Large Language Models

CSCI 601 471/671
NLP: Self-Supervised Models

https://self-supervised.cs.jhu.edu/sp2023/

[Slide credit: Chris Tanner, Jacob Devlin and many others]

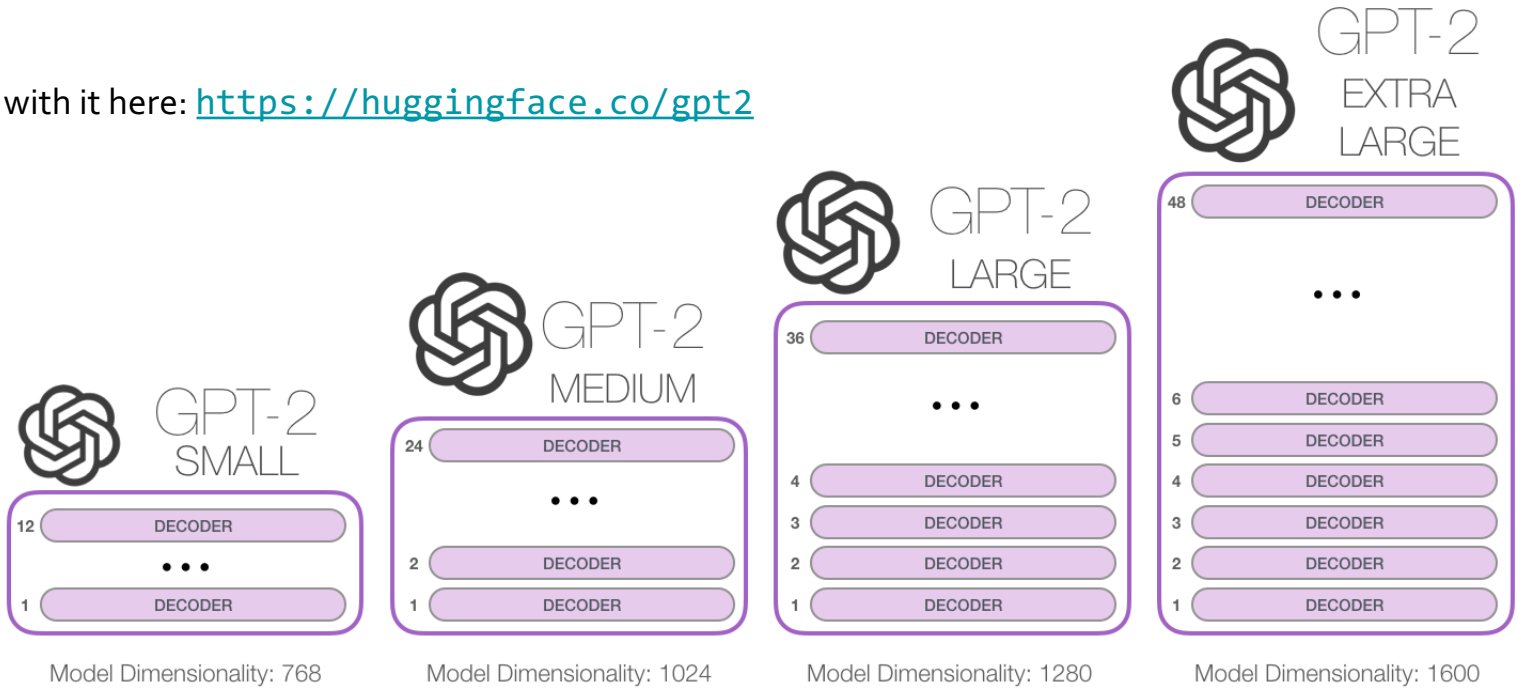
Logistics Update

- The midterm:
  - How was it?

- HW6 is out!
  - No Google-colabs anymore!
  - You can do it as a team (one submission per team).
  - **Please** start the programming portion early!

GPT2: Model Sizes

Play with it here: https://huggingface.co/gpt2



117M parameters 345M 762M 1542M

[Image by http://jalammar.github.io/illustrated-gpt2/]

GPT-3: A Very Large Language Model (2020)

- More layers & parameters
- Bigger dataset
- Longer training
- Larger embedding/hidden dimension
- Larger context window

[Slide credit: Sbhya Chhabria & Michael Tang]

GPT3: Try it yourself!

https://beta.openai.com/playground

```python
import openai
openai.api_key = ("sk-3kFAtzisBypel                                    ")
my_prompt = '''The sun is [MASK].

    Replace [MASK] with the most probable 5 words to replace, and give me their probabilities.'''
# Here set parameters as you like
response = openai.Completion.create(
  engine="text-davinci-002",
  prompt=my_prompt,
  temperature=0,
  max_tokens=100,
)

print(response['choices'][0]['text'])
```

GPT3: Try it yourself!

Ada	Babbage	Curie	Davinci
Fastest			Most powerful
$0.0004 / 1K tokens	$0.005 / 1K tokens	$0.0020 / 1K tokens	$0.0200 / 1K tokens

Fine-tuned models

Create your own custom models by fine-tuning our base models with your training data. Once you fine-tune a model, you'll be billed only for the tokens you use in requests to that model.

Learn more about fine-tuning ↗

Model	Training	Usage
Ada	$0.0004 / 1K tokens	$0.0016 / 1K tokens
Babbage	$0.0006 / 1K tokens	$0.0024 / 1K tokens
Curie	$0.0030 / 1K tokens	$0.0120 / 1K tokens
Davinci	$0.0300 / 1K tokens	$0.1200 / 1K tokens

Other Available [Decoder] LMs

EleutherAI: GPT-Neo (6.7B), GPT-J (6B), GPT-NeoX (20B)

https://huggingface.co/EleutherAI
https://6b.eleuther.ai/

Meta/Facebook: OPT (Open Pre-trained Transformer), various sizes up to 175B

https://huggingface.co/facebook/opt-125m

BLOOM, 176B model https://huggingface.co/bigscience/bloom

LLaMA, 65B https://github.com/facebookresearch/llama

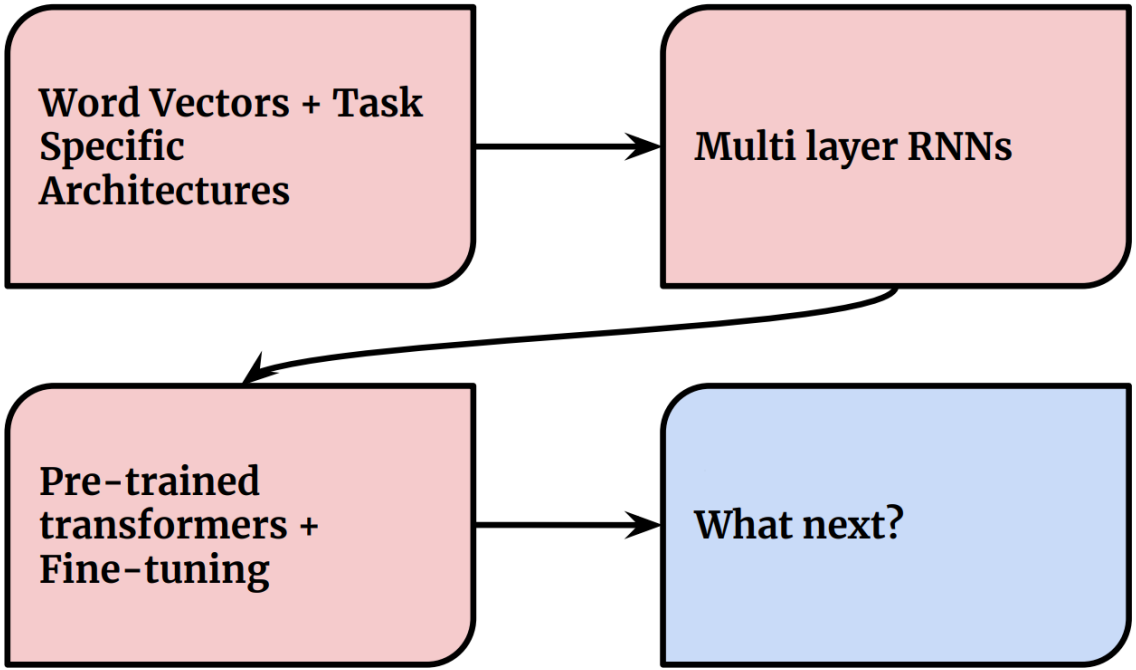
In-context Learning

CSCI 601 471/671
NLP: Self-Supervised Models

https://self-supervised.cs.jhu.edu/sp2023/

[Slide credit: Iz Beltagy, Arman Cohan, Robert Logan IV, Sewon Min, Sameer Singh, and many others]

The Phases of Paradigms



“I have an extremely large collection of clean labeled data”

- No one

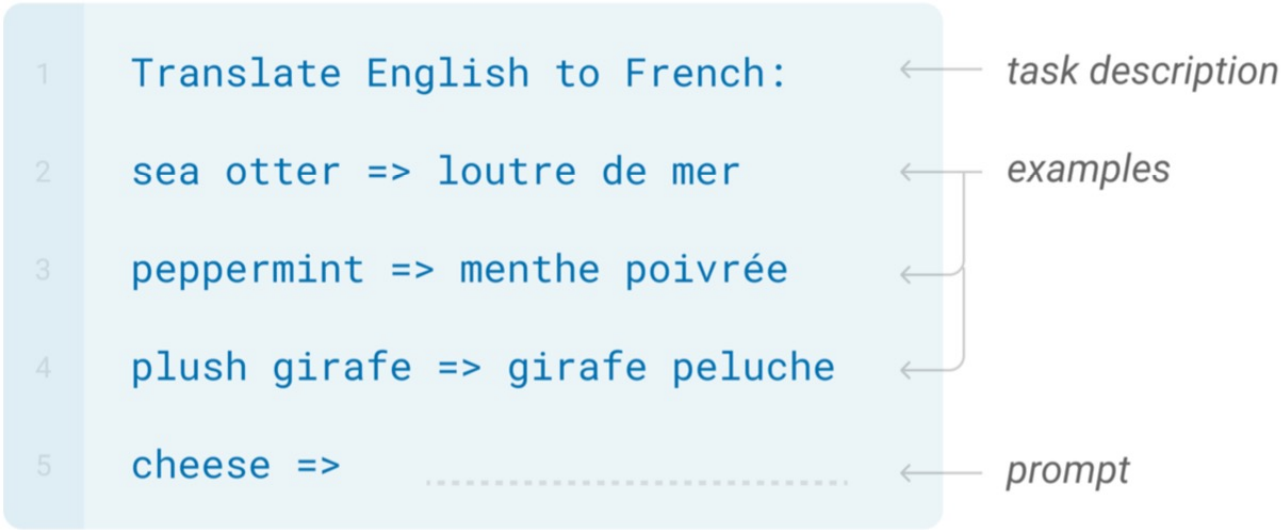
[Colin Raffel]

Limitations of Pre-training -> Fine-tuning

- Often you need a large labeled data
  - Though more pre-training can reduce the need for labeled data

- End up with many copies (or sub-copies) of the same model

In-context Learning



In-context Learning

Reverse words in a sentence

This is great
Great is this

The man on the moon
Moon the on man the

Will this really work
Work really this will

I hope this is a big achievement
Achievement big I hope this is

The king came home on a horse
Home horse king came the

https://gpt3experiments.substack.com/p/is-gpt-3-really-doing-few-shot-learning

In-context Learning

Context (passage and previous question/answer pairs)

Tom goes everywhere with Catherine Green, a 54-year-old secretary. He moves around her office at work and goes shopping with her. "Most people don't seem to mind Tom," says Catherine, who thinks he is wonderful. "He's my fourth child," she says. She may think of him and treat him that way as her son. He moves around buying his food, paying his health bills and his taxes, but in fact Tom is a dog.

Catherine and Tom live in Sweden, a country where everyone is expected to lead an orderly life according to rules laid down by the government, which also provides a high level of care for its people. This level of care costs money.

People in Sweden pay taxes on everything, so aren't surprised to find that owning a dog means more taxes. Some people are paying as much as 500 Swedish kronor in taxes a year for the right to keep their dog, which is spent by the government on dog hospitals and sometimes medical treatment for a dog that falls ill. However, most such treatment is expensive, so owners often decide to offer health and even life _ for their dog.

In Sweden dog owners must pay for any damage their dog does. A Swedish Kennel Club official explains what this means: if your dog runs out on the road and gets hit by a passing car, you, as the owner, have to pay for any damage done to the car, even if your dog has been killed in the accident.

Q: How old is Catherine?
A: 54

Q: where does she live?
A:

Model answer: Stockholm
Turker answers: Sweden, Sweden, in Sweden, Sweden

GPT3: Try it yourself!

https://beta.openai.com/playground

In-Context (Few Shot) Prompting

- Popularized by GPT-3 (but predates that model)
- Perform a task based on a few examples provided in the inference time.
- The model identifies patterns in examples and replicates it

GPT-3: Language Models are Few-Shot Learners

Zero-shot

The model predicts the answer given only a natural language description of the task. No gradient updates are performed.

```
1  Translate English to French:        ← task description
2  cheese =>                           ← prompt
```

One-shot

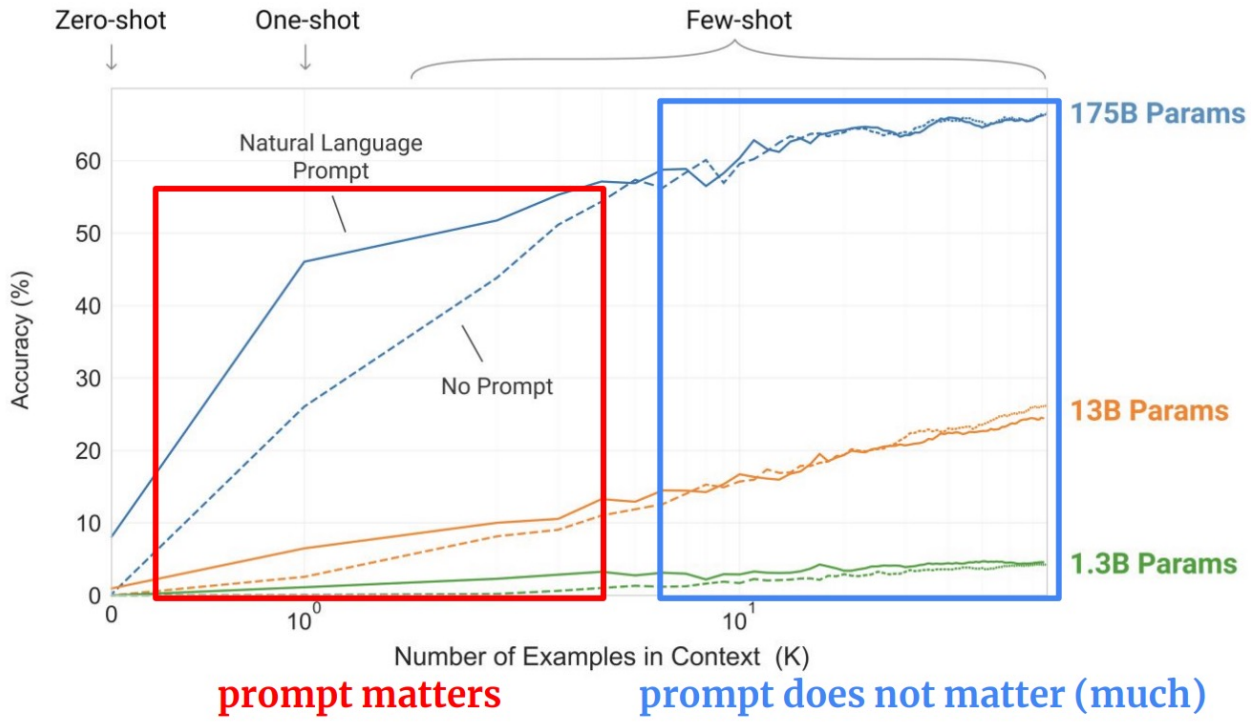
In addition to the task description, the model sees a single example of the task. No gradient updates are performed.

```
1  Translate English to French:        ← task description
2  sea otter => loutre de mer          ← example
3  cheese =>                           ← prompt
```

Few-shot

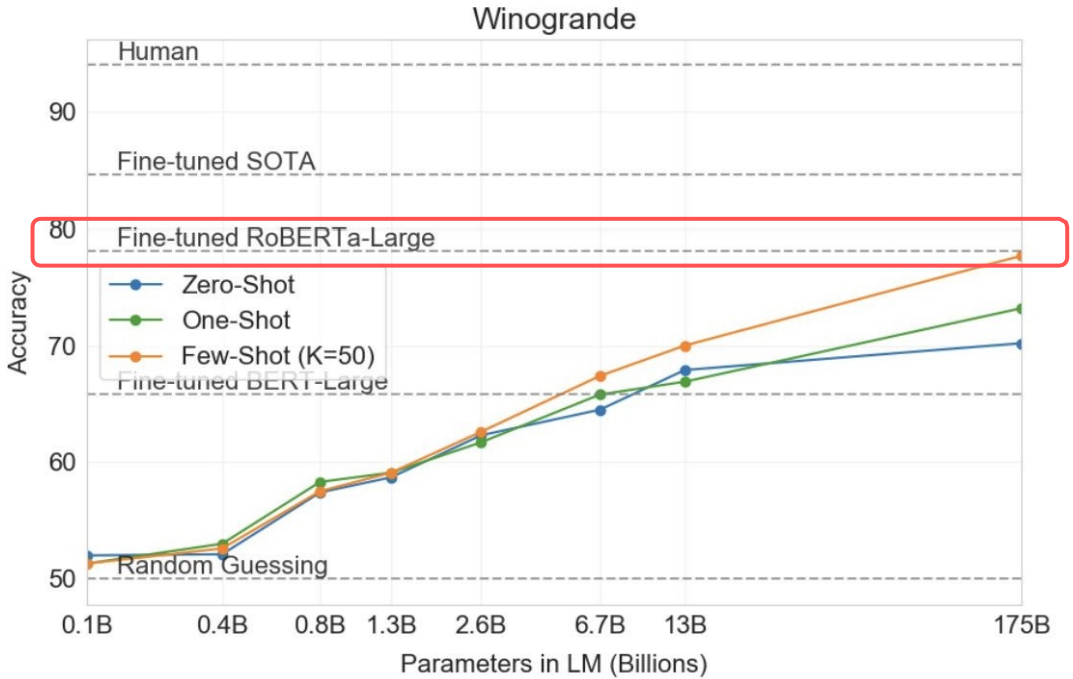
In addition to the task description, the model sees a few examples of the task. No gradient updates are performed.

```
1  Translate English to French:        ← task description
2  sea otter => loutre de mer          ← examples
3  peppermint => menthe poivrée        ←
4  plush girafe => girafe peluche      ←
5  cheese =>                           ← prompt
```



In-context learning results

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



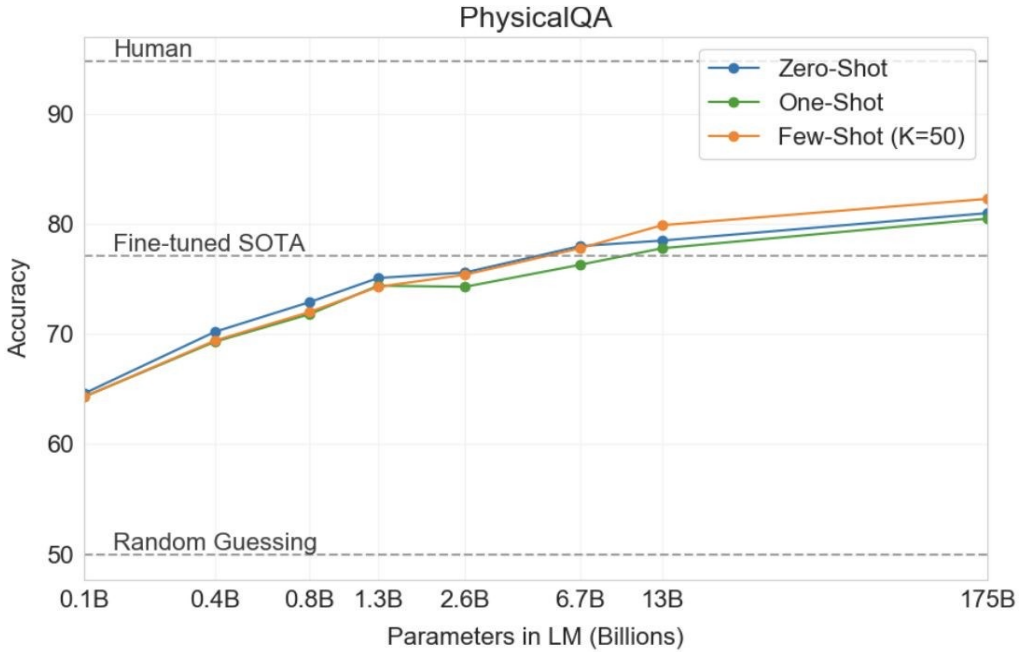
✓	Robert woke up at 9:00am while Samuel woke up at 6:00am, so **he** had *less* time to get ready for school.	**Robert** / Samuel
	Robert woke up at 9:00am while Samuel woke up at 6:00am, so **he** had *more* time to get ready for school.	Robert / **Samuel**

In-context learning results

To separate egg whites from the yolk using a water bottle, you should…

a. ***Squeeze*** the water bottle and press it against the yolk. ***Release,*** which creates suction and lifts the yolk.

b. ***Place*** the water bottle and press it against the yolk. ***Keep pushing,*** which creates suction and lifts the yolk.

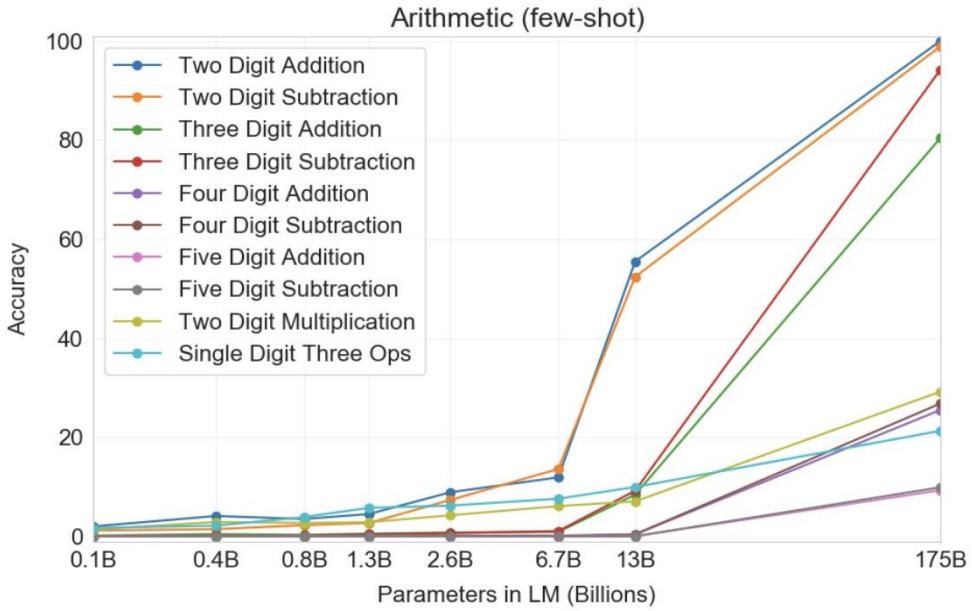


Brown et al. 2020. "Language Models are Few-Shot Learners"

In-context learning results

- Example:
  - Q: What is 48 plus 76?
  - A: 124

- Observations:
  - Scale is important
  - Number of digits correlate with their difficulty.
  - Multiplication is harder than summation!



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

The Phases of Our Understanding

"Language modeling is a useful subtask for many NLP tasks"
– everyone, pre-2018

"Language modeling is a useful supertask for many NLP tasks"
– everyone, post-2018