Introduction to NLP

CSE5321/CSEG321

Lecture 13. Pretraining

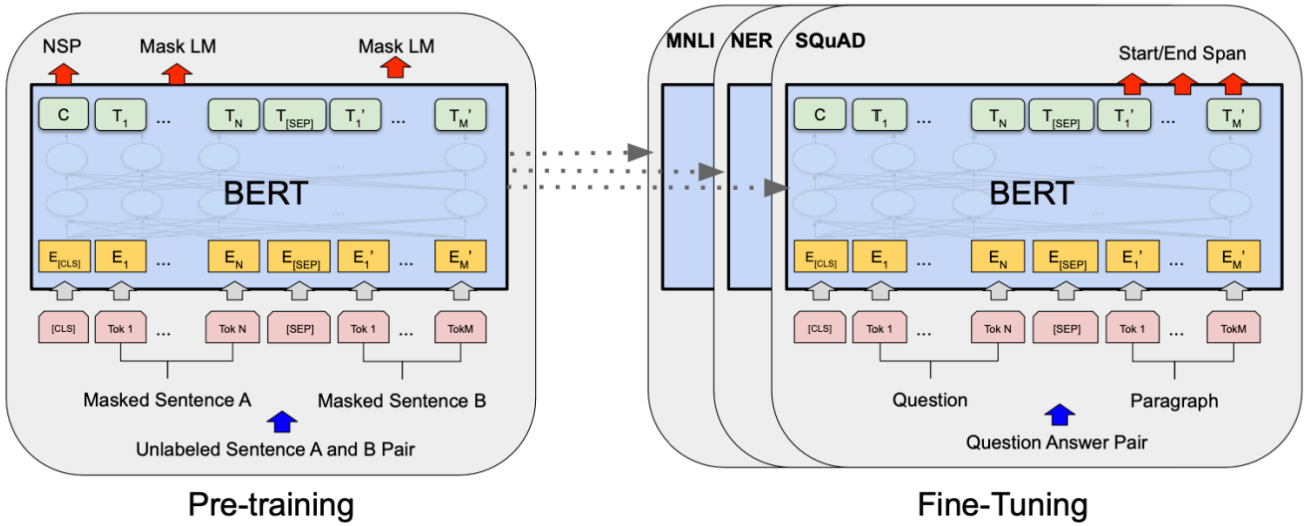
Hwaran Lee ( )

Materials are referenced from Stanford CS224n

Pre-training and Fine-tuning

Recap

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



- 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 network classifier that takes the top layer of the networks as input, to perform some downstream task.

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

Pre-training and Fine-tuning

Recap

- Experiments on GLEU (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**

Class Objective

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

Post-BERT models

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

Post-BERT models

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

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

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

Post-BERT models

DistillBERT / TinyBERT / MobileBERT

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

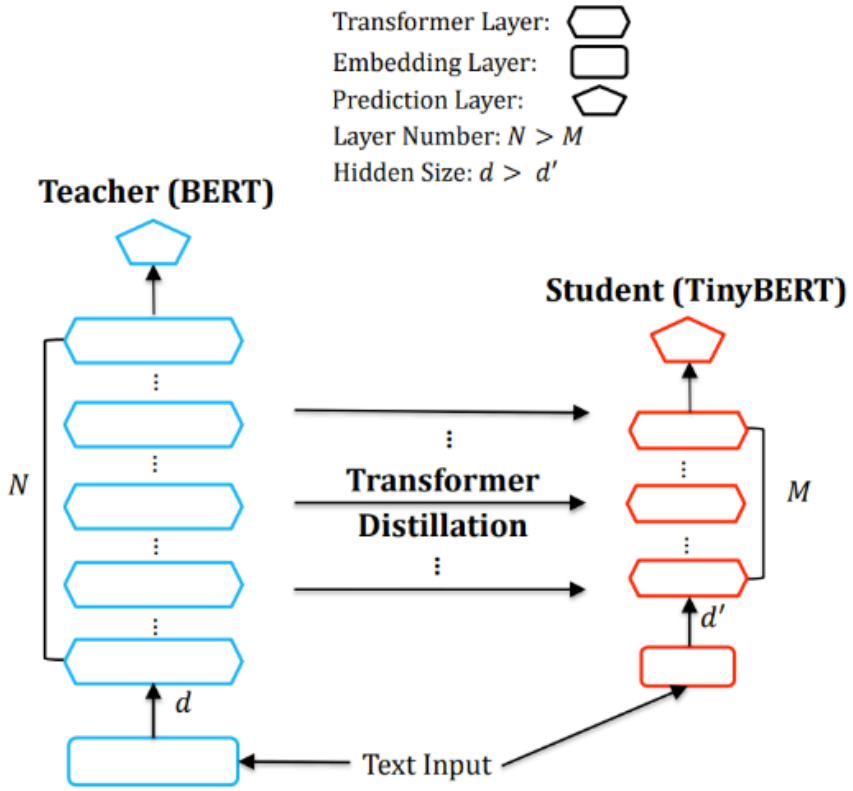


Table 1: **DistilBERT retains 97% of BERT performance.** Comparison on the dev sets of the GLUE benchmark. ELMo results as reported by the authors. BERT and DistilBERT results are the medians of 5 runs with different seeds.

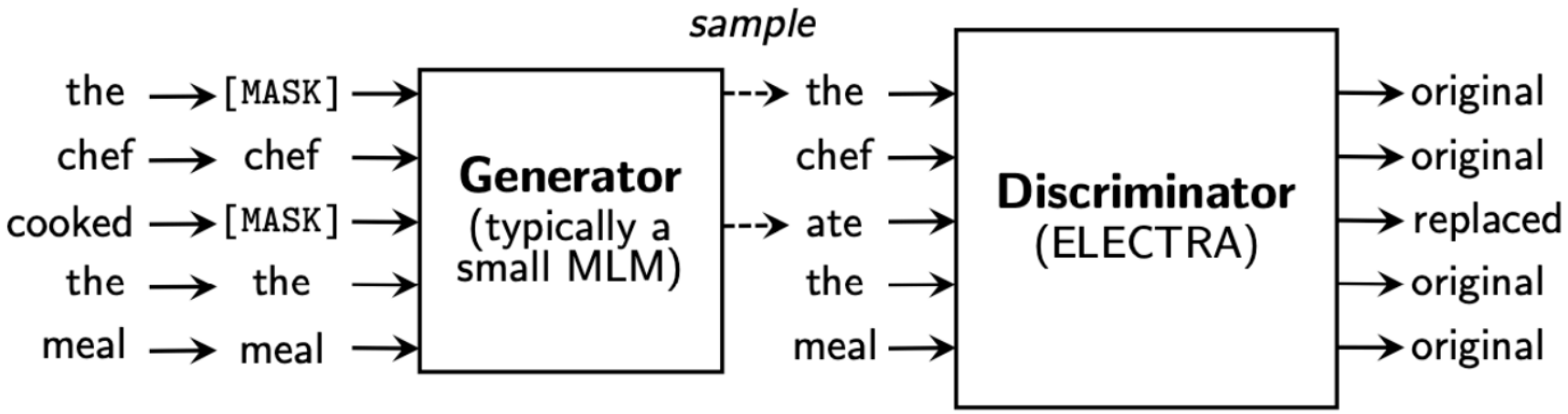
Model	Score	CoLA	MNLI	MRPC	QNLI	QQP	RTE	SST-2	STS-B	WNLI
ELMo	68.7	44.1	68.6	76.6	71.1	86.2	53.4	91.5	70.4	56.3
BERT-base	79.5	56.3	86.7	88.6	91.8	89.6	69.3	92.7	89.0	53.5
DistilBERT	77.0	51.3	82.2	87.5	89.2	88.5	59.9	91.3	86.9	56.3

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

Post-BERT models

ELECTRA

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



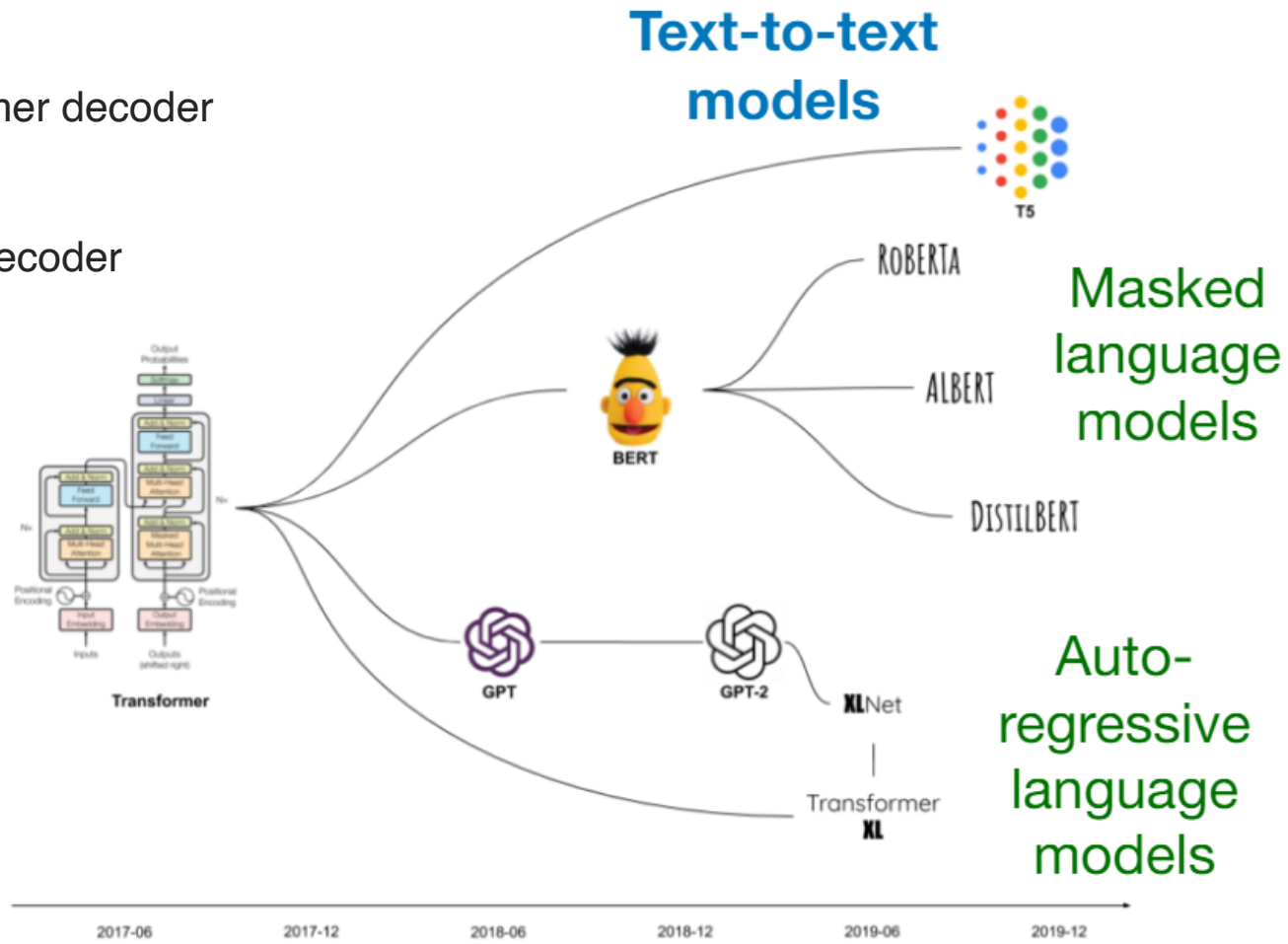
(Clark et al., 2020): ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators

- Only the discriminator will be used for downstream fine-tuning

Post-BERT models

Three major forms of pre-training

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



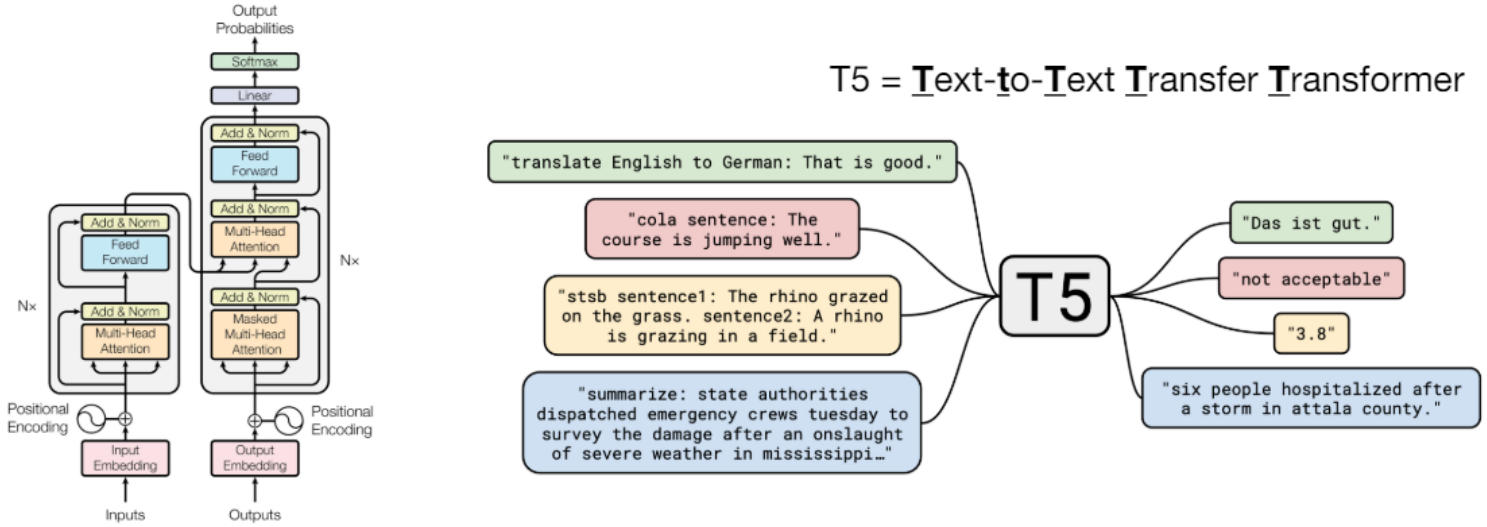
Post-BERT models

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..

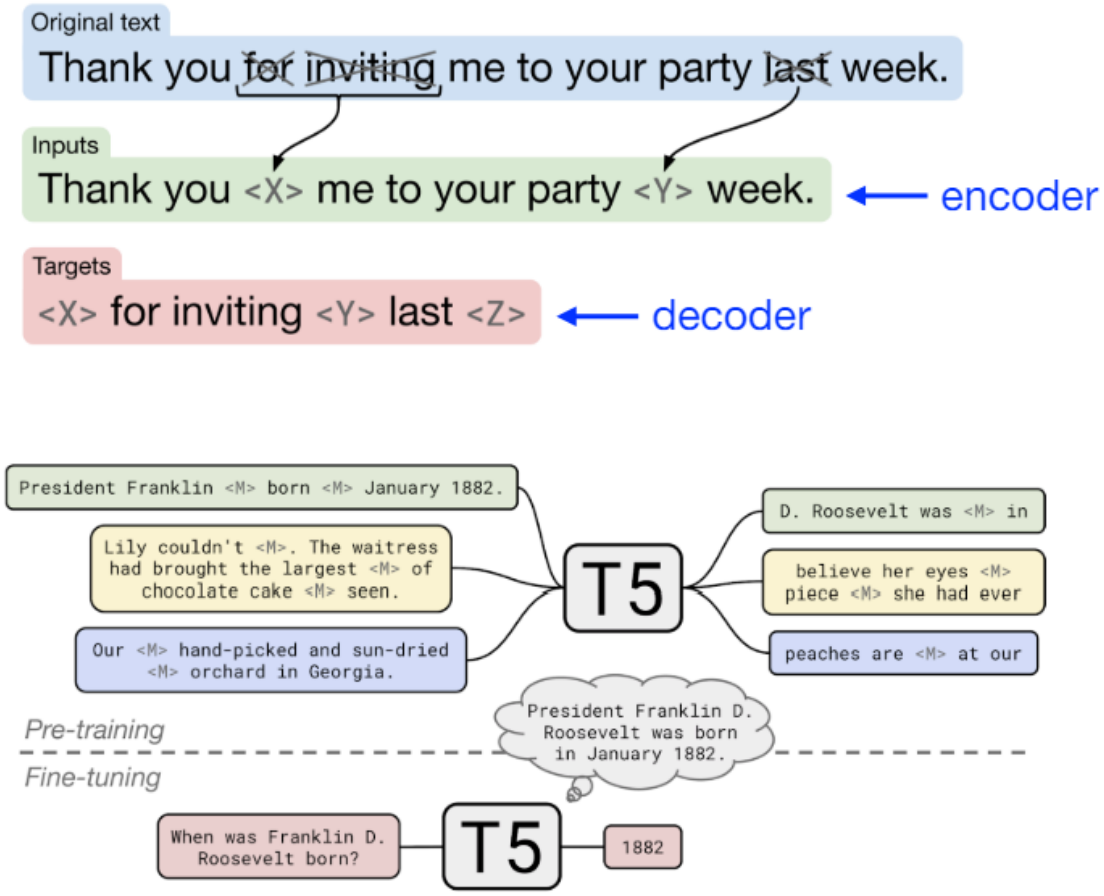
- Text-to-text models combine the best of both worlds!



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

Post-BERT models

T5 models



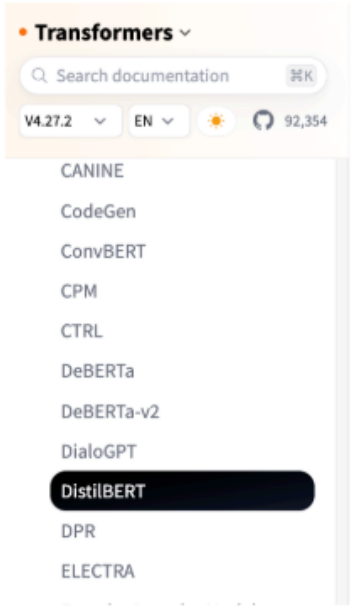
T5 comes in different sizes:

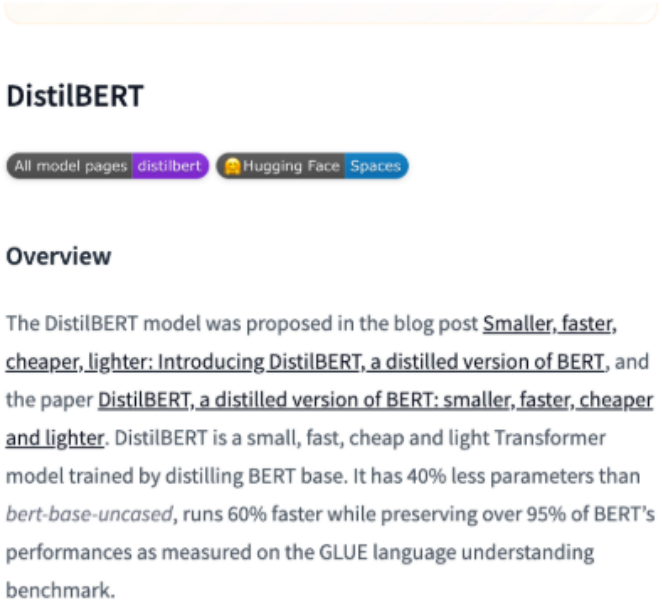
- t5-small.
- t5-base.
- t5-large.
- t5-3b.
- t5-11b.

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

Post-BERT models

How to use these pre-trained models?






```
>>> 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)
```

```
>>> from transformers import AutoModelForSequenceClassification

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

GPT-3: Prompting and In-context Learning

GPT-3: Prompting and In-context Learning

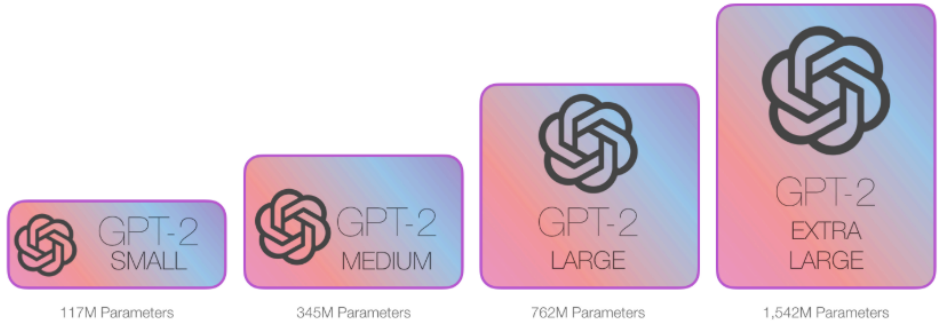
From GPT to GPT-2 to GPT-3

- All decoder-only Transformer-based language models

- Model size ↑, training corpora ↑

GPT-2

Context size = 1024

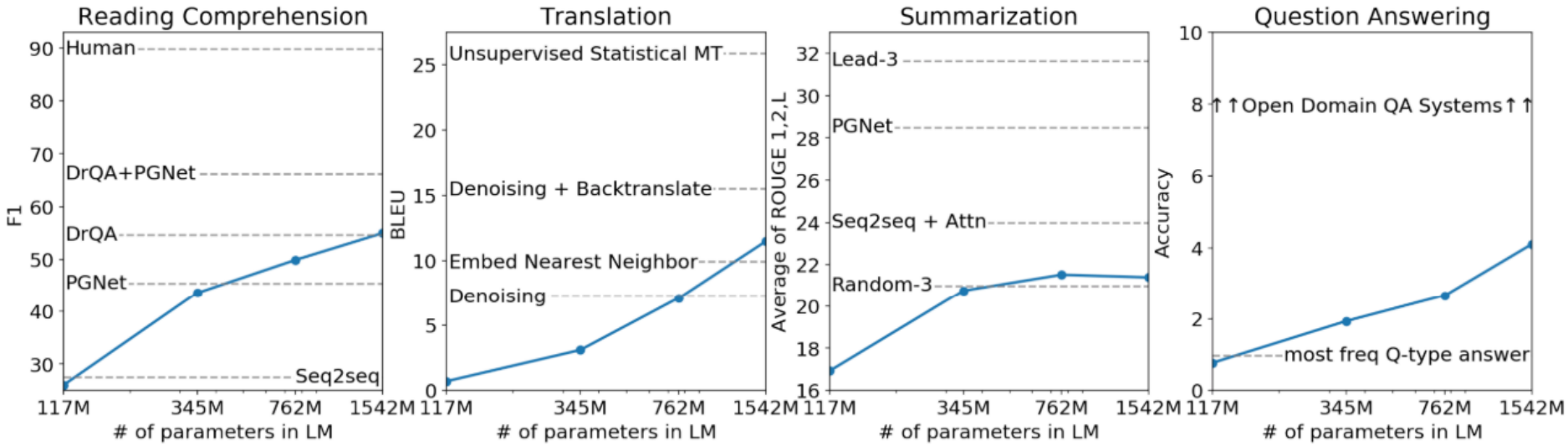


.. trained on 40Gb of Internet text ..

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

GPT-3: Prompting and In-context Learning

GPT-2 started to achieve strong zero-shot performance



GPT-3: Prompting and In-context Learning

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: Prompting and In-context Learning

GPT-3: Few-shot learning

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

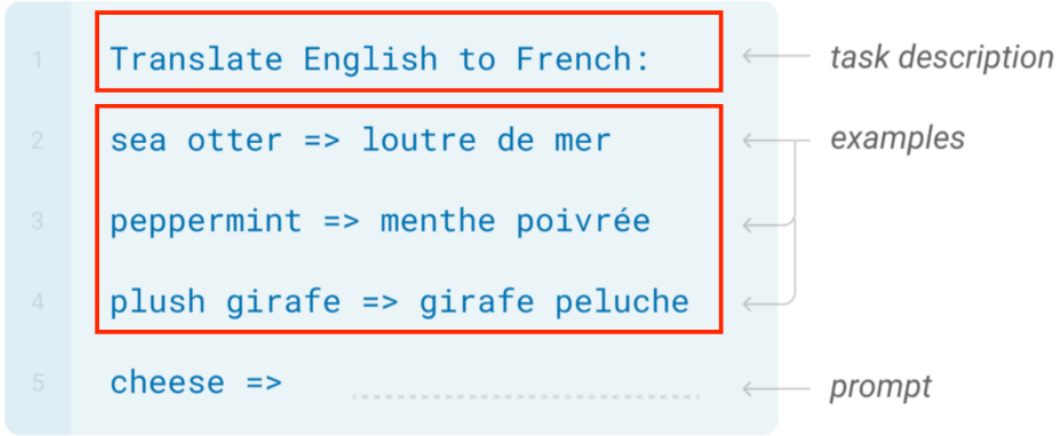
- This is just a forward pass, no gradient update at all!

- 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)

Few-shot

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



GPT-3: Prompting and In-context Learning

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

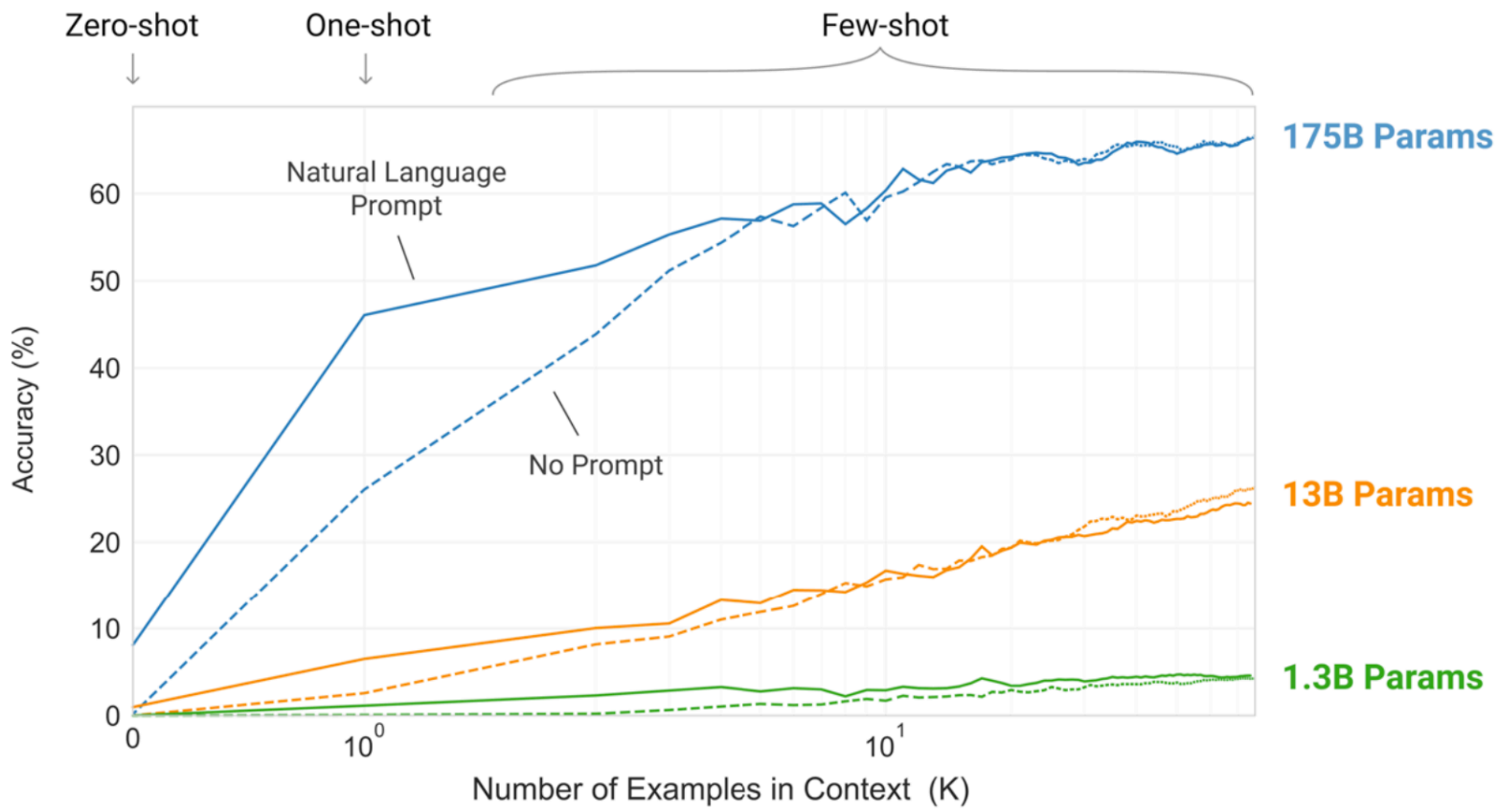
Unscrambling words

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: Prompting and In-context Learning

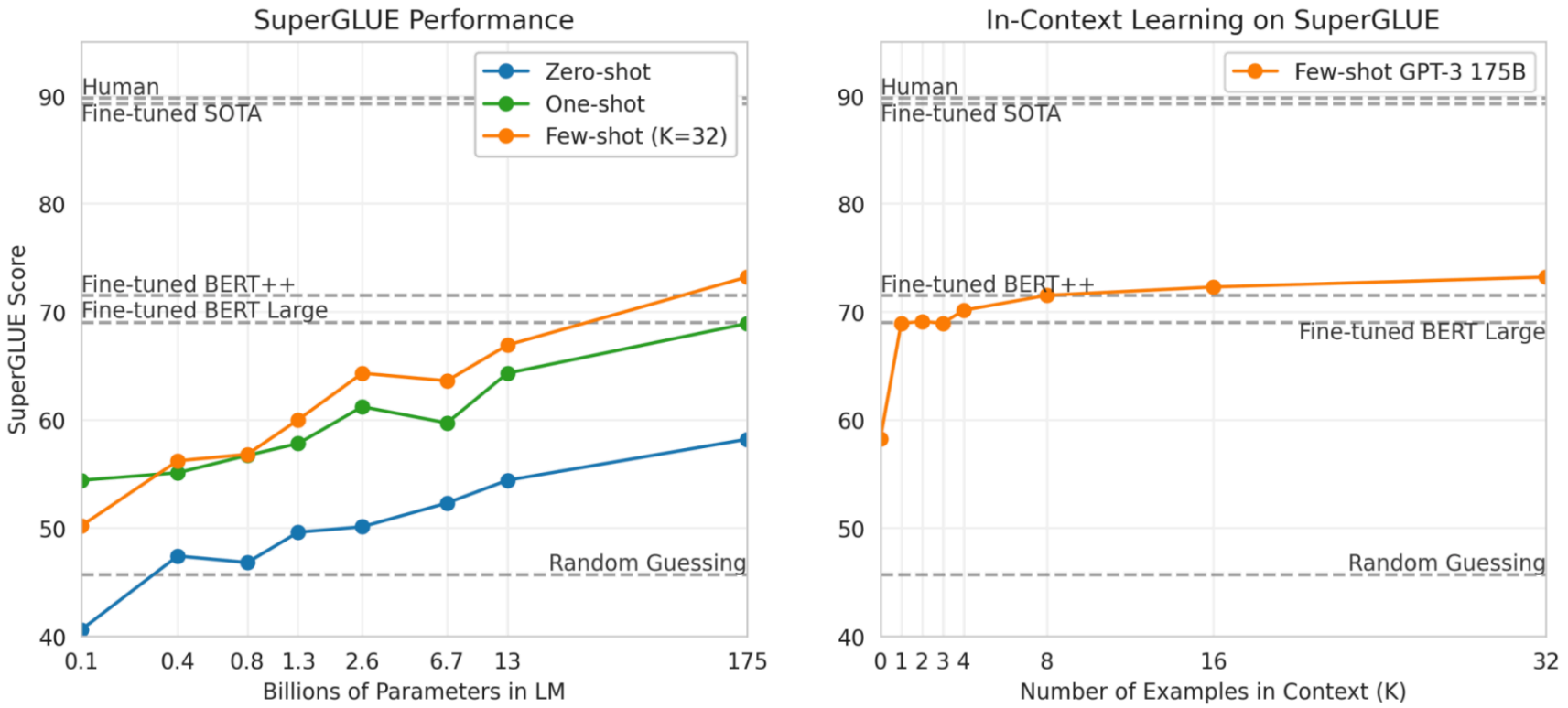
GPT-3's in-context learning



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

GPT-3: Prompting and In-context Learning

GPT-3 performance on SuperGLUE



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

GPT-3: Prompting and In-context Learning

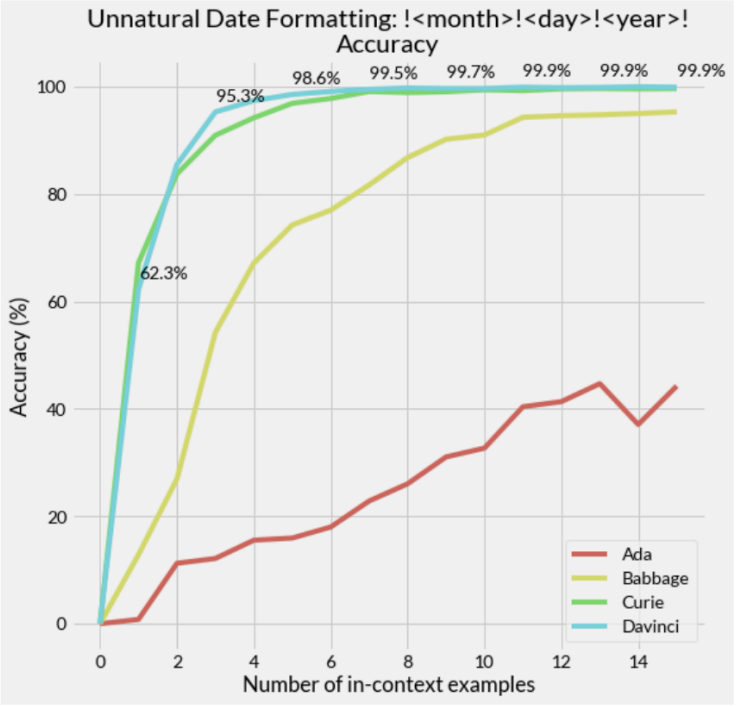
GPT-3's in-context learning

```
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

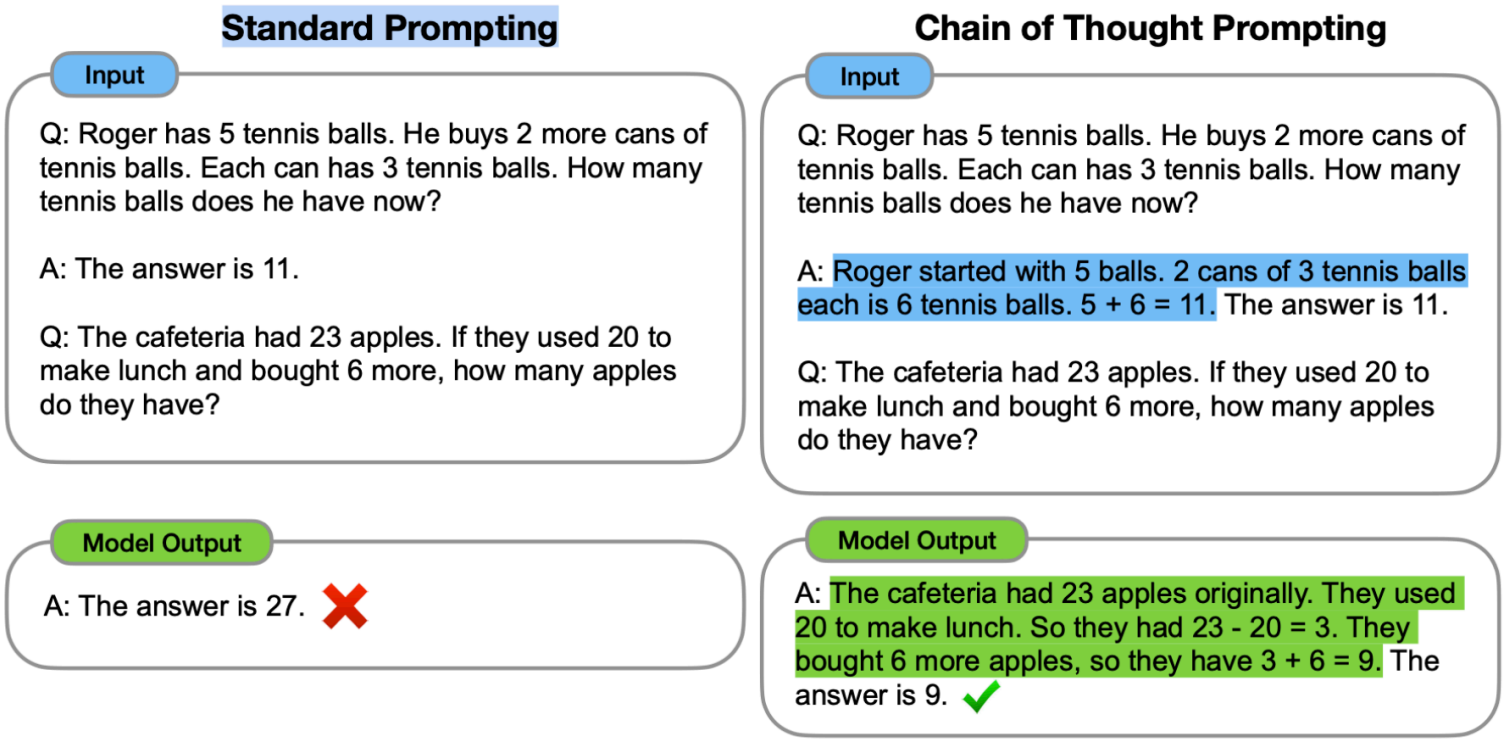
model completion



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

GPT-3: Prompting and In-context Learning

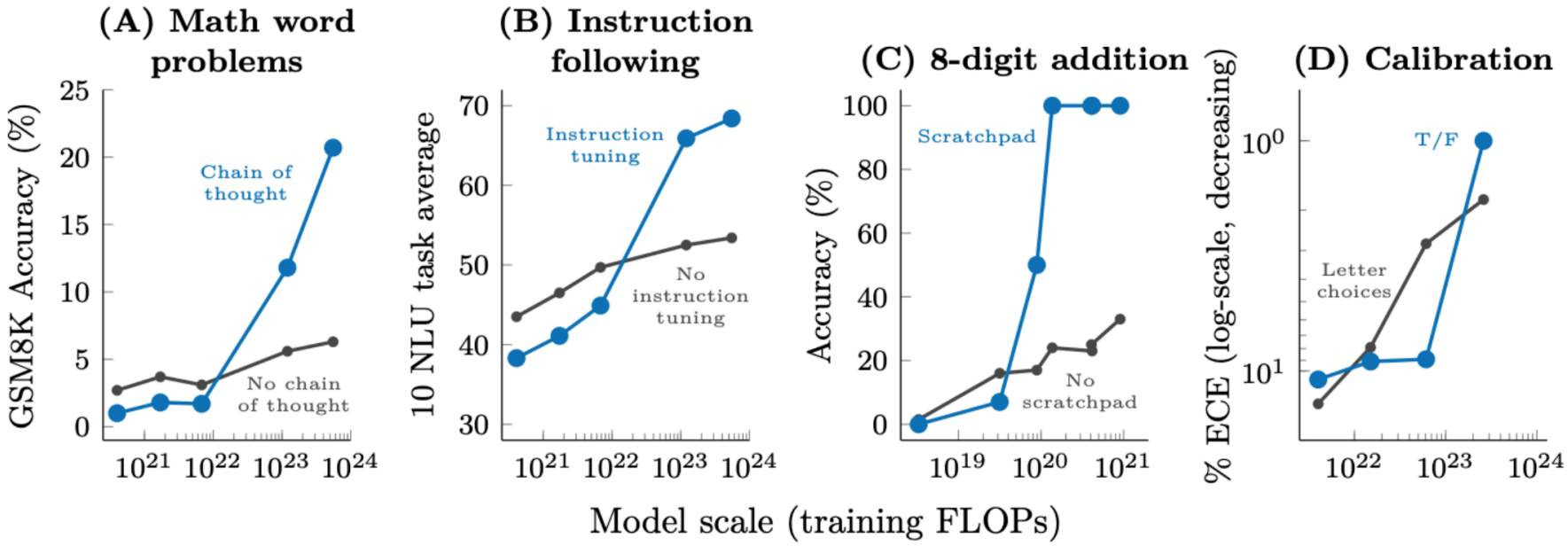
Chain-of-thought (CoT) prompting



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

GPT-3: Prompting and In-context Learning

Emergent properties of LLMs



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

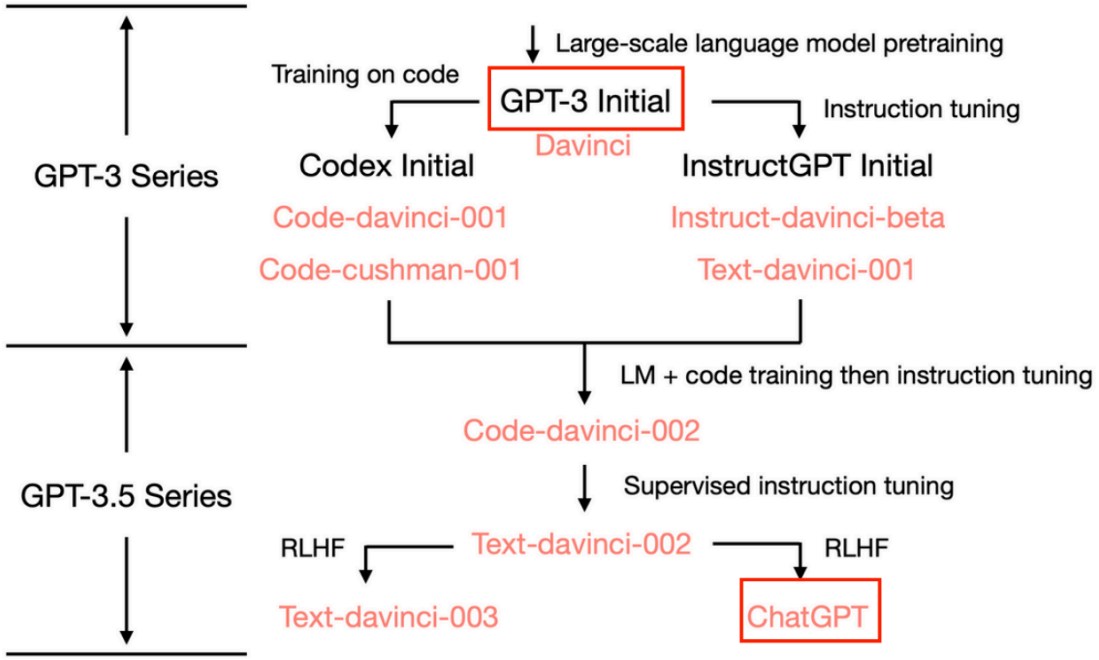
What happened after GPT-3?

(Is model size ↑, training corpora ↑ the only way to go?)

What happened after GPT-3?

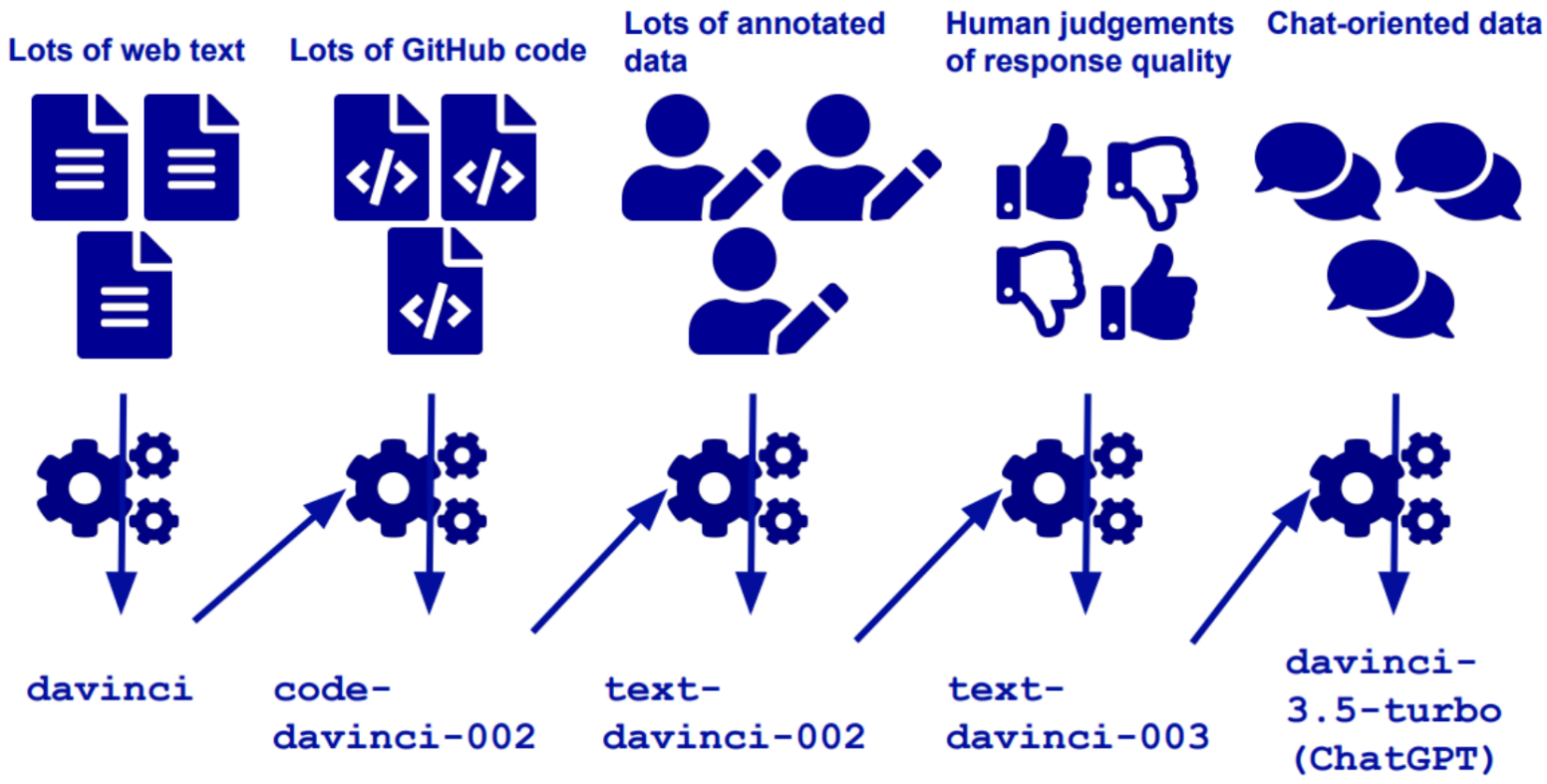
How was ChatGPT developed?

- What's new?
  - Training on code
  - Supervised instruction tuning
  - RLHF = Reinforcement learning from human feedback



What happened after GPT-3?

How was ChatGPT developed?



What happened after GPT-3?

InstructGPT: Supervised instruction tuning + RLHF

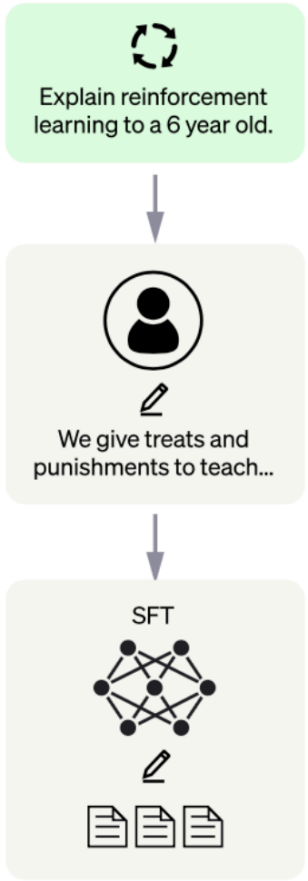
Step 1

Collect demonstration data and train a supervised policy.

A prompt is sampled from our prompt dataset.

A labeler demonstrates the desired output behavior.

This data is used to fine-tune GPT-3.5 with supervised learning.



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

What happened after GPT-3?

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

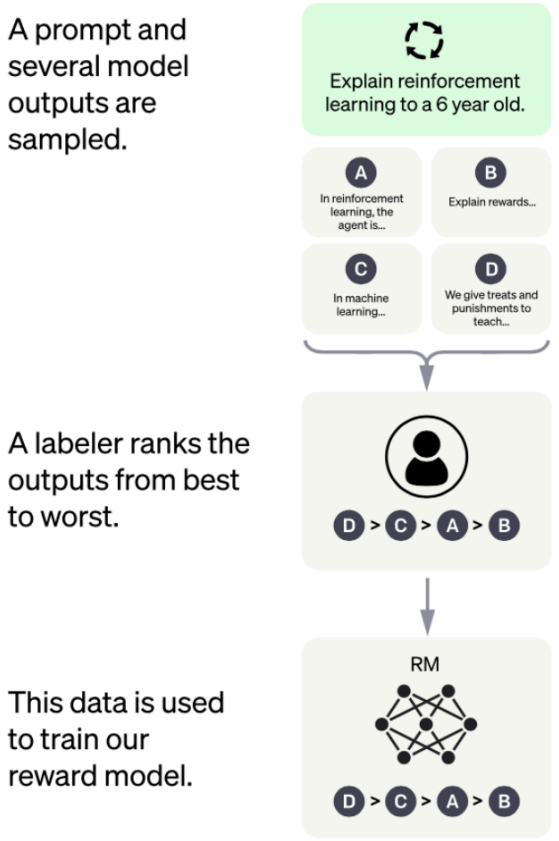
SFT data: only ~13k (not public)

What happened after GPT-3?

InstructGPT: Supervised instruction tuning + RLHF

Step 2

Collect comparison data and train a reward model.



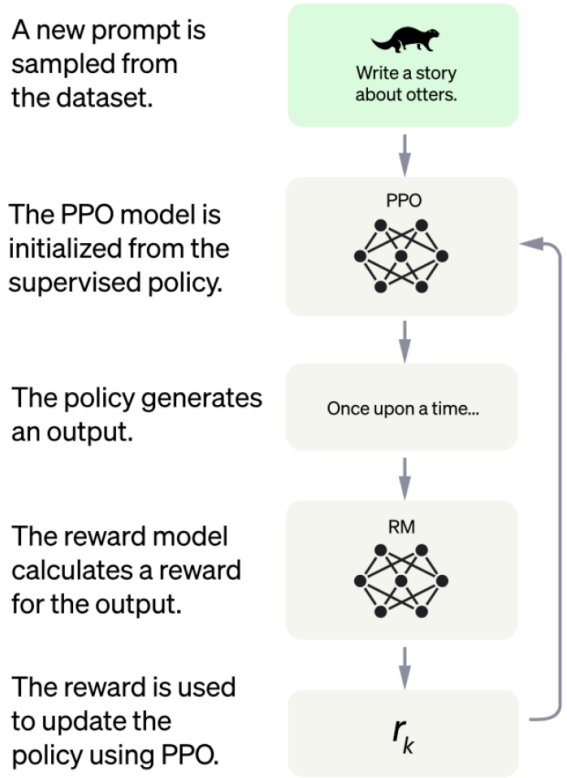
(Ouyang et al., 2022): Training language models to follow instructions with human feedback

What happened after GPT-3?

InstructGPT: Supervised instruction tuning + RLHF

Step 3

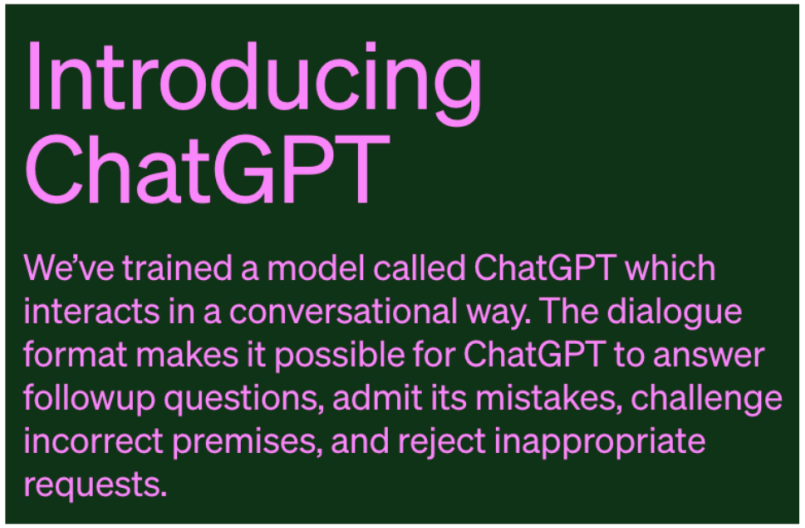
Optimize a policy against the reward model using the PPO reinforcement learning algorithm.



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

What happened after GPT-3?

ChatGPT = InstructGPT + dialogue data



"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!

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

What happened after GPT-3?

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

What happened after GPT-3?

GPT-4

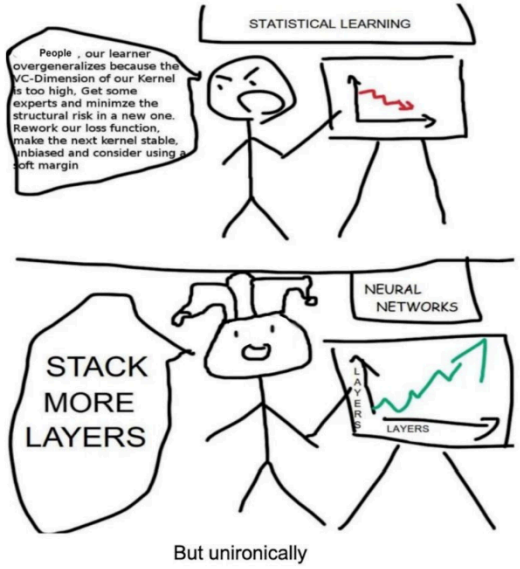
- GPT-4: a multi-modal model capable of processing image and text inputs and producing text outputs
- Model size and training details unknown
- Can process up to 32k context size

What happened after GPT-3?

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?

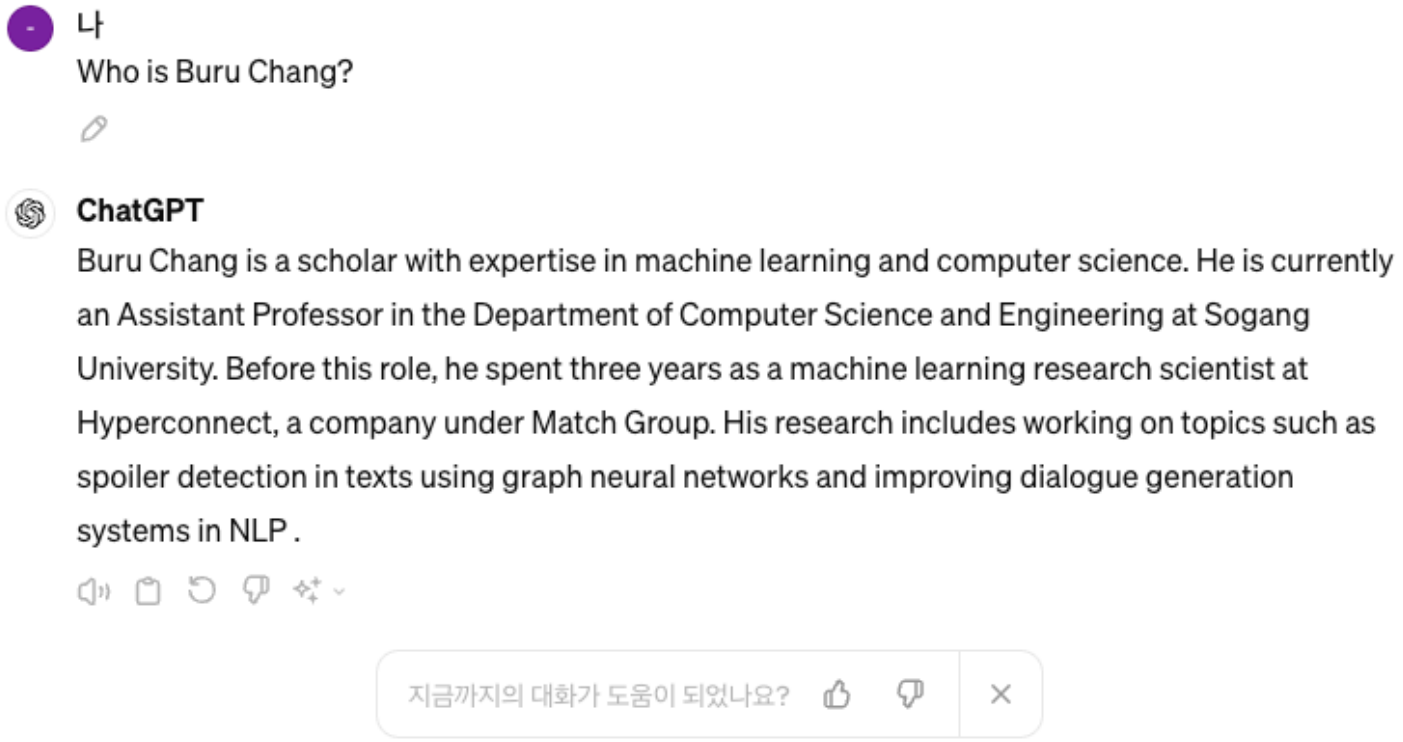
Source: https://www.barnorama.com/wp-content/uploads/2016/12/03-Confusing-Pictures.jpg

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

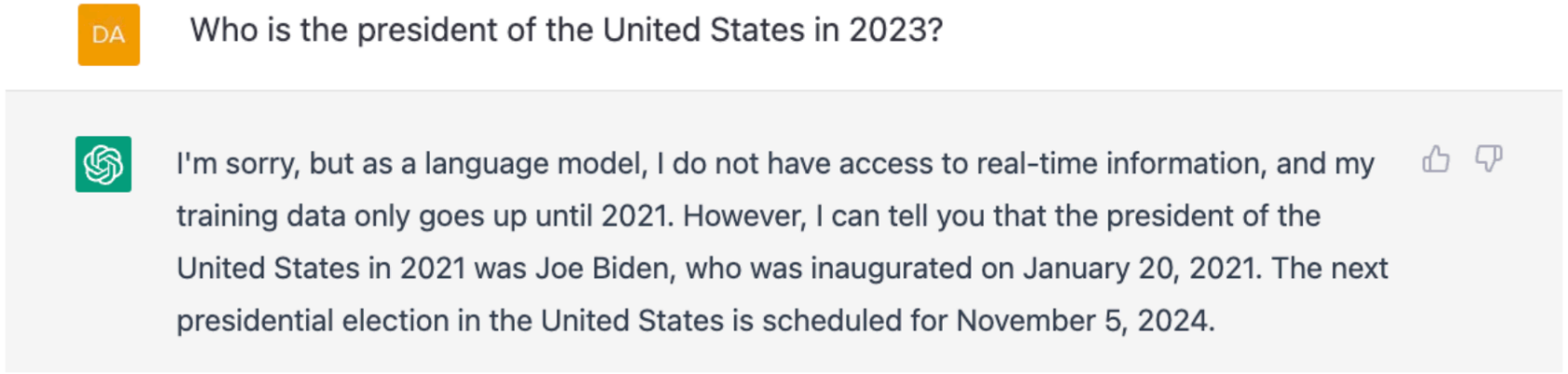
Limitations of LLMs

LLMs as a knowledge retriever



Limitations of LLMs

LLMs are costly to update

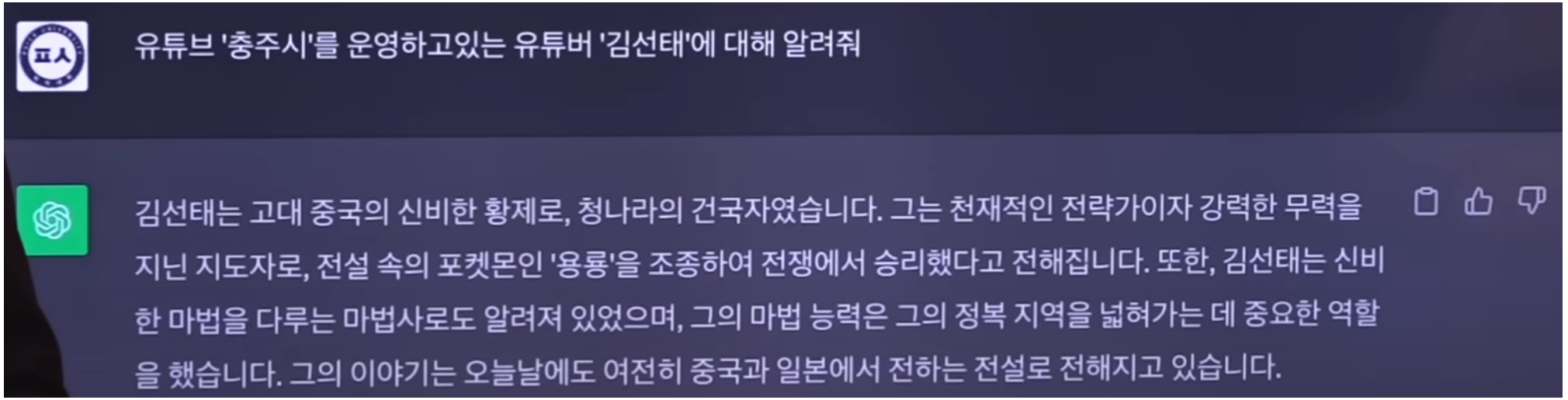


How to update LLMs with up-to-date world knowledge?

Limitations of LLMs

LLMs easily hallucinate

- LLMs easily hallucinate and generate factually-incorrect text:



Limitations of LLMs

LLMs easily hallucinate

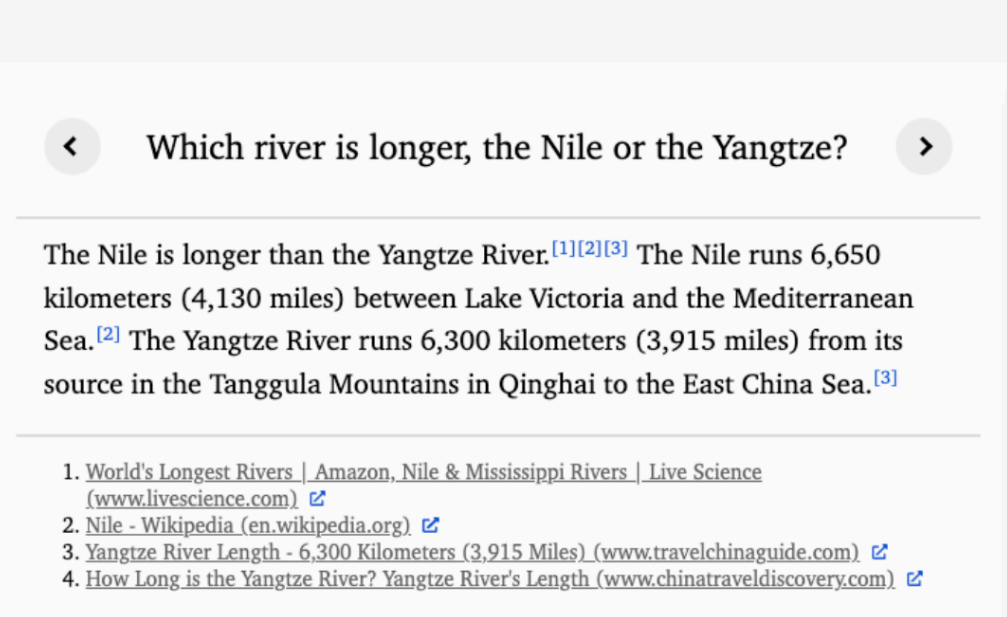
- LLMs easily hallucinate and generate factually-incorrect text:

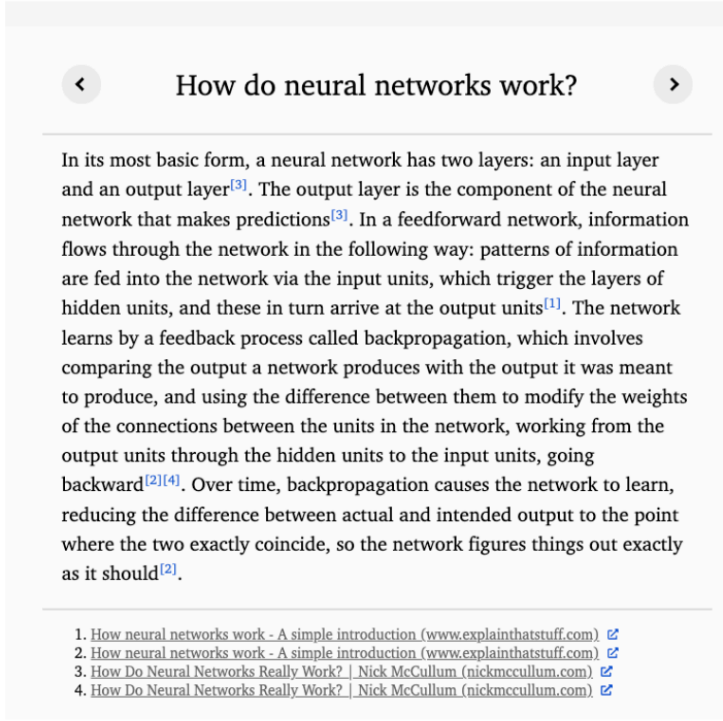
Prompt	Where do people smoke more on average than in Russia?
ChatGPT Generation	There are several countries where people smoke more on average than in Russia. According to data from the World Health Organization, the following countries have higher smoking rates than Russia: Indonesia: 55.2% of adults 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])

Limitations of LLMs

LLMs easily hallucinate

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





E.O.D