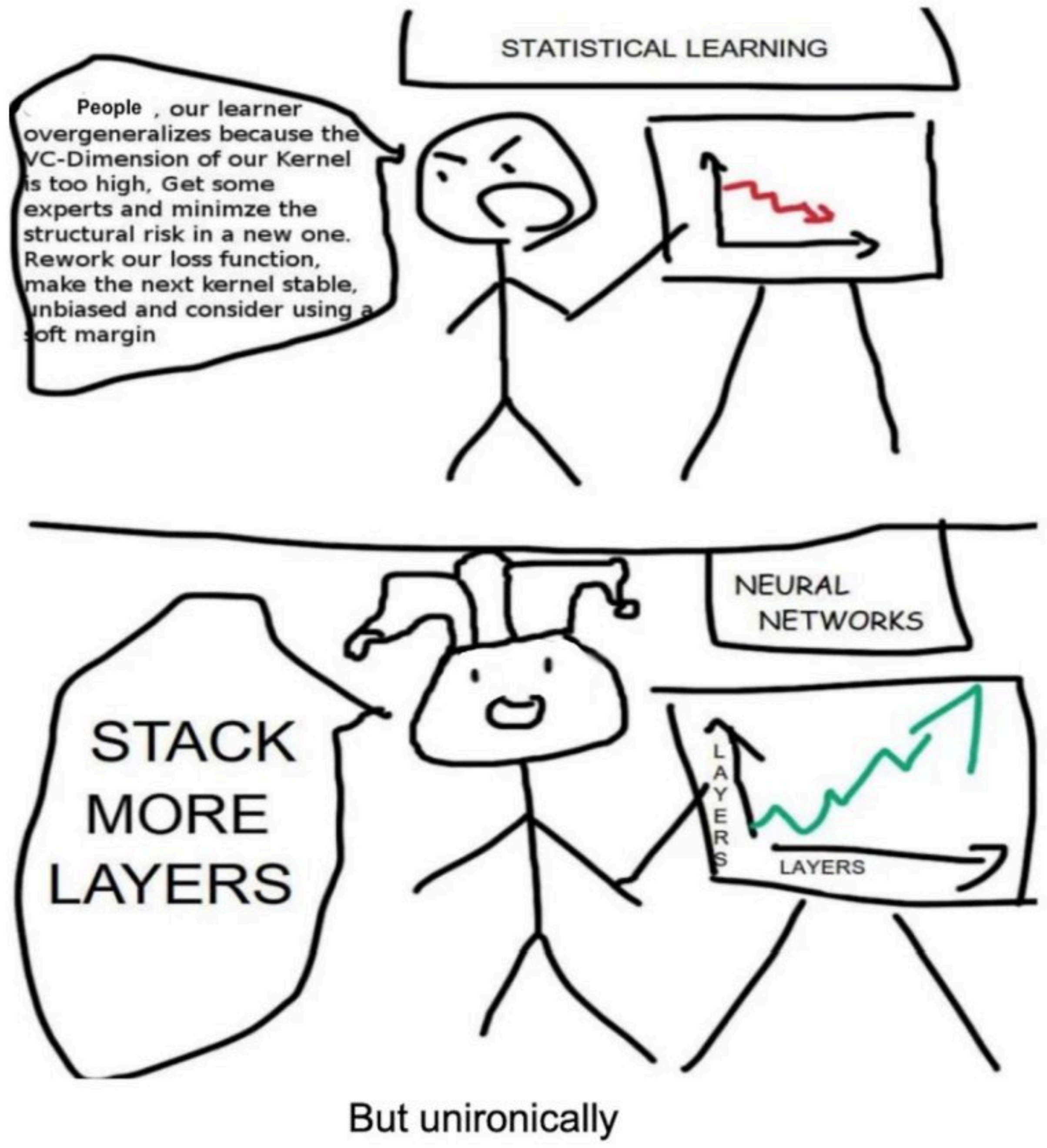
GPT-4: [Produces LATEX compiling to following picture.]



GPT-4

GPT-4 visual input example, Moar Layers:

User Can you explain why this is funny. Think about it step-by-step.



GPT-4 The comic is satirizing the difference in approaches to improving model performance between statistical learning and neural networks.

In statistical learning, the character is shown to be concerned with overfitting and suggests a series of complex and technical solutions, such as minimizing structural risk, reworking the loss function, and using a soft margin.

In contrast, the neural networks character simply suggests adding more layers to the model. This is often seen as a common solution to improving

GPT-4 visual input example, Extreme Ironing:

User What is unusual about this image?

Source: `https://www.barnorama.com/wp-content/uploads/2016/12/03-Confusing-Pictures.jpg`

GPT-4 The unusual thing about this image is that a man is ironing clothes on an ironing board attached to the roof of a moving taxi.

Limitations of LLMs

LLMs as a knowledge retriever

Who invented the telephone?

Alexander Graham Bell is credited with inventing the first practical telephone in 1876.

Who is the first person to go to Mariana Trench?

The first person to go to the Mariana Trench was the American oceanographer and adventurer Don Walsh, who descended to its deepest point, the Challenger Deep, in 1960.

Who is the president of the United States in 2023?

It is not possible to answer this question at this time since it is too far in the future.

LLMs are costly to update

DA Who is the president of the United States in 2023?

I'm sorry, but as a language model, I do not have access to real-time information, and my training data only goes up until 2021. However, I can tell you that the president of the United States in 2021 was Joe Biden, who was inaugurated on January 20, 2021. The next presidential election in the United States is scheduled for November 5, 2024.

How to update LLMs with up-to-date world knowledge?

LLMs easily hallucinate

- LLMs easily **hallucinate** and **generate factually-incorrect** text:

DA

Who is Danqi Chen?

Danqi Chen is a computer science researcher and an assistant professor at Princeton University. He is known for his work in the fields of natural language processing and machine learning. Chen's research focuses on developing machine learning models that can understand natural language text, including how to represent language meaning and how to answer questions posed in natural language. Some of his notable contributions include the development of the BiDAF model for machine reading comprehension and the creation of the CoQA dataset for conversational question answering. Chen has also received several awards and honors for his work, including the ACL Outstanding Paper Award and the EMNLP Best Paper Award.

LLMs easily hallucinate

- LLMs easily **hallucinate** and **generate factually-incorrect** text:

Prompt	Where do people smoke more on average than in Russia?
ChatGPT Generation	There are several countries where people smoke more on average than in Russia. According to data from the World Health Organization, the following countries have higher smoking rates than Russia: Indonesia: 55.2% of adults smoke China: 52.9% of adults smoke Serbia: 47.5% of adults smoke Serbia and Montenegro: 45.6% of adults smoke Montenegro: 45.6% of adults smoke Greece: 44.5% of adults smoke Bosnia and Herzegovina: 41.8% of adults smoke Kazakhstan: 41.3% of adults smoke Kosovo: 41.3% of adults smoke Moldova: 40.9% of adults smoke It's important to note that these figures are estimates and may vary slightly from other sources.
Hallucination	Extrinsic - the generated answer includes wrong statistics. (Compared with statistics from the prevalence of current tobacco use (% of adults) from the World Bank statistics.[16])

(Bang et al., 2023): A Multitask, Multilingual, Multimodal Evaluation of ChatGPT on Reasoning, Hallucination, and Interactivity

LLMs are easy to hallucinate

- **Solutions**: generate text with citations that can be fact-checked

Which river is longer, the Nile or the Yangtze?

The Nile is longer than the Yangtze River.[1][2][3] The Nile runs 6,650 kilometers (4,130 miles) between Lake Victoria and the Mediterranean Sea.[2] The Yangtze River runs 6,300 kilometers (3,915 miles) from its source in the Tanggula Mountains in Qinghai to the East China Sea.[3]

1. World's Longest Rivers | Amazon, Nile & Mississippi Rivers | Live Science (www.livescience.com)
2. Nile - Wikipedia (en.wikipedia.org)
3. Yangtze River Length - 6,300 Kilometers (3,915 Miles) (www.travelchinaguide.com)
4. How Long is the Yangtze River? Yangtze River's Length (www.chinatraveldiscovery.com)

How do neural networks work?

In its most basic form, a neural network has two layers: an input layer and an output layer[3]. The output layer is the component of the neural network that makes predictions[3]. In a feedforward network, information flows through the network in the following way: patterns of information are fed into the network via the input units, which trigger the layers of hidden units, and these in turn arrive at the output units[1]. The network learns by a feedback process called backpropagation, which involves comparing the output a network produces with the output it was meant to produce, and using the difference between them to modify the weights of the connections between the units in the network, working from the output units through the hidden units to the input units, going backward[2][4]. Over time, backpropagation causes the network to learn, reducing the difference between actual and intended output to the point where the two exactly coincide, so the network figures things out exactly as it should[2].

1. How neural networks work - A simple introduction (www.explainthatstuff.com)
2. How neural networks work - A simple introduction (www.explainthatstuff.com)
3. How Do Neural Networks Really Work? | Nick McCullum (nickmccullum.com)
4. How Do Neural Networks Really Work? | Nick McCullum (nickmccullum.com)

https://openai.com/research/webgpt

LLMs are costly to train and deploy

LLM.int8(): 8-bit Matrix Multiplication for Transformers at Scale

Tim Dettmers[λ*] **Mike Lewis**[†] **Younes Belkada**[§∓] **Luke Zettlemoyer**[†λ]

Quantization seems promising

Memory reduction by 2x

SparseGPT: Massive Language Models Can be Accurately Pruned in One-Shot

Elias Frantar[1] **Dan Alistarh**[1 2]

They can prune 175B models by 50% (unstructured pruning) without loss in perplexity