Large Language Models

CSC413 Tutorial 9

Yongchao Zhou

Overview

- What are LLMs?
- Why LLMs?
- Emergent Capabilities
  - Few-shot In-context Learning
  - Advanced Prompt Techniques
- LLM Training
  - Architectures
  - Objectives
- LLM Finetuning
  - Instruction finetuning
  - RLHF
  - Bootstrapping
- LLM Risks

What are Language Models?

- Narrow Sense
  - A probabilistic model that assigns a probability to every finite sequence (grammatical or not)

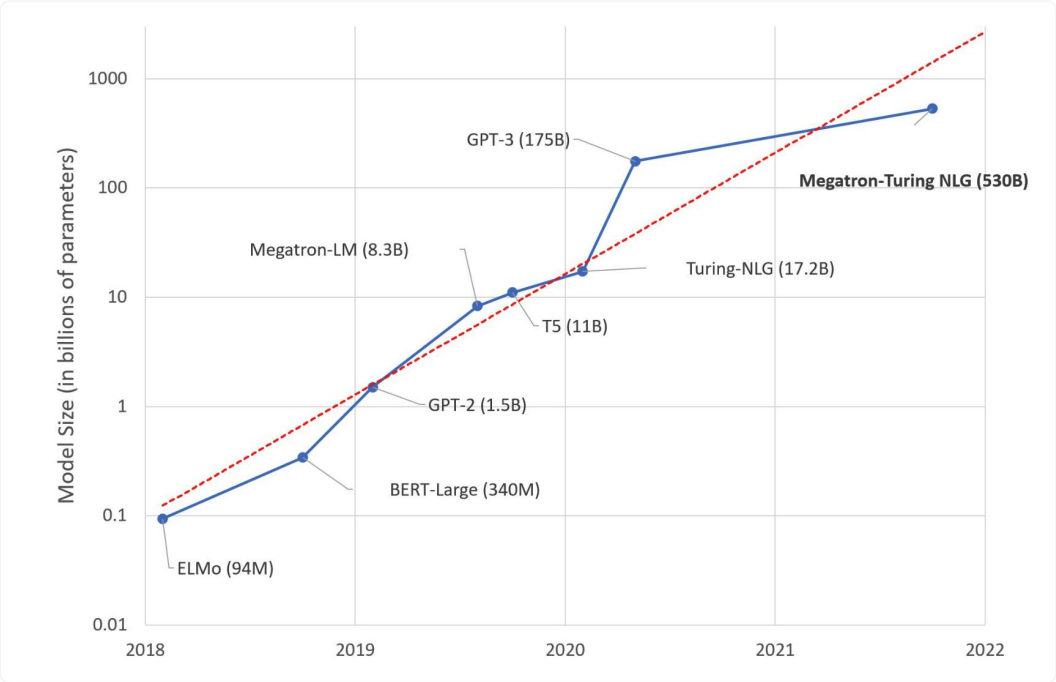
Sentence: “the cat sat on the mat”

$$P(\text{the cat sat on the mat}) = P(\text{the}) * P(\text{cat}|\text{the}) * P(\text{sat}|\text{the cat}) \\ *P(\text{on}|\text{the cat sat}) * P(\text{the}|\text{the cat sat on}) \\ *P(\text{mat}|\text{the cat sat on the})$$

Implicit order

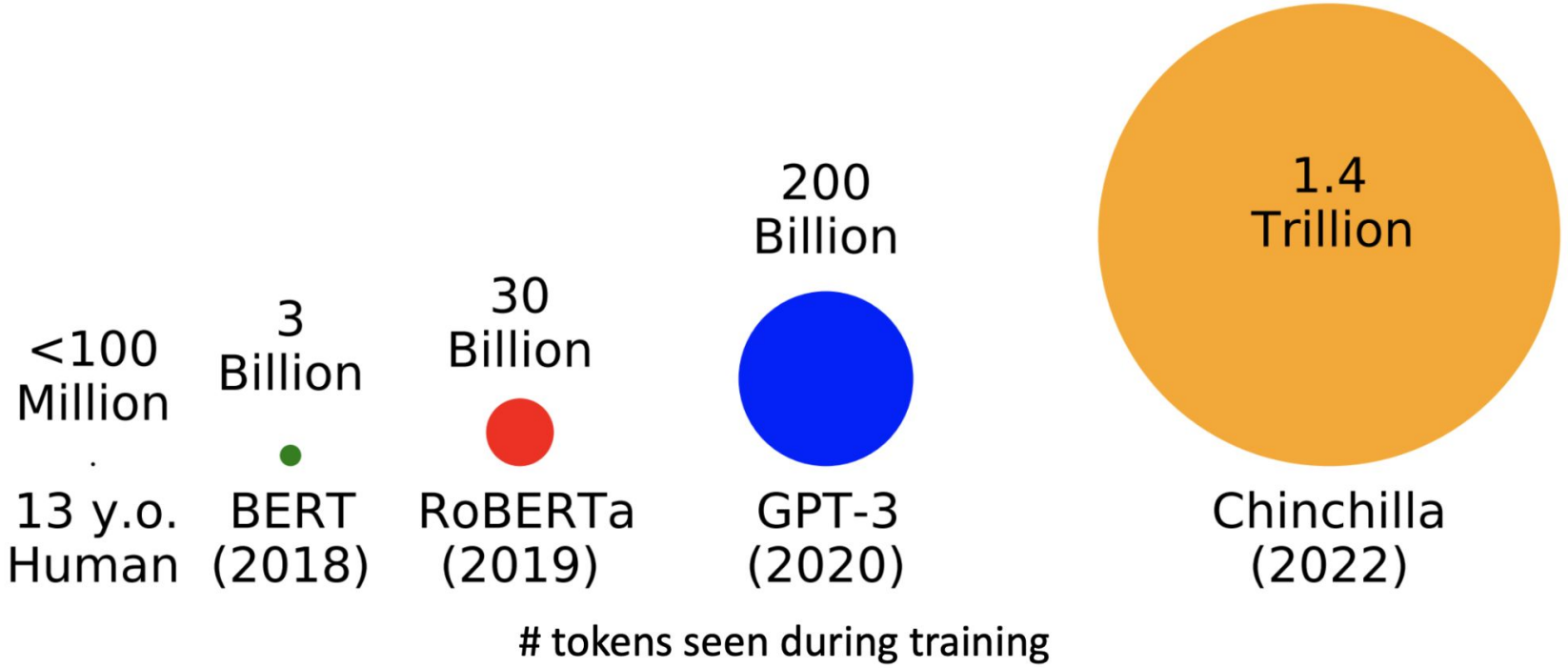
- Broad Sense
  - Decoder-only models (GPT-X, OPT, LLaMA, PaLM)
  - Encoder-only models (BERT, RoBERTa, ELECTRA)
  - Encoder-decoder models (T5, BART)

Large Language Models - **Billions of Parameters**



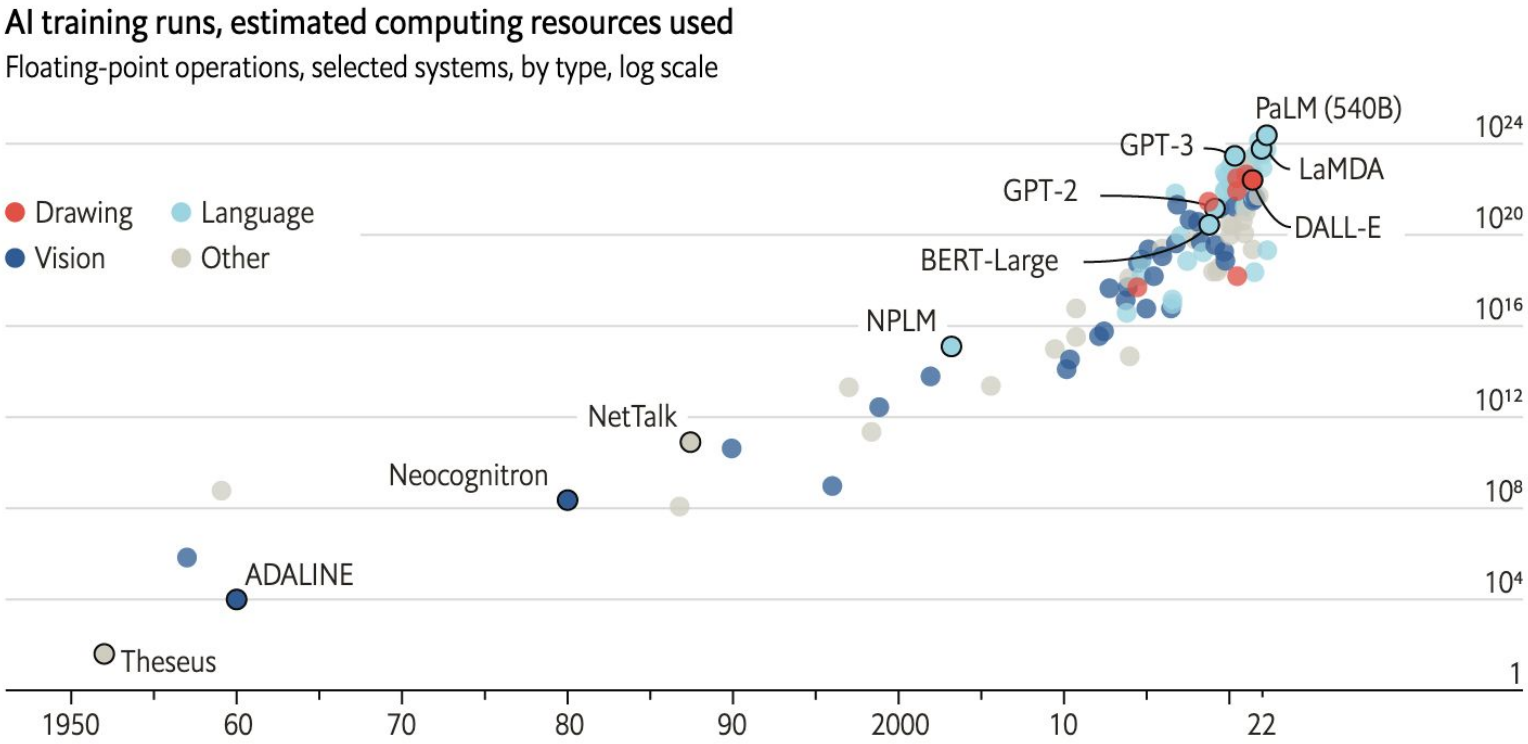
https://huggingface.co/blog/large-language-models

Large Language Models - **Hundreds of Billions of Tokens**



https://babylm.github.io/

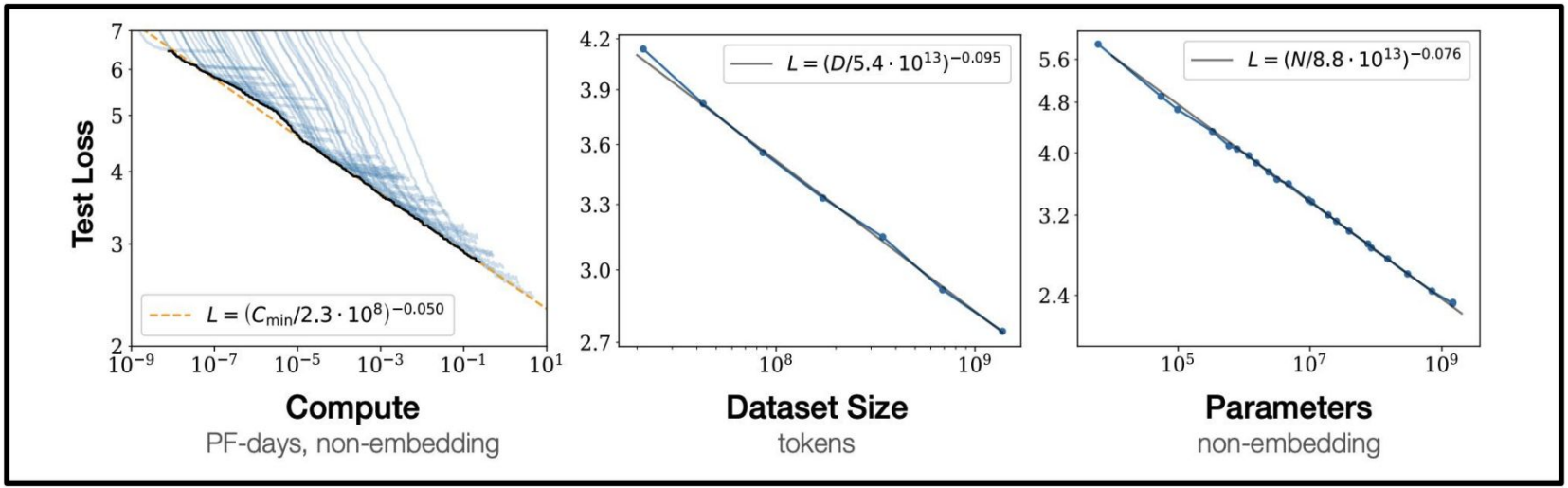
Large Language Models - yottaFlops of Compute



https://web.stanford.edu/class/cs224n/slides/cs224n-2023-lecture11-prompting-rlhf.pdf

Why LLMs?

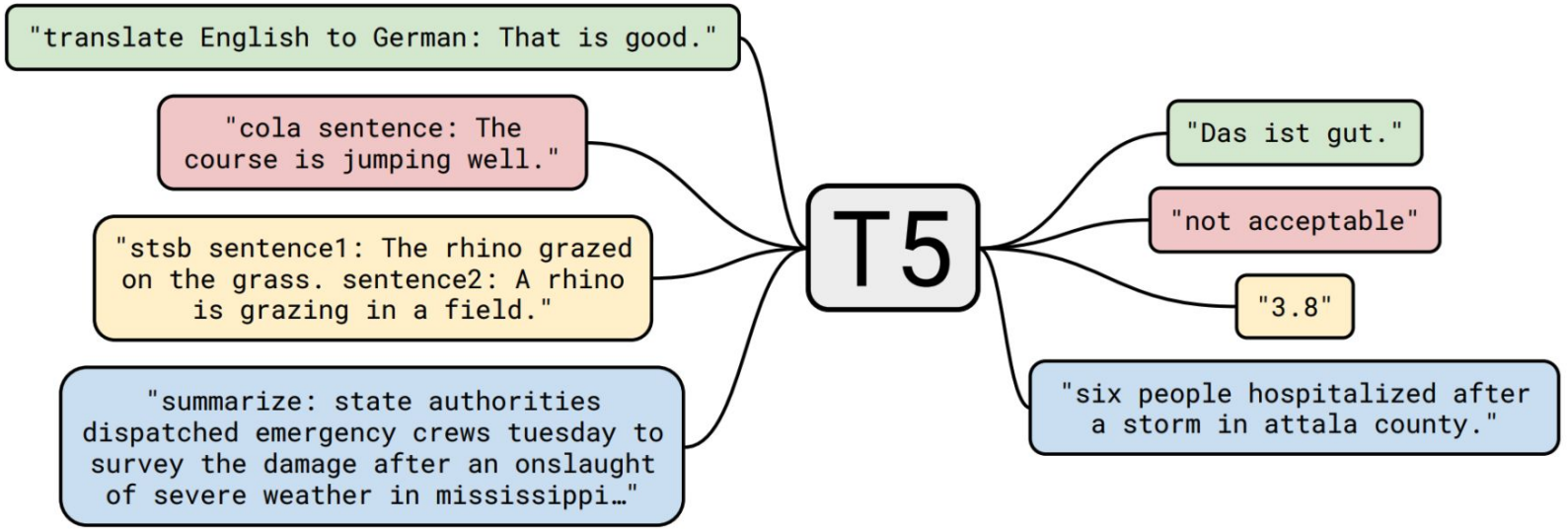
- **Scaling Law for Neural Language Models**
  - Performance depends strongly on scale! We keep getting better performance as we scale the model, data, and compute up!



https://arxiv.org/pdf/2001.08361.pdf

Why LLMs?

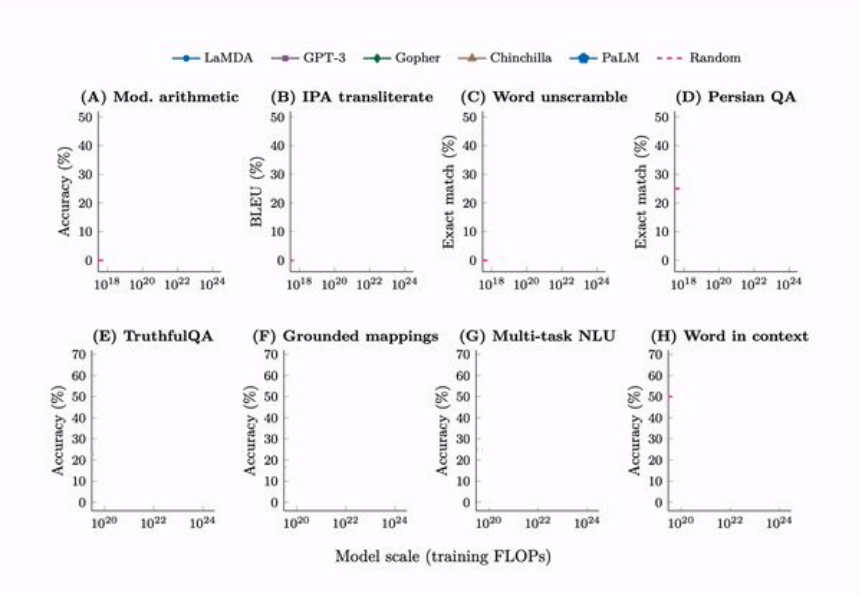
- **Generalization**
  - We can now use one single model to solve many NLP tasks



https://arxiv.org/pdf/1910.10683.pdf

Why LLMs?

- **Emergent Abilities**
  - Some ability of LM is not present in smaller models but is present in larger models



https://docs.google.com/presentation/d/1yzbmYB5E7G8lY2-KzhmArmPYwwl7o7CUST1xRZDUu1Y/edit?resourcekey=0-6_TnUMoKWCk_FN2BiPxmbw#slide=id.g1fc34b3ac18_0_27

Emergent Capability - In-Context Learning

Traditional fine-tuning (not used for GPT-3)

Fine-tuning

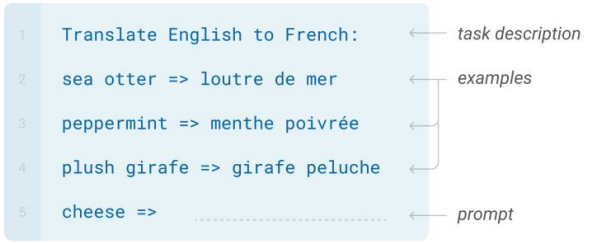
The model is trained via repeated gradient updates using a large corpus of example tasks.

One-shot

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

Few-shot

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



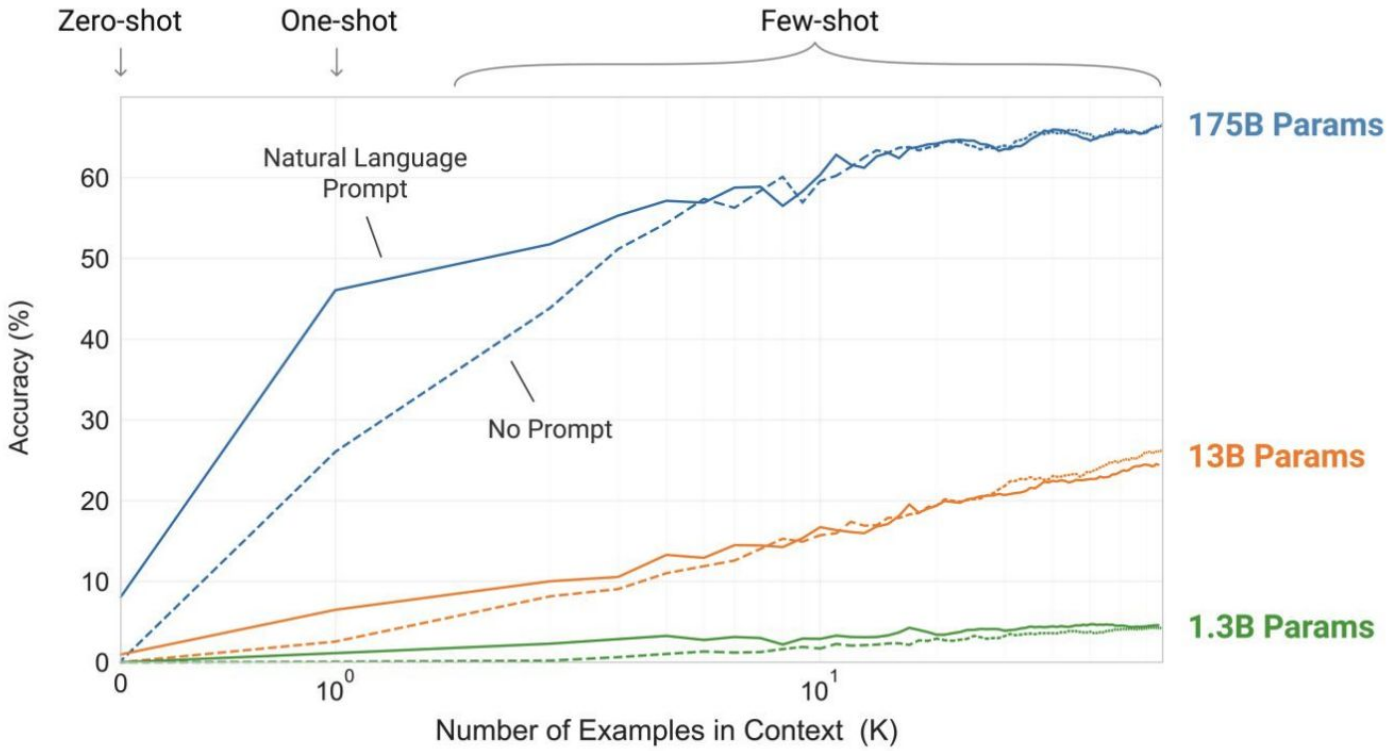
https://arxiv.org/pdf/2005.14165.pdf

Emergent Capability - In-Context Learning

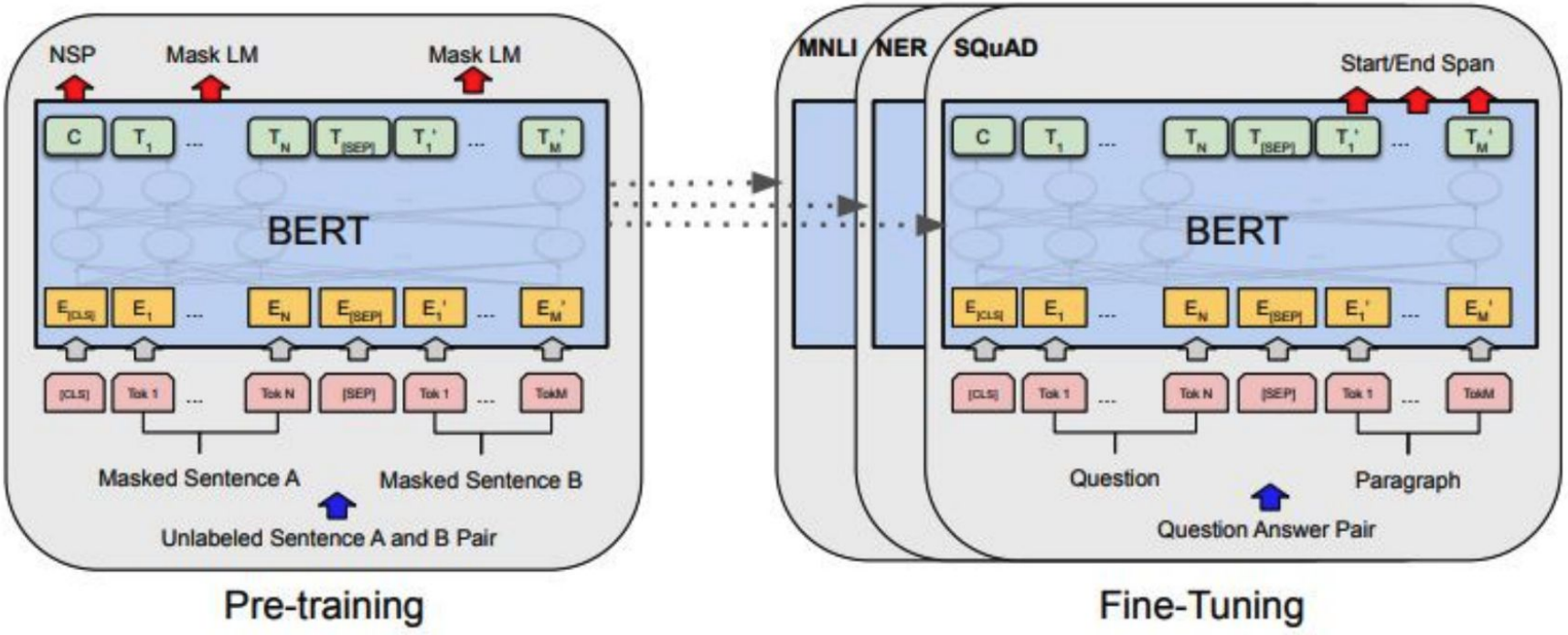
	No Prompt	Prompt
Zero-shot (0s)	skicts = sticks	Please unscramble the letters into a word, and write that word: skicts = sticks
1-shot (1s)	chiar = chair skicts = sticks	Please unscramble the letters into a word, and write that word: chiar = chair skicts = sticks
Few-shot (FS)	chiar = chair [...] pciinc = picnic skicts = sticks	Please unscramble the letters into a word, and write that word: chiar = chair [...] pciinc = picnic skicts = sticks

https://www.cs.princeton.edu/courses/archive/fall22/cos597G/lectures/lec04.pdf

Emergent Capability - **In-Context Learning**



Pretraining + Fine-tuning Paradigm



Pretraining + Prompting Paradigm

- Fine-tuning (FT)
  - + Strongest performance
  - - Need curated and labeled dataset for each new task (typically 1k-100k ex.)
  - - Poor generalization, spurious feature exploitation
- Few-shot (FS)
  - + Much less task-specific data needed
  - + No spurious feature exploitation
  - - Challenging
- One-shot (1S)
  - + "Most natural," e.g. giving humans instructions
  - - Challenging
- Zero-shot (OS)
  - + Most convenient
  - - Challenging, can be ambiguous

Stronger task-specific performance

More convenient, general, less data

Emergent Capability - Chain of Thoughts Prompting

Standard Prompting

Model 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

Model 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. ✓

https://arxiv.org/pdf/2201.11903.pdf

Emergent Capability - Chain of Thoughts Prompting

Math Word Problems (free response)

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.

Math Word Problems (multiple choice)

Q: How many keystrokes are needed to type the numbers from 1 to 500? Answer Choices: (a) 1156 (b) 1392 (c) 1480 (d) 1562 (e) 1788

A: There are 9 one-digit numbers from 1 to 9. There are 90 two-digit numbers from 10 to 99. There are 401 three-digit numbers from 100 to 500. 9 + 90(2) + 401(3) = 1392. The answer is (b).

CSQA (commonsense)

Q: Sammy wanted to go to where the people were. Where might he go? Options: (a) race track (b) populated areas (c) desert (d) apartment (e) roadblock

A: The answer must be a place with a lot of people. Race tracks, desert, apartments, and roadblocks don't have a lot of people, but populated areas do. So the answer is (b).

StrategyQA

Q: Yes or no: Would a pear sink in water?

A: The density of a pear is about 0.6 g/cm^3, which is less than water. Thus, a pear would float. So the answer is no.

Date Understanding

Q: The concert was scheduled to be on 06/01/1943, but was delayed by one day to today. What is the date 10 days ago in MM/DD/YYYY?

A: One day after 06/01/1943 is 06/02/1943, so today is 06/02/1943. 10 days before today is 05/23/1943. So the answer is 05/23/1943.

Sports Understanding

Q: Is the following sentence plausible? "Joao Moutinho caught the screen pass in the NFC championship."

A: Joao Moutinho is a soccer player. The NFC championship is part of American football, not soccer. So the answer is no.

SayCan (Instructing a robot)

Human: How would you bring me something that isn't a fruit?

Explanation: the user wants something to eat that isn't a fruit. An energy bar is not a fruit, so I will bring the user an energy bar.
Plan: 1. find(energy bar) 2. pick(energy bar) 3. find(user) 4. put(energy bar) 5. done().

Last Letter Concatenation

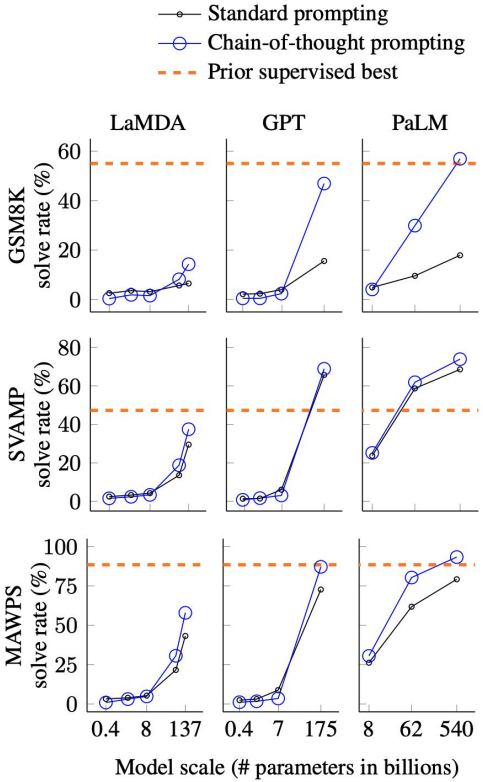
Q: Take the last letters of the words in "Lady Gaga" and concatenate them.

A: The last letter of "Lady" is "y". The last letter of "Gaga" is "a". Concatenating them is "ya". So the answer is ya.

Coin Flip (state tracking)

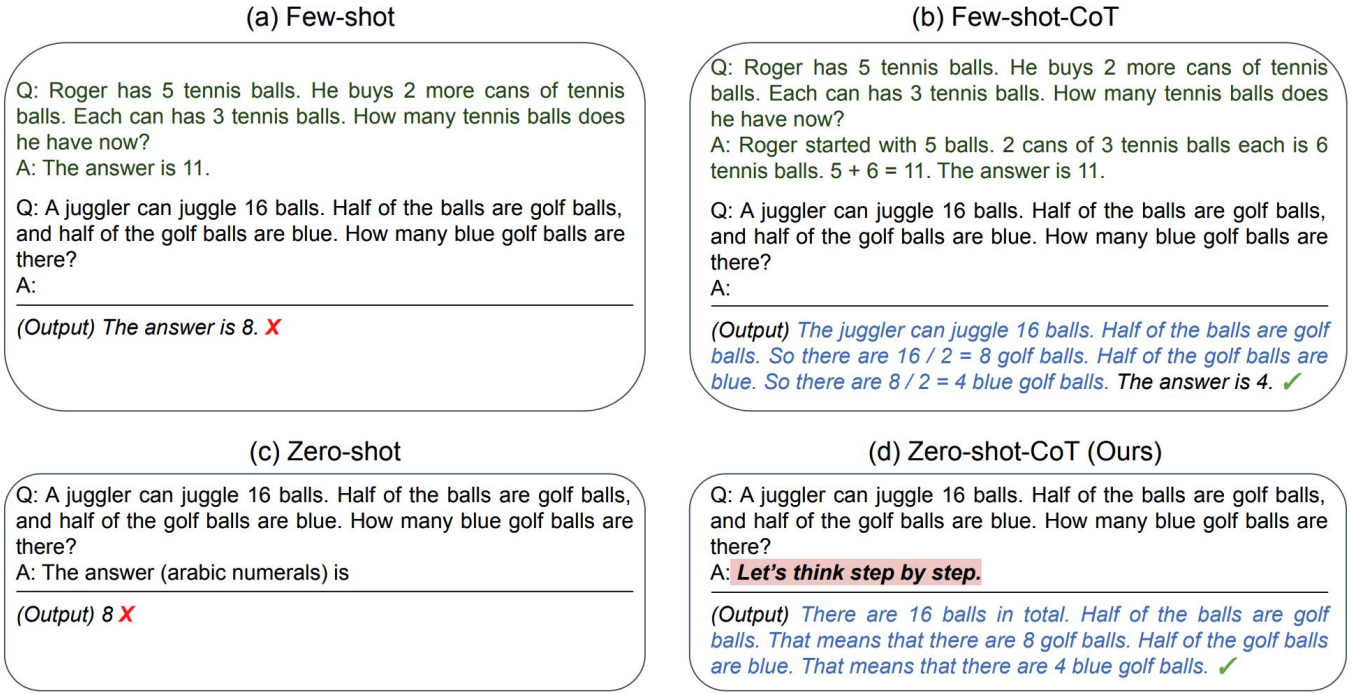
Q: A coin is heads up. Maybelle flips the coin. Shalonda does not flip the coin. Is the coin still heads up?

A: The coin was flipped by Maybelle. So the coin was flipped 1 time, which is an odd number. The coin started heads up, so after an odd number of flips, it will be tails up. So the answer is no.



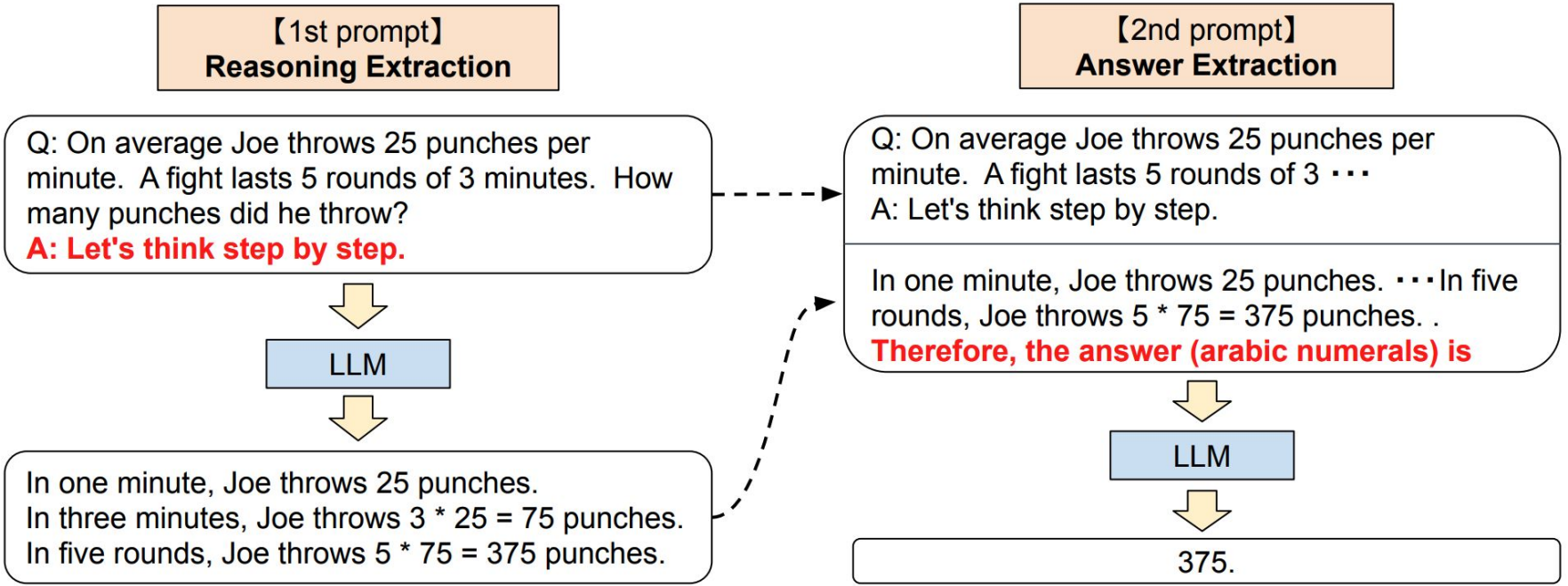
https://arxiv.org/pdf/2201.11903.pdf

Emergent Capability - **Zero Shot CoT Prompting**



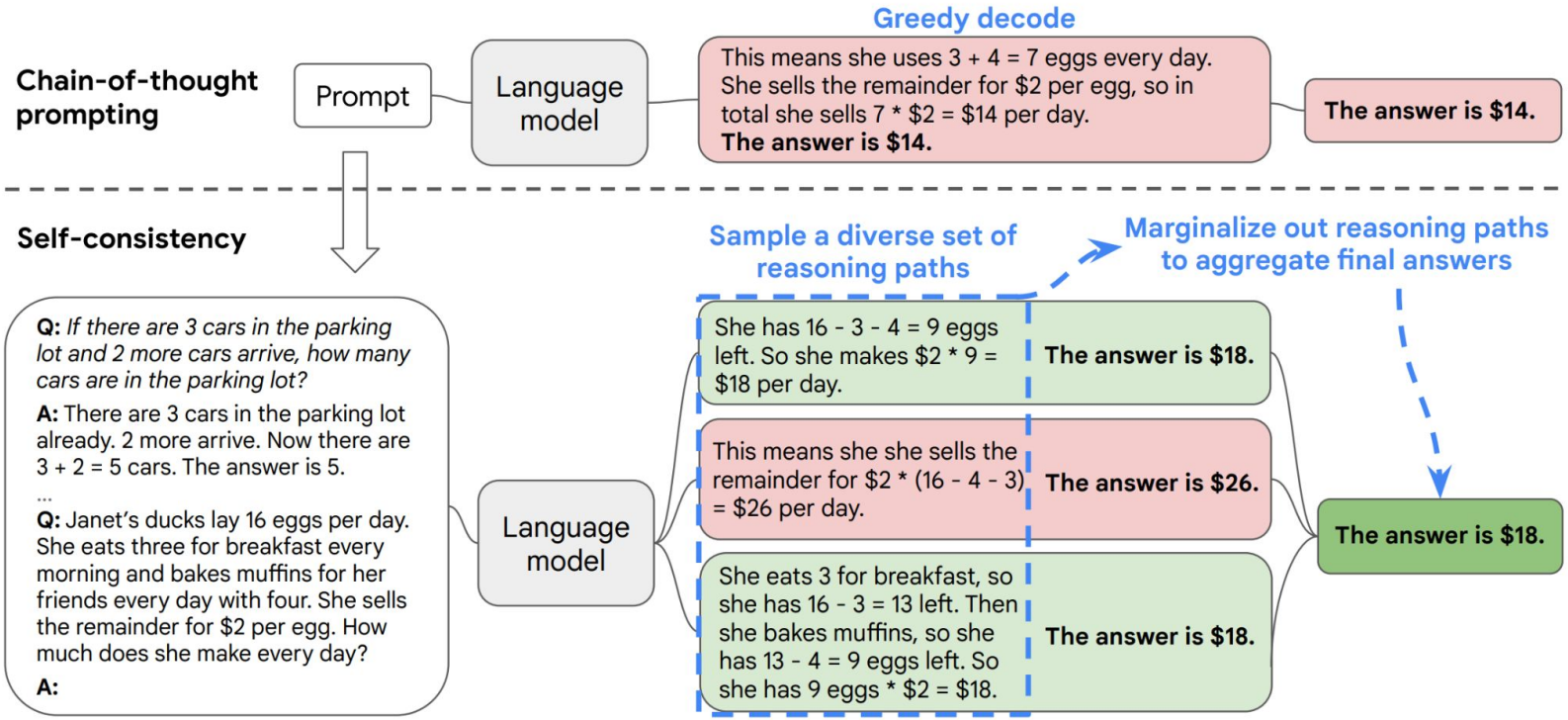
https://arxiv.org/pdf/2205.11916.pdf

Emergent Capability - **Zero Shot CoT Prompting**



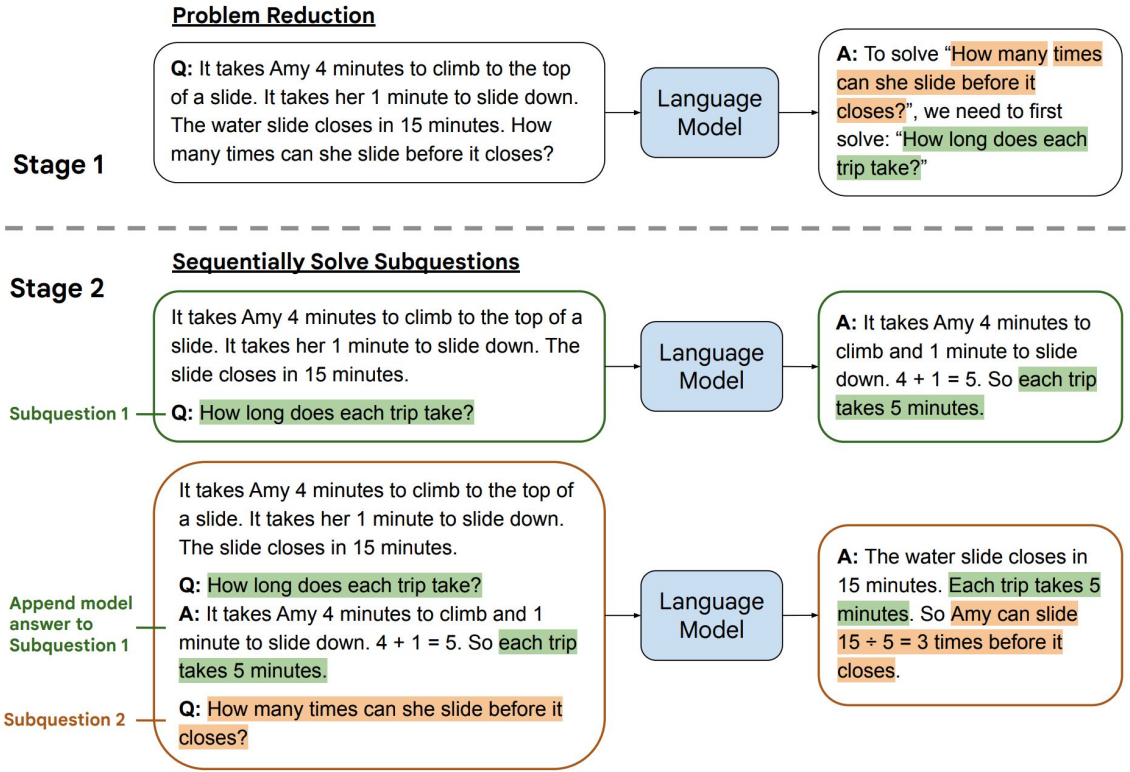
https://arxiv.org/pdf/2205.11916.pdf

Emergent Capability - **Self-Consistency Prompting**



https://arxiv.org/pdf/2203.11171.pdf

Emergent Capability - Least-to-Most Prompting



https://arxiv.org/pdf/2205.10625.pdf

Emergent Capability - Augmented Prompting Abilities

Advanced Prompting Techniques

- Zero-shot CoT Prompting
- Self-Consistency
- Divide-and-Conquer

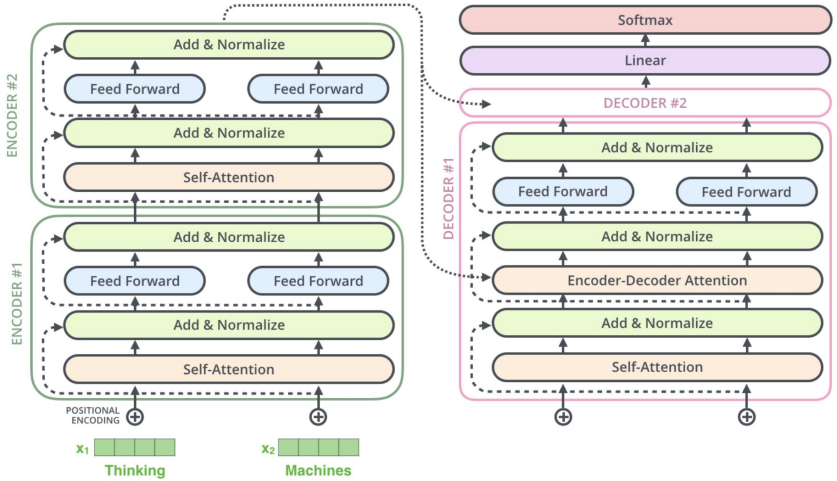
Ask a human to

- Explain the rationale
- Double check the answer
- Decompose to easy subproblems

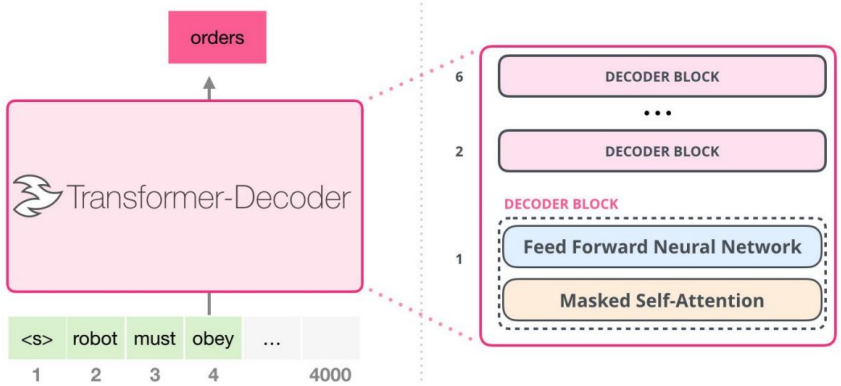
Large Language Models demonstrate some human-like behaviors!

Training Architectures

Encoder-decoder models (T5, BART)

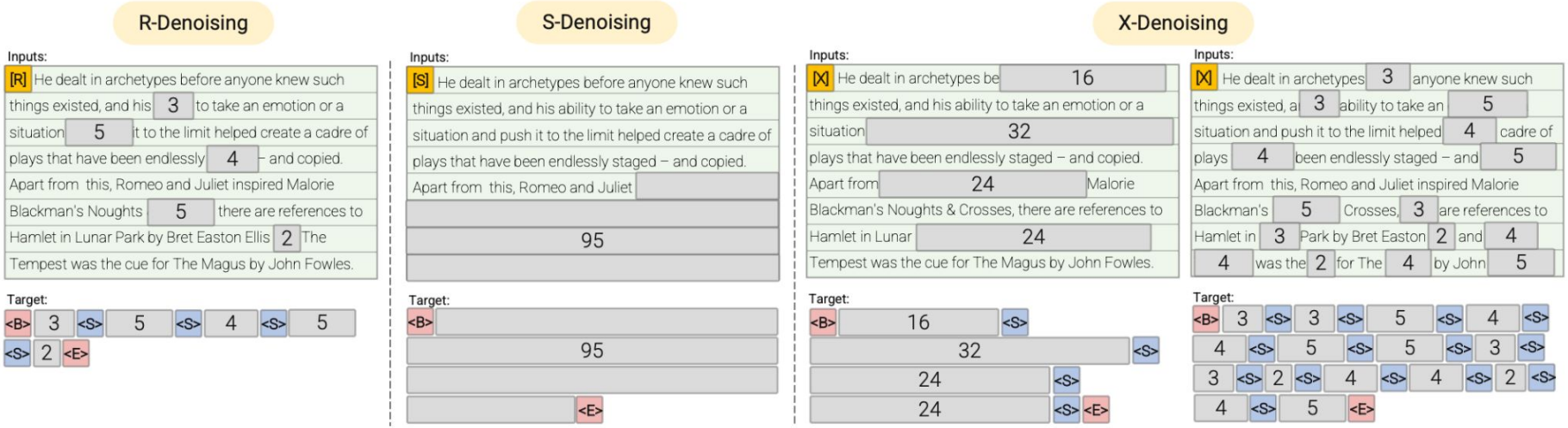


Decoder-only models (GPT-X, PaLM)



http://jalammar.github.io/illustrated-transformer/

Training Objectives - **UL2**



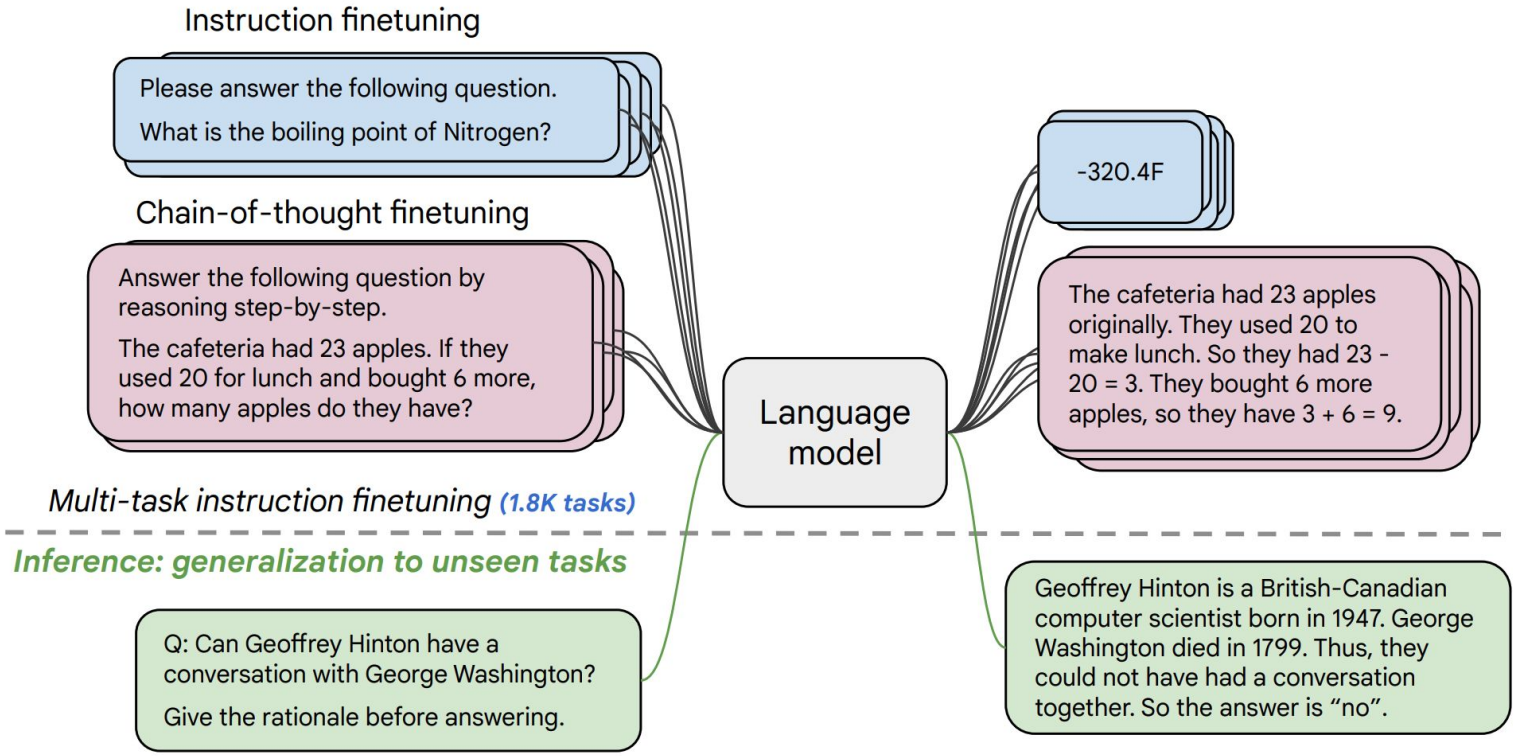
https://arxiv.org/pdf/2205.05131.pdf

What kinds of things does pretraining learn?

- *Stanford University is located in _____, California.* [Trivia]
- *I put ___ fork down on the table.* [syntax]
- *The woman walked across the street, checking for traffic over ___ shoulder.* [coreference]
- *I went to the ocean to see the fish, turtles, seals, and _____.* [lexical semantics/topic]
- *Overall, the value I got from the two hours watching it was the sum total of the popcorn and the drink. The movie was ___.* [sentiment]
- Iroh went into the kitchen to make some tea. Standing next to Iroh, Zuko pondered his destiny. Zuko left the _____. [some reasoning – this is harder]
- I was thinking about the sequence that goes 1, 1, 2, 3, 5, 8, 13, 21, ____ [some basic arithmetic; they don't learn the Fibonnaci sequence]

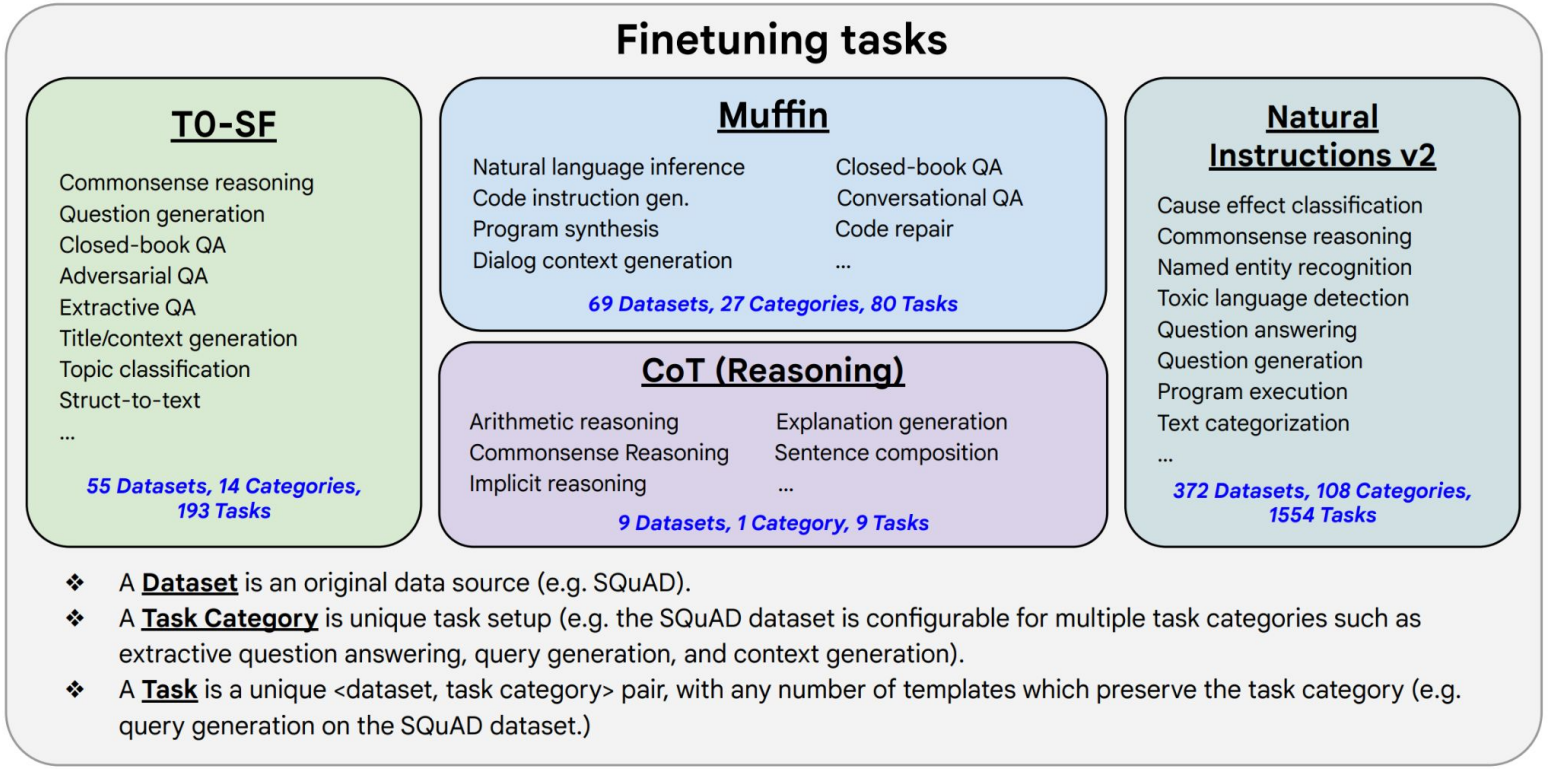
https://web.stanford.edu/class/cs224n/slides/cs224n-2023-lecture11-prompting-rlhf.pdf

Finetune - **Instruction Finetune**



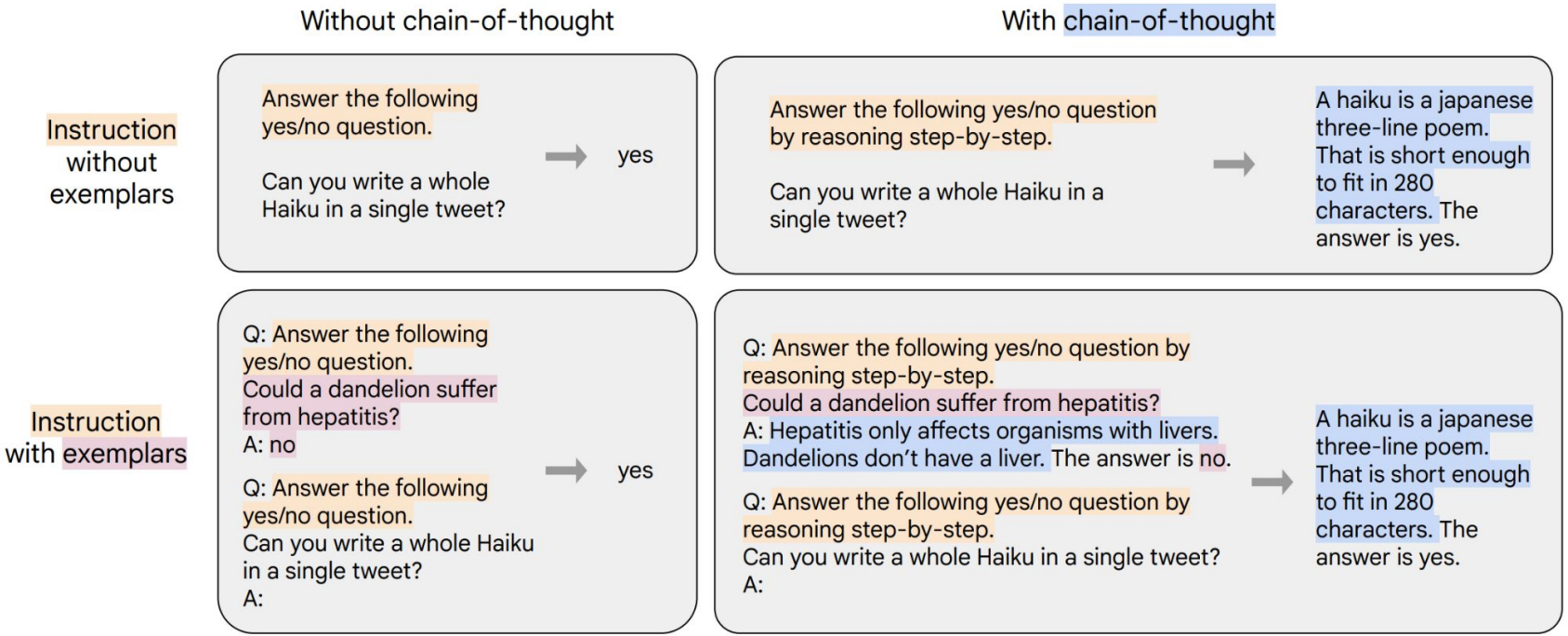
https://arxiv.org/pdf/2210.11416.pdf

Finetune - **Instruction Finetune**



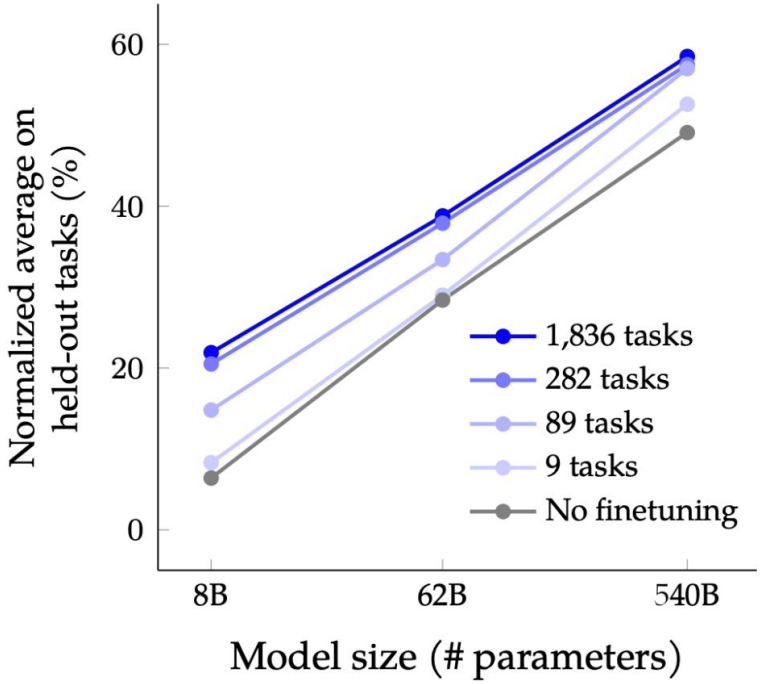
https://arxiv.org/pdf/2210.11416.pdf

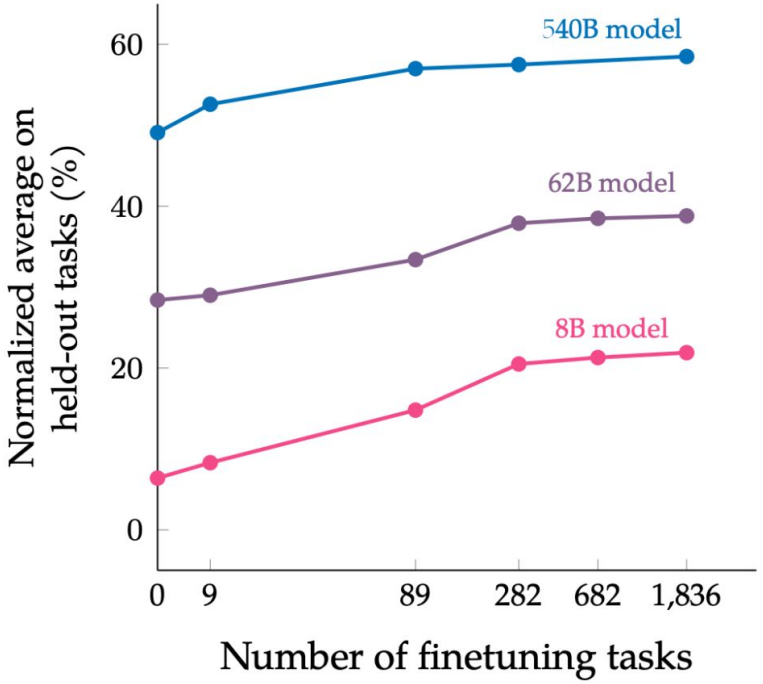
Finetune - **Instruction Finetune**



https://arxiv.org/pdf/2210.11416.pdf

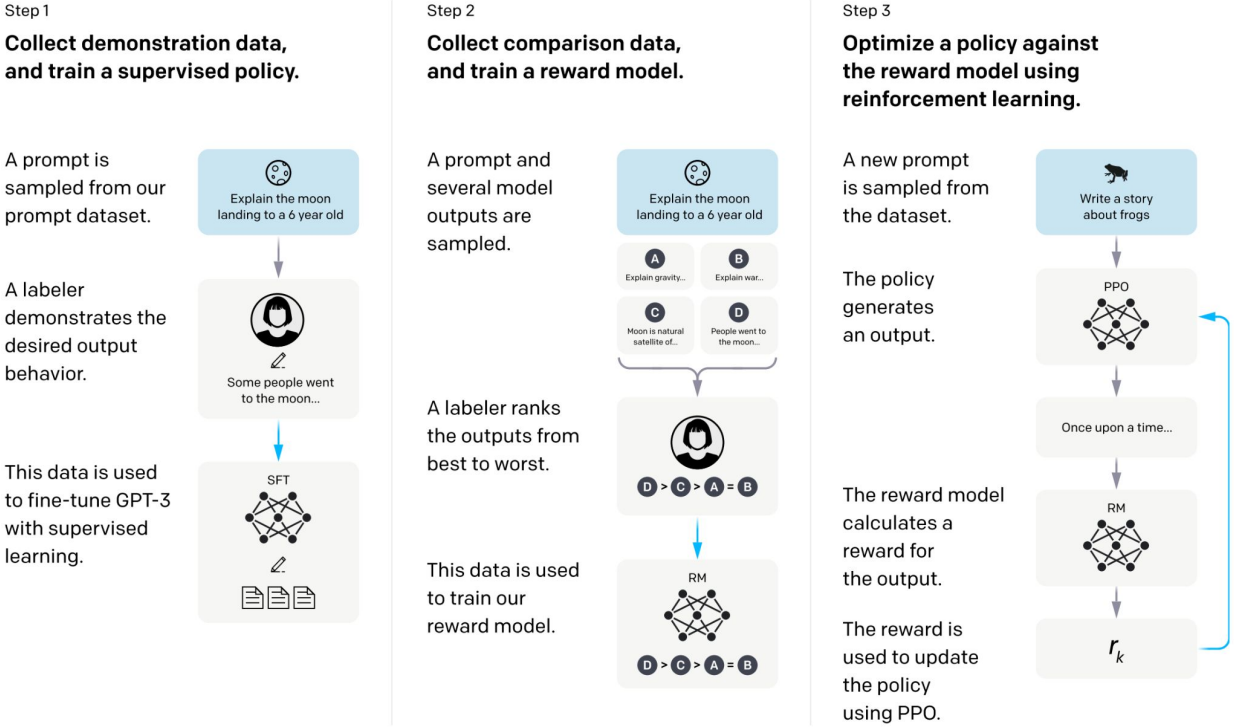
Finetune - **Instruction Finetune**





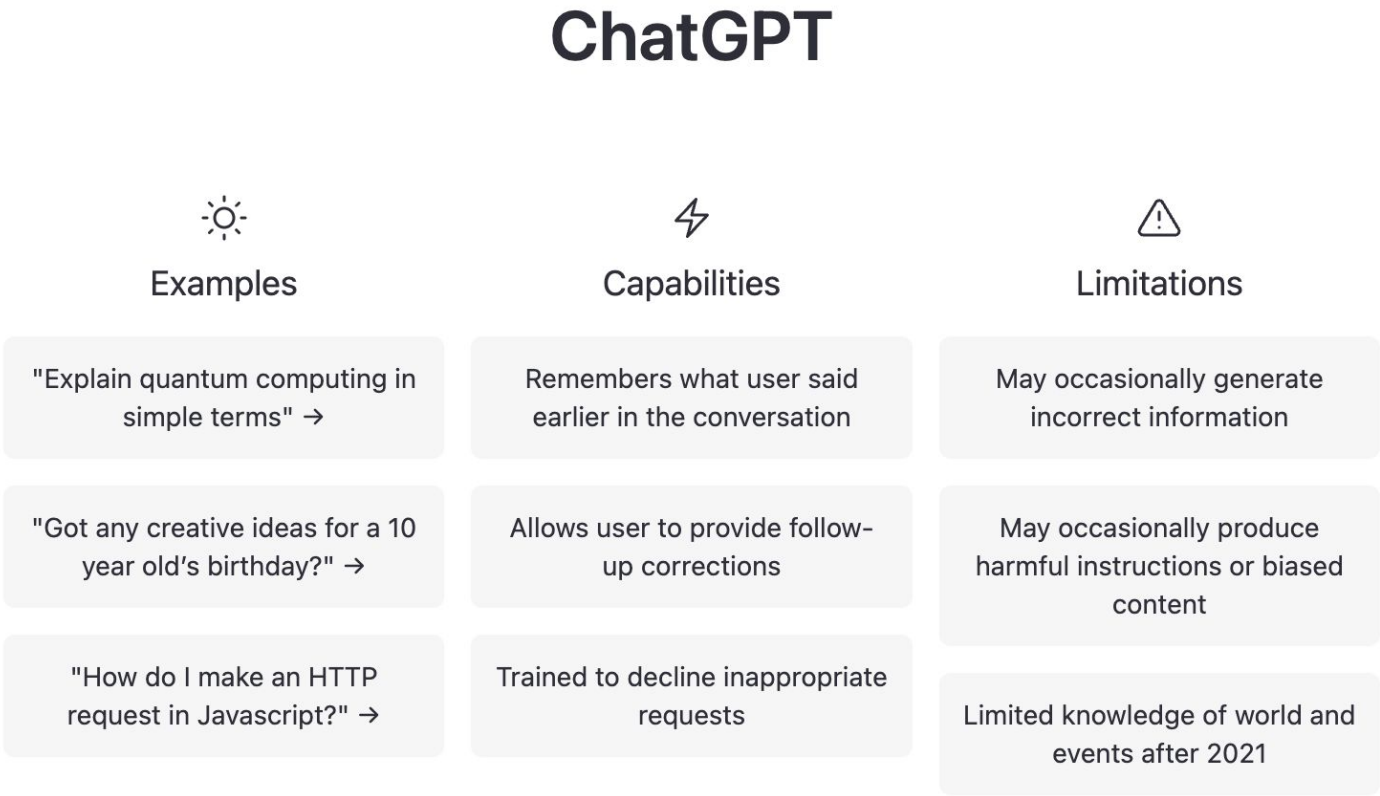
https://arxiv.org/pdf/2210.11416.pdf

Finetune - **RLHF**

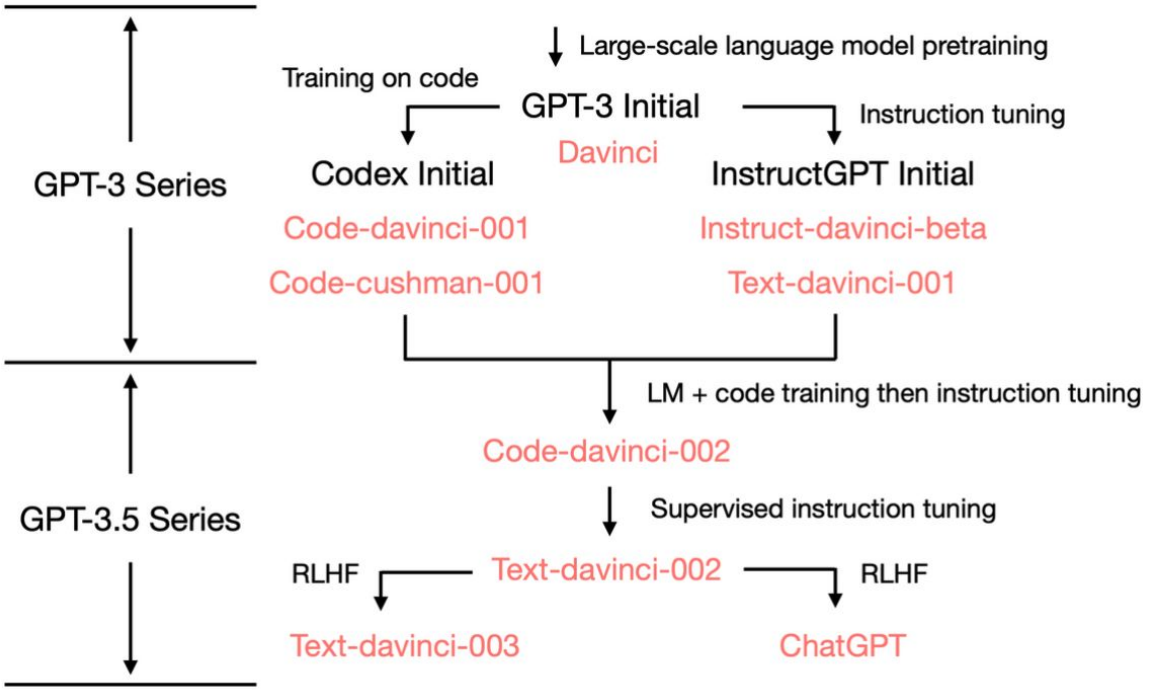


https://arxiv.org/pdf/2203.14465.pdf

Application - ChatGPT

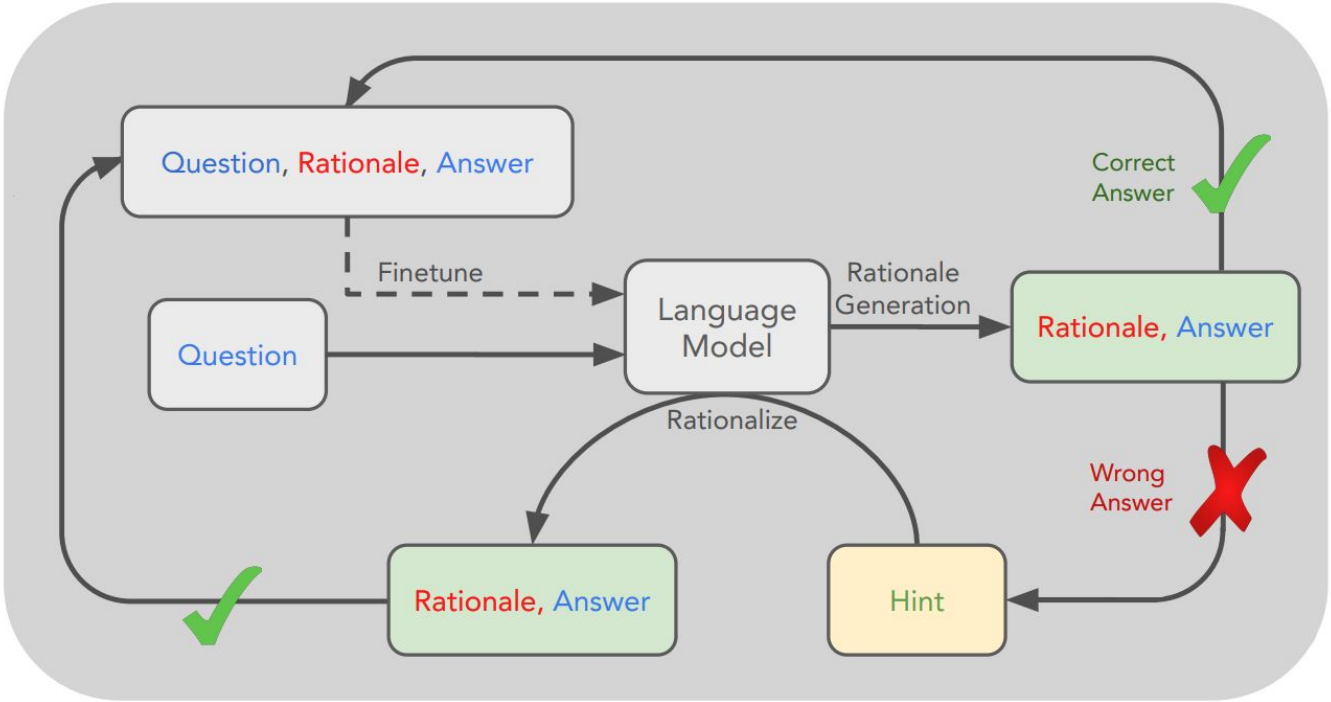


Application - **ChatGPT**



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

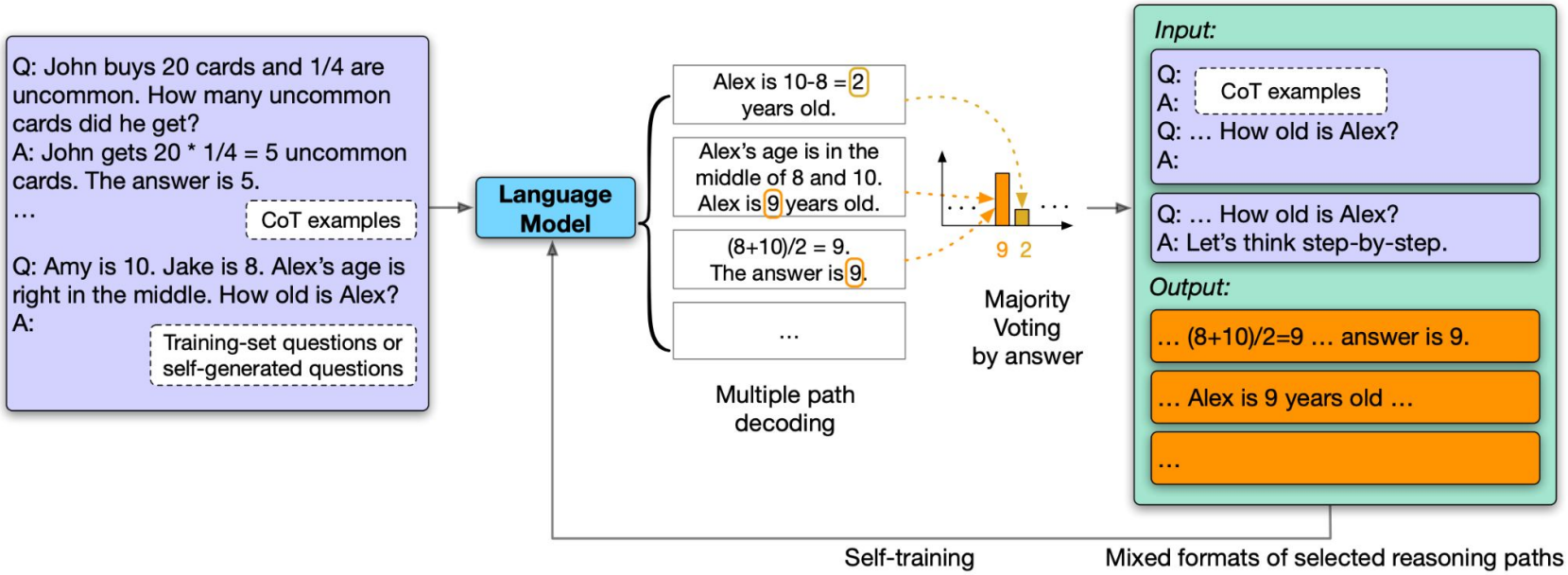
Finetune - **Bootstrapping**




```
Q: What can be used
to carry a small dog?
Answer Choices:
(a) swimming pool
(b) basket
(c) dog show
(d) backyard
(e) own home
A: The answer must be
something that can be
used to carry a small
dog. Baskets are
designed to hold things.
Therefore, the answer
is basket (b).
```

https://arxiv.org/pdf/2203.14465.pdf

Finetune - **Bootstrapping**



https://arxiv.org/pdf/2210.11610.pdf

Large Language models Risks

- LLMs make mistakes

  (falsehoods, hallucinations)
- LLMs can be misused

  (misinformation, spam)
- LLMs can cause harms

  (toxicity, biases, stereotypes)
- LLMs can be attacked

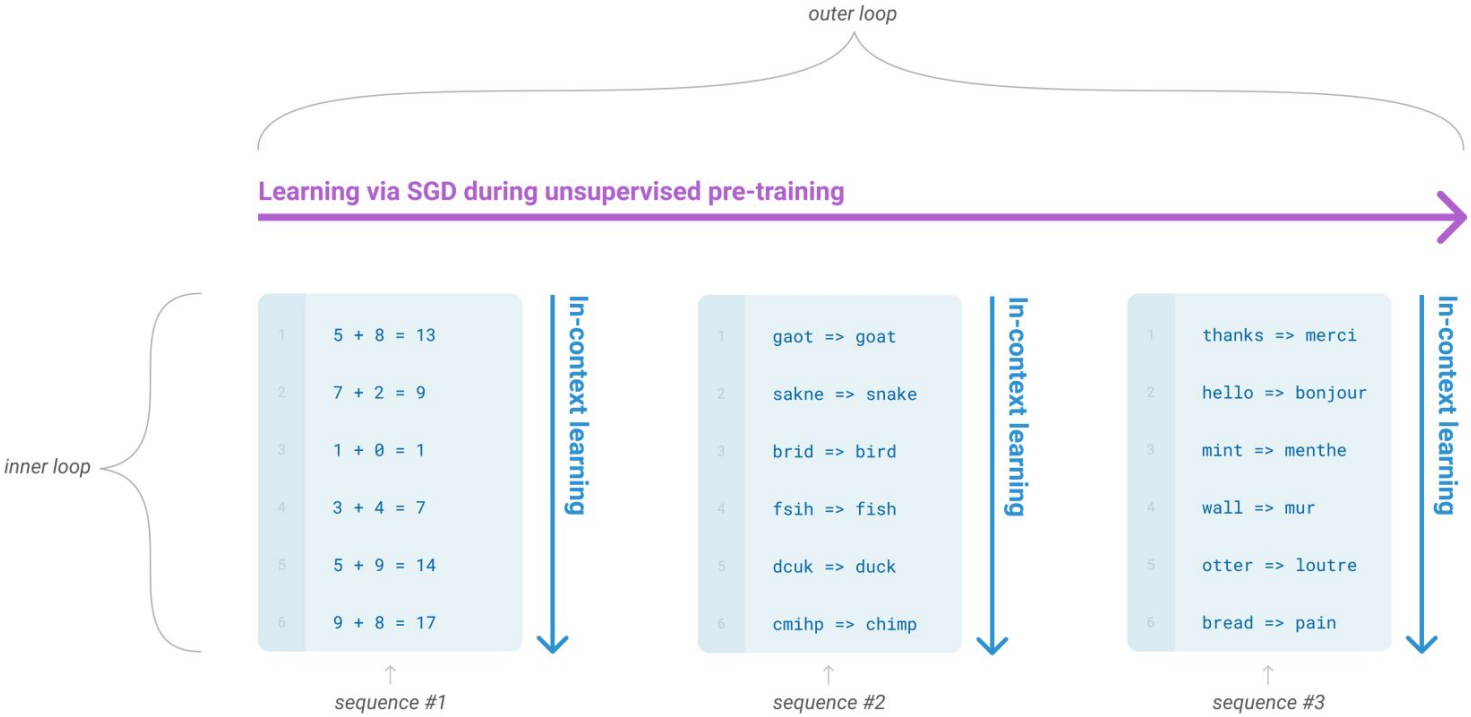
  (adversarial examples, poisoning, prompt injection)
- LLMs can be useful as defenses

  (content moderation, explanations)

Resources for further reading

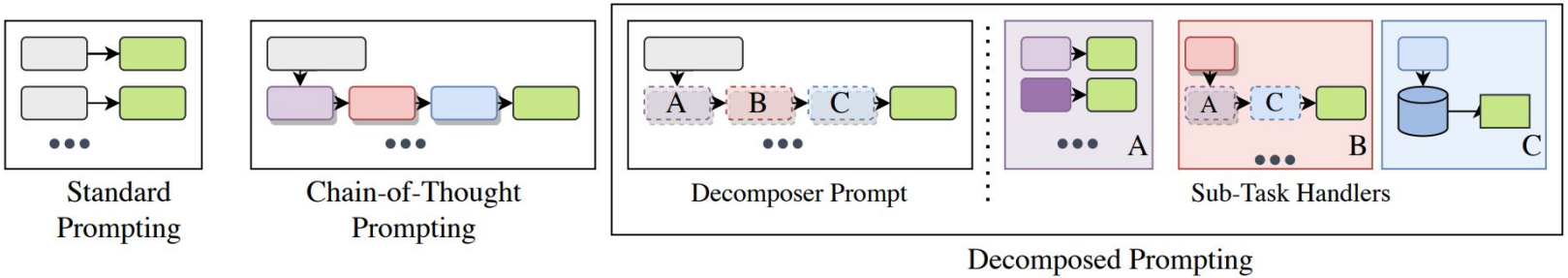
- https://web.stanford.edu/class/cs224n/
- https://stanford-cs324.github.io/winter2022/
- https://stanford-cs324.github.io/winter2023/
- https://www.cs.princeton.edu/courses/archive/fall22/cos597G/
- https://rycolab.io/classes/llm-s23/
- https://yaofu.notion.site/How-does-GPT-Obtain-its-Ability-Tracing-Emergent-Abilities-of-Language-Models-to-their-Sources-b9a57ac0fcf74f30a1ab9e3e36fa1dc1
- https://www.jasonwei.net/blog/emergence

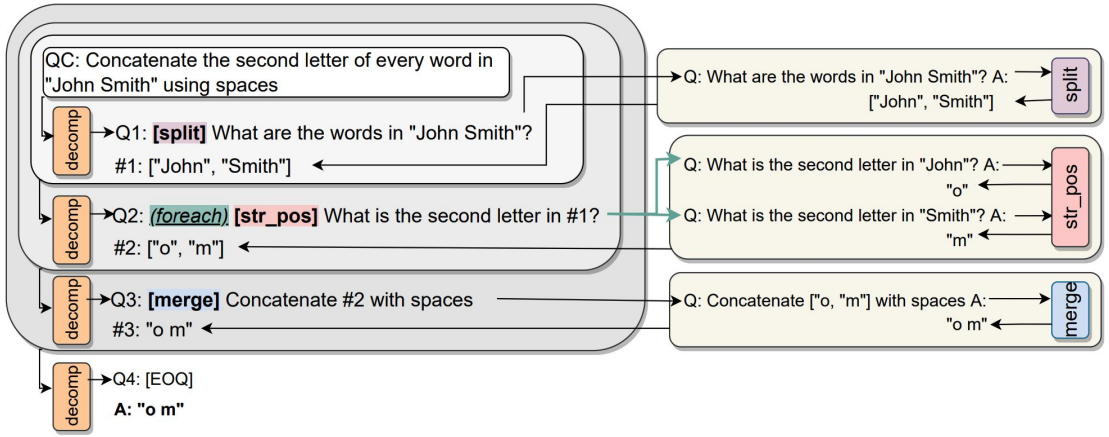
Emergent Capability - **In-Context Learning**



https://arxiv.org/pdf/2005.14165.pdf

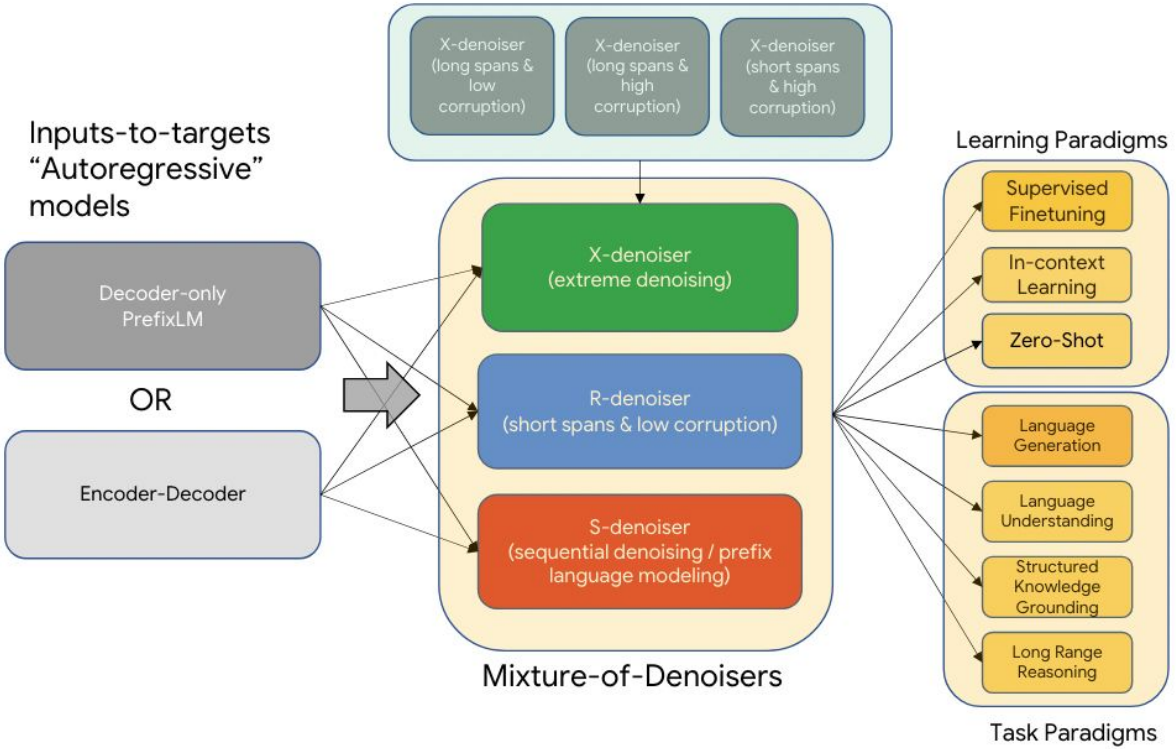
Emergent Capability - Decomposed Prompting





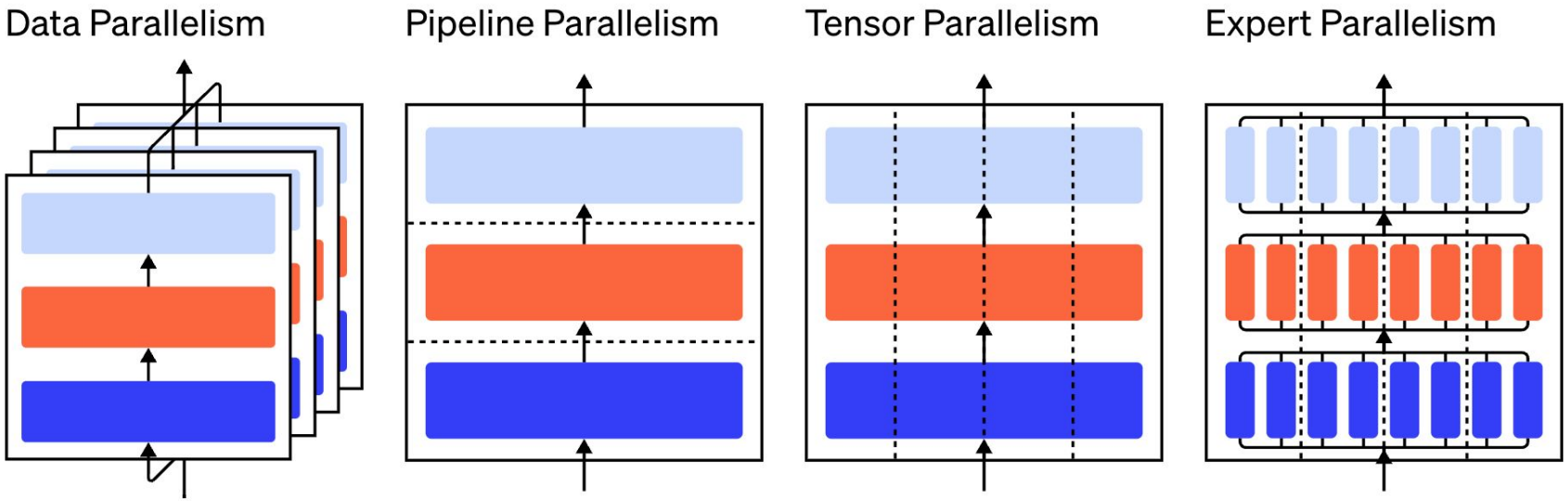
https://arxiv.org/pdf/2210.02406.pdf

Training Objectives - UL2



https://arxiv.org/pdf/2205.05131.pdf

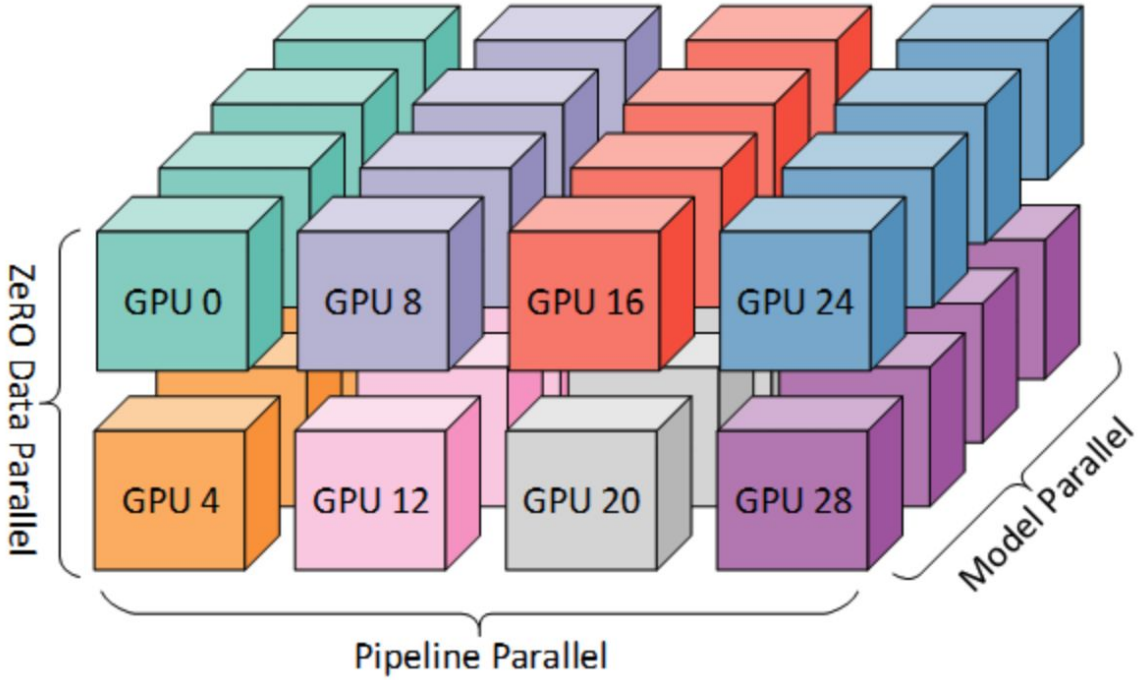
Training Techniques - **Parallelism**



An illustration of various parallelism strategies on a three-layer model. Each color refers to one layer and dashed lines separate different GPUs.

https://openai.com/research/techniques-for-training-large-neural-networks

Training Techniques - **Parallelism**



https://www.microsoft.com/en-us/research/blog/deepspeed-extreme-scale-model-training-for-everyone/