



Cornell Bowers C-IS

College of Computing and Information Science

Deep Learning

Week [03]: [LLMs]

Logistics

- HW2 is out
 - Due next Thursday
- We released a feedback form on Canvas
 - Due Friday February 23
 - Part of your participation grade!
- We are talking about projects today!
 - HW3 will be a shorter

Project

- Aim: to get hands on experience with implementing modern deep learning methods
- Find a recent deep learning research paper that was published in the last 1-5 years
- Reproduce a specific result from the paper
- To be completed in groups of 2-3

Project Grading

- Project proposal
 - We want to make sure the papers you've chosen are reasonable
 - Sign up on spreadsheet, we don't want duplicates!
- Final presentation
 - The last 2-3 classes will be devoted to presentations
 - Present the paper you've chosen to the class
 - Discuss whether you were able to replicate results from the paper and describe any obstacles
- Final deliverable
 - Github repo with re-implementation
 - Readme with method description, instructions to run the code and results
 - More detail will be provided later

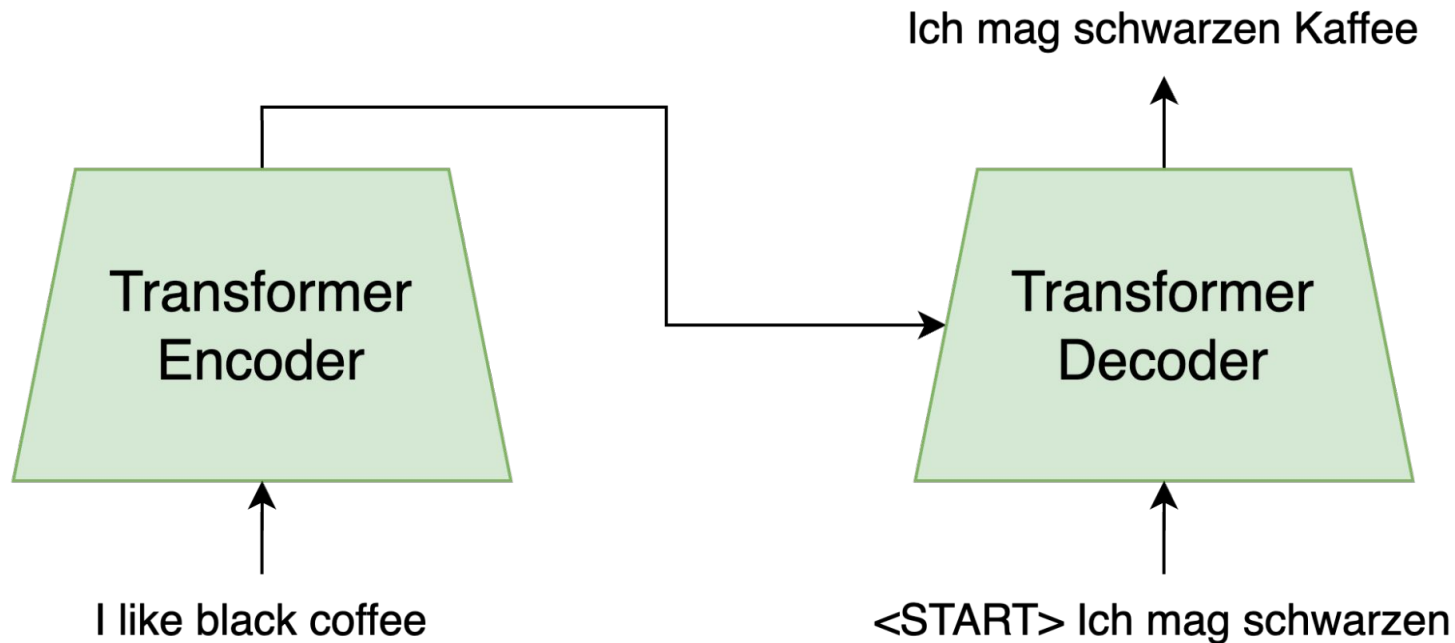
Project Proposal

A page long project proposal due **March 7**. It should contain the following:

- Paper selection:
 - Title, authors, and publication venue of the chosen paper
 - Brief summary of the chosen paper
 - Brief justification of why you choose this paper
- Result Selection
 - Tell us which result you want to replicate
- Re-implementation Plan
 - Describe architecture, method, and metrics
 - Details about how much compute and time is required to replicate results
- **Detailed instruction will be provided on canvas**

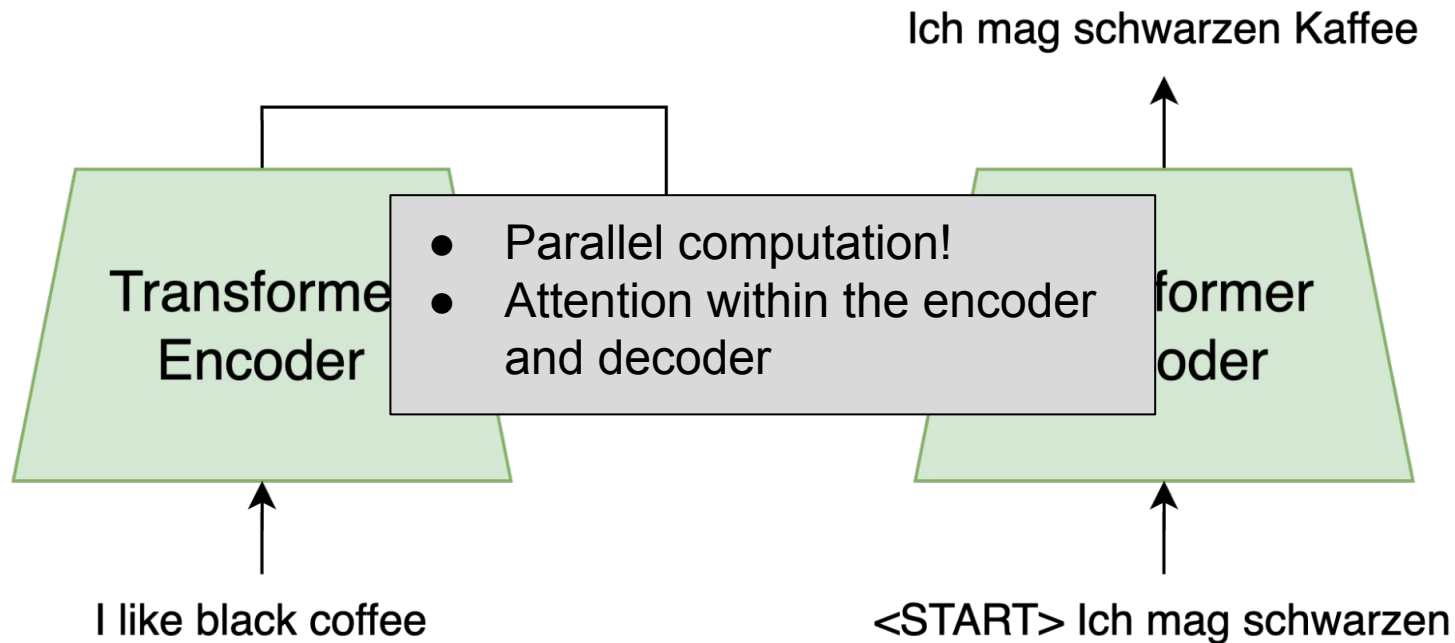
Transformer Architecture

Introduced for seq2seq tasks like Machine translation, summarization, question answering, etc.

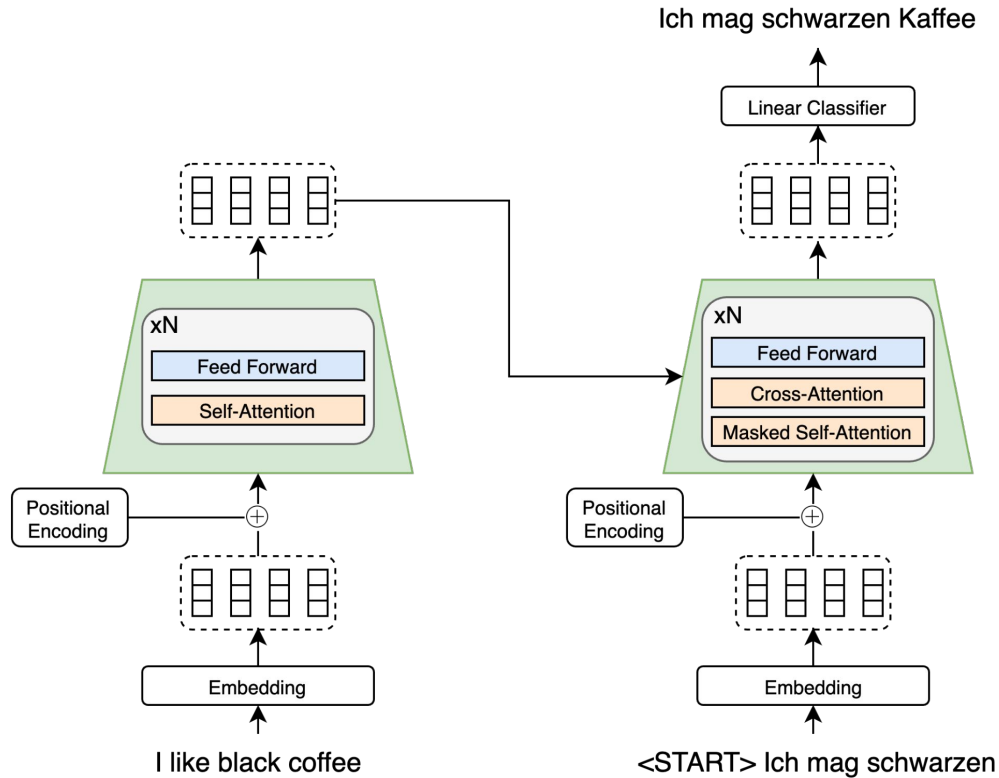


Transformer Architecture

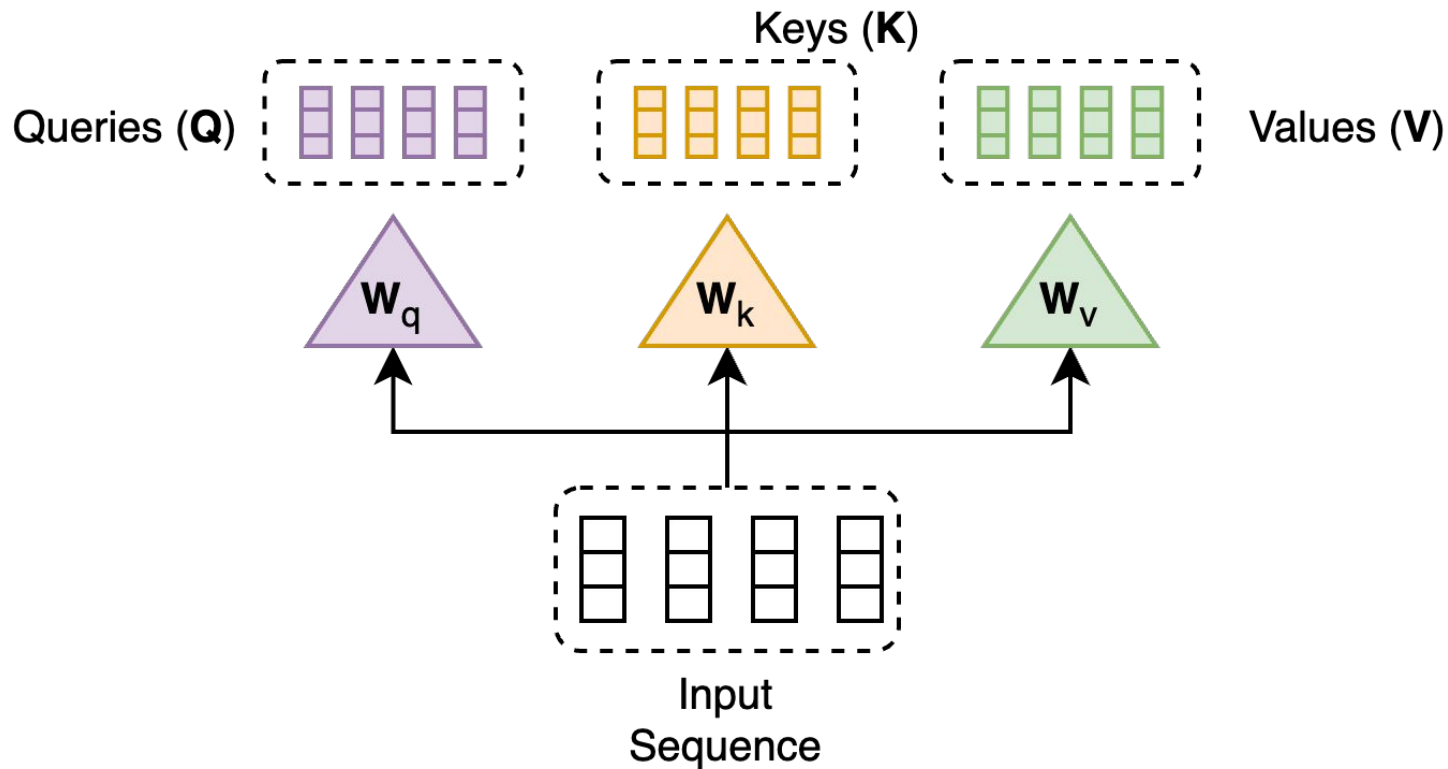
Introduced for seq2seq tasks like Machine translation, summarization, question answering, etc.



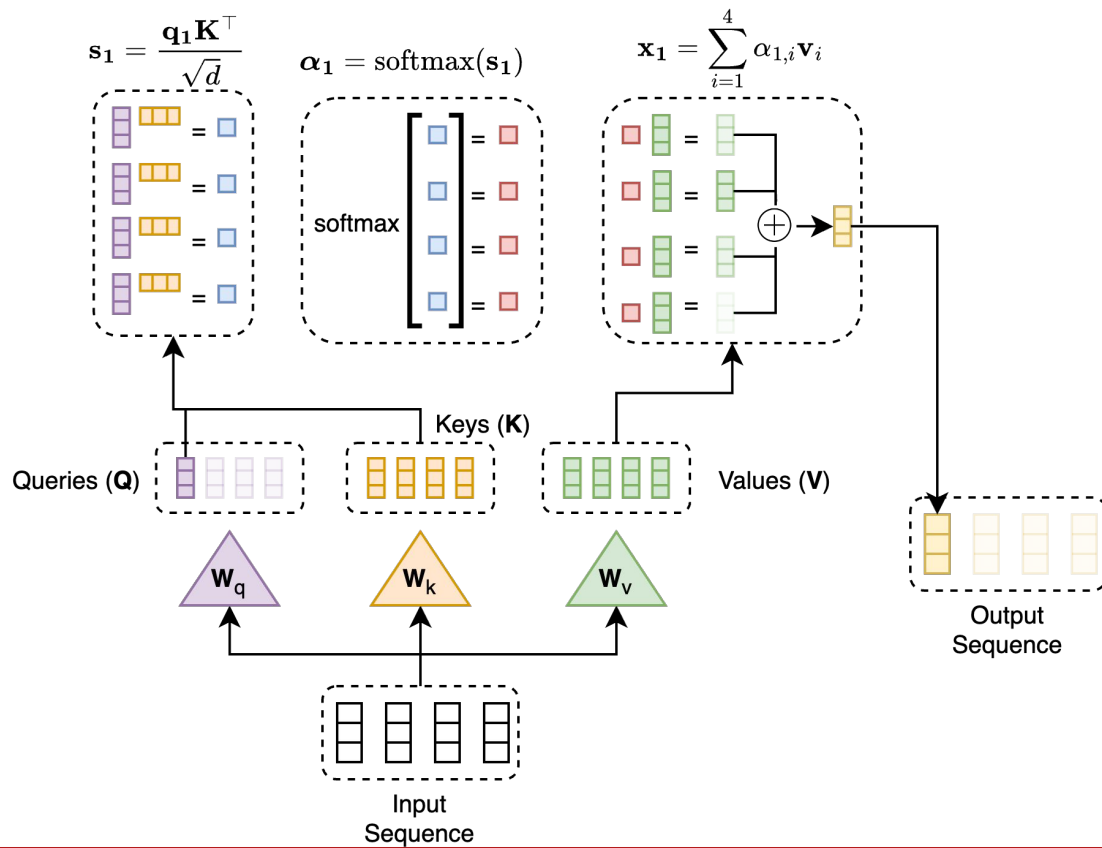
Transformer Architecture



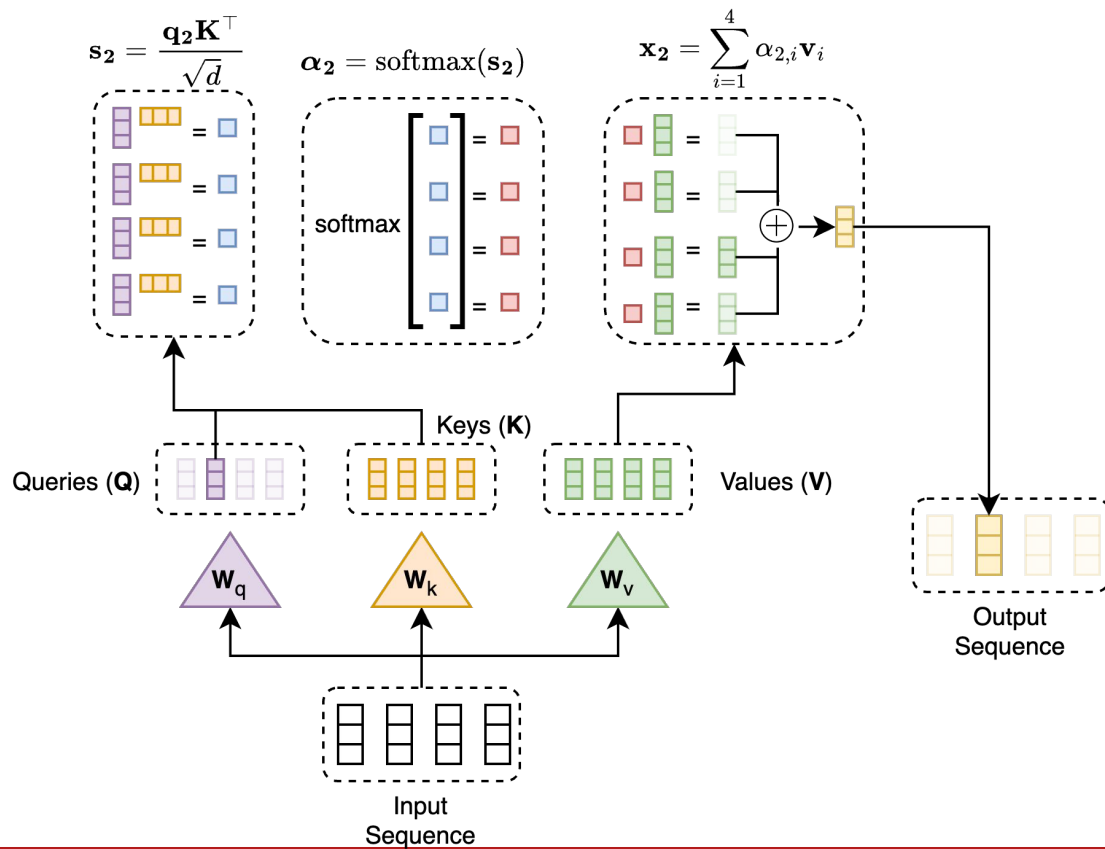
Self-Attention



Self-Attention



Self-Attention



Self-Attention: General Formula

$$\textit{Attention}(Q, K, V) = \textit{Softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V$$

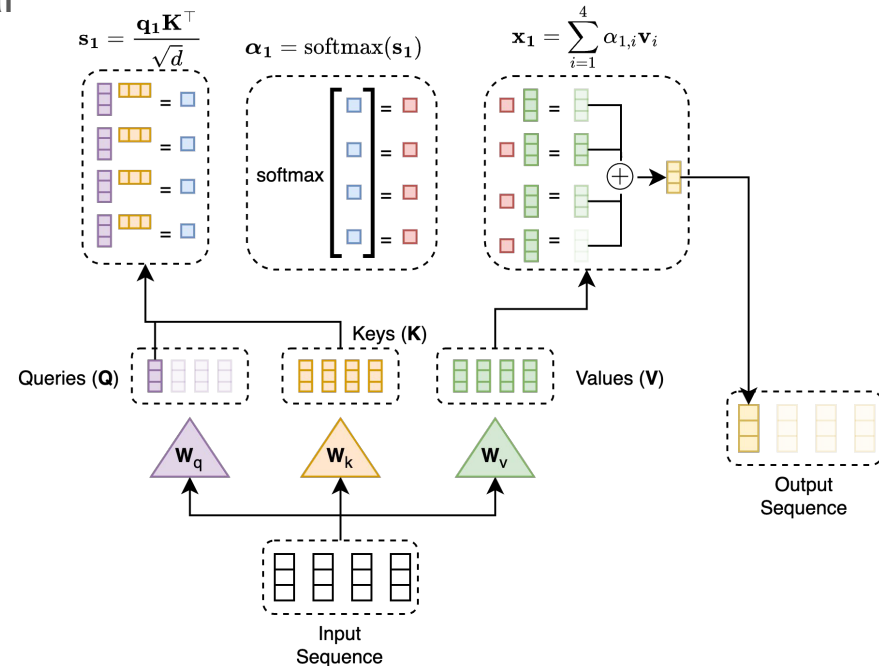
Discuss:

Assume input sequence is $(n \times d)$,

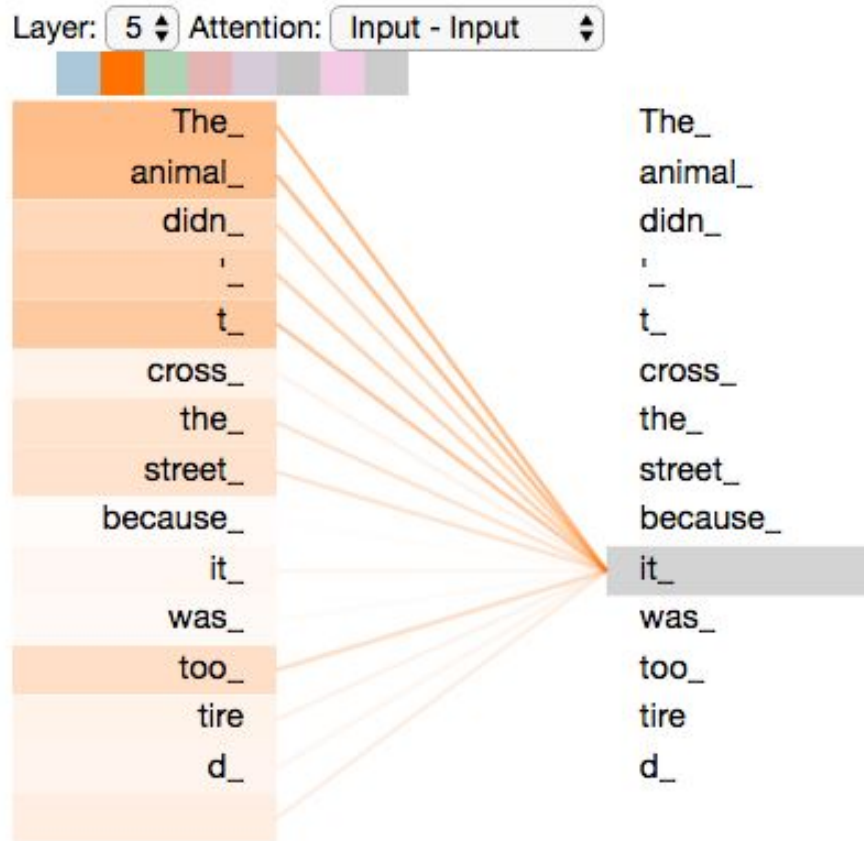
Queries, keys, and values are d -dimensional

What is the size of W_q , W_k , W_v ?

What is the size of Q , K , V ?



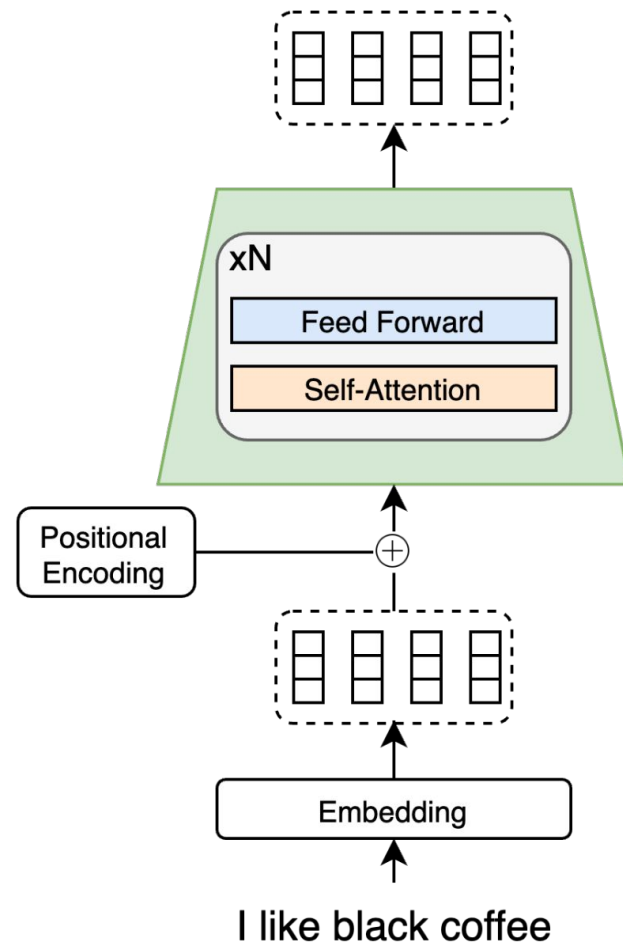
Self-Attention



Position-wise Feed-forward Networks

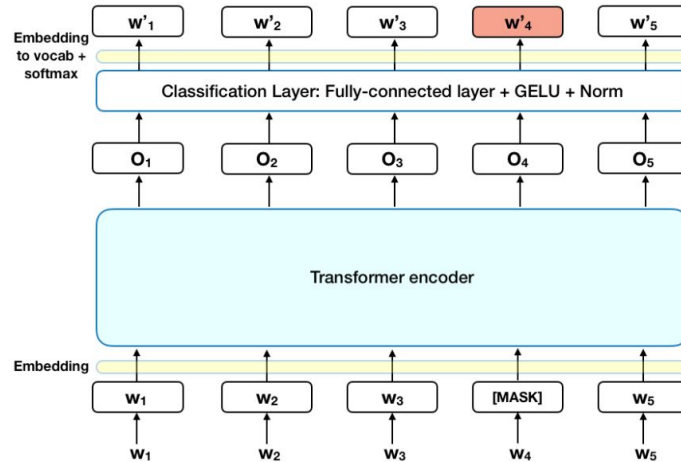
- Purpose
 - Applies non-linear transformations to the output of the attention layer
- Equation
 - $\text{FFN}(x) = \max(0, xW_1 + b_1)W_2 + b_2$
 - where W and b are learned weights and biases
- These FFN is applied separately to each position

Transformer Encoder



BERT (Bidirectional Encoder Representations from Transformers)

- Bidirectional Context
- Pre-trained on the language, and then fine-tuned



Self-Supervised Learning

- Masked Language Modelling

- Mask out k% of the input words, and then predict the masked words
- The man went to the [MASK] to [MASK] a gallon of milk
- Multi-class classification problem

- Next sentence prediction

- To learn relationships between sentences, predict whether Sentence B is actual sentence that proceeds Sentence A, or a random sentence

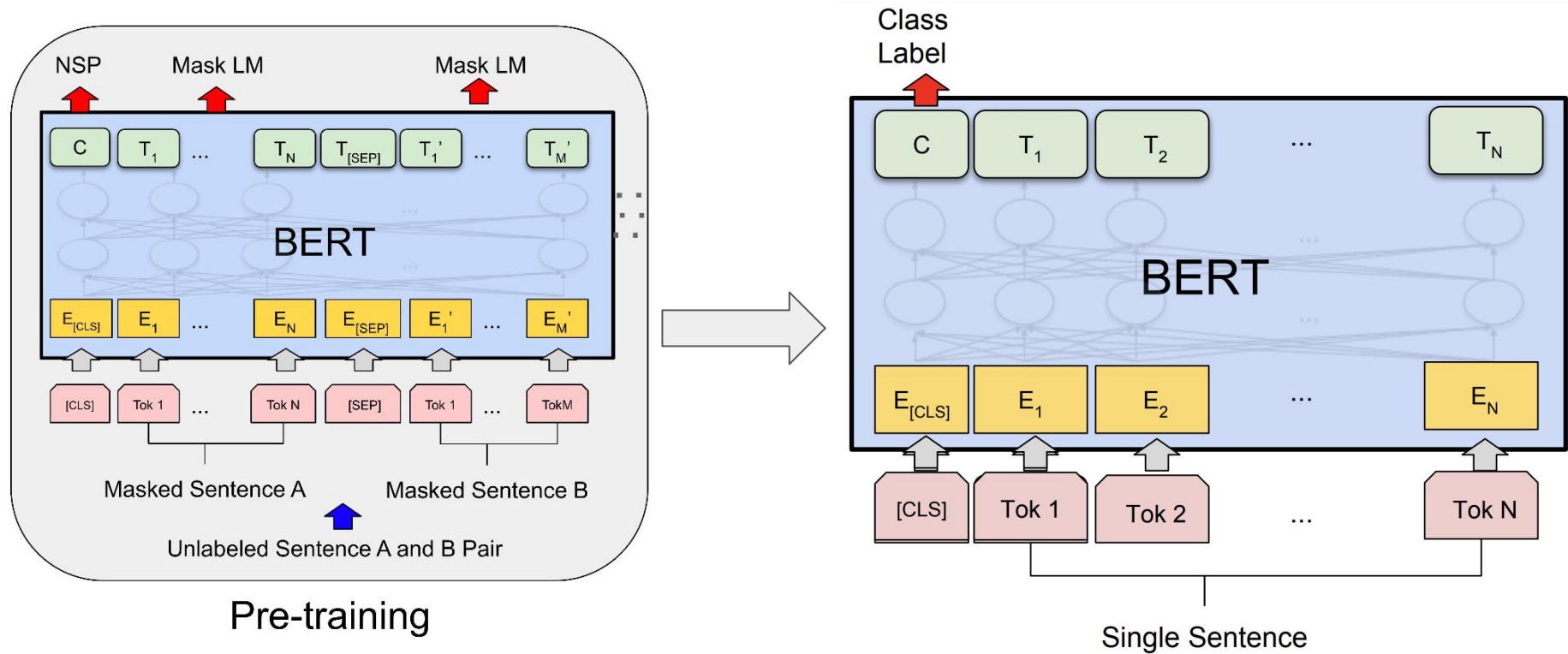
Sentence A = The man went to the store.
Sentence B = He bought a gallon of milk.
Label = IsNextSentence

Sentence A = The man went to the store.
Sentence B = Penguins are flightless.
Label = NotNextSentence

Masked Language Modeling

- Masked language modeling is really hard!
 - The man went to the [MASK] to [MASK] a gallon of milk
- World knowledge:
 - The capital of Paris is [MASK]
- Sentiment classification:
 - A sentimental mess that never rings true. All in all, the movie was [MASK].

Pre-training to Fine-tuning Pipeline



Works Really Well!

System	MNLI-(m/mm) 392k	QQP 363k	QNLI 108k	SST-2 67k	CoLA 8.5k	STS-B 5.7k	MRPC 3.5k	RTE 2.5k	Average
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
BERT _{BASE}	84.6/83.4	71.2	90.5	93.5	52.1	85.8	88.9	66.4	79.6
BERT _{LARGE}	86.7/85.9	72.1	92.7	94.9	60.5	86.5	89.3	70.1	82.1

Pre-Train then Fine-tune Paradigm

- Pre-train a large model on a lot of data with some self-supervised objective.
 - E.g. Masked language modeling
- Fine-tune on (smaller) downstream datasets
- Benefits:
 - Transfer learning: Leveraging knowledge from self-supervised objective to improve performance
 - Avoids overfitting: Large models would overfit if trained from scratch

Fine-tuning Limitations

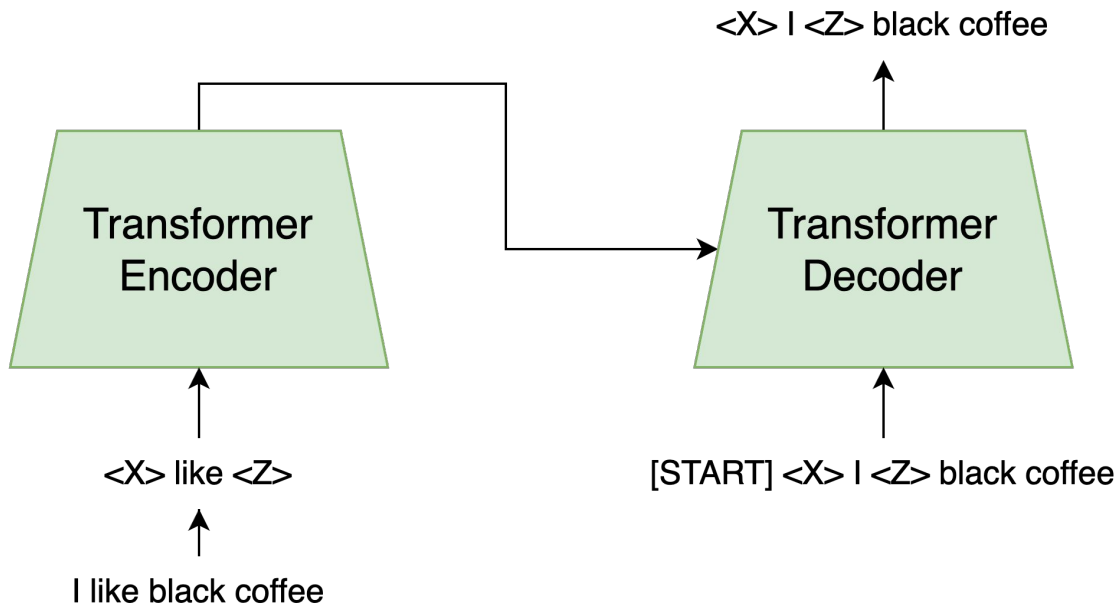
- Memory overhead
 - Training takes more memory than inference
 - Fine-tuning large models requires expensive GPUs
- Optimization
 - Might need to adjust optimization hyperparameters for each downstream task
 - Learning rates, batch sizes, etc.

Question:

- What are some limitations of BERT?
 - Think about the different kinds of language tasks you might want to perform

T5: Text-to-Text Transfer Transformer

- Pre-trained Encoder Decoder Language Model
 - Can generate text!



T5 Span Corruption Objective

- Very similar to masked language modeling
- Key differences:
 - Mask spans with single “sentinel” tokens
 - Generate the masked text with a decoder

Original text

Thank you ~~for inviting~~ me to your party ~~last~~ week.

Inputs

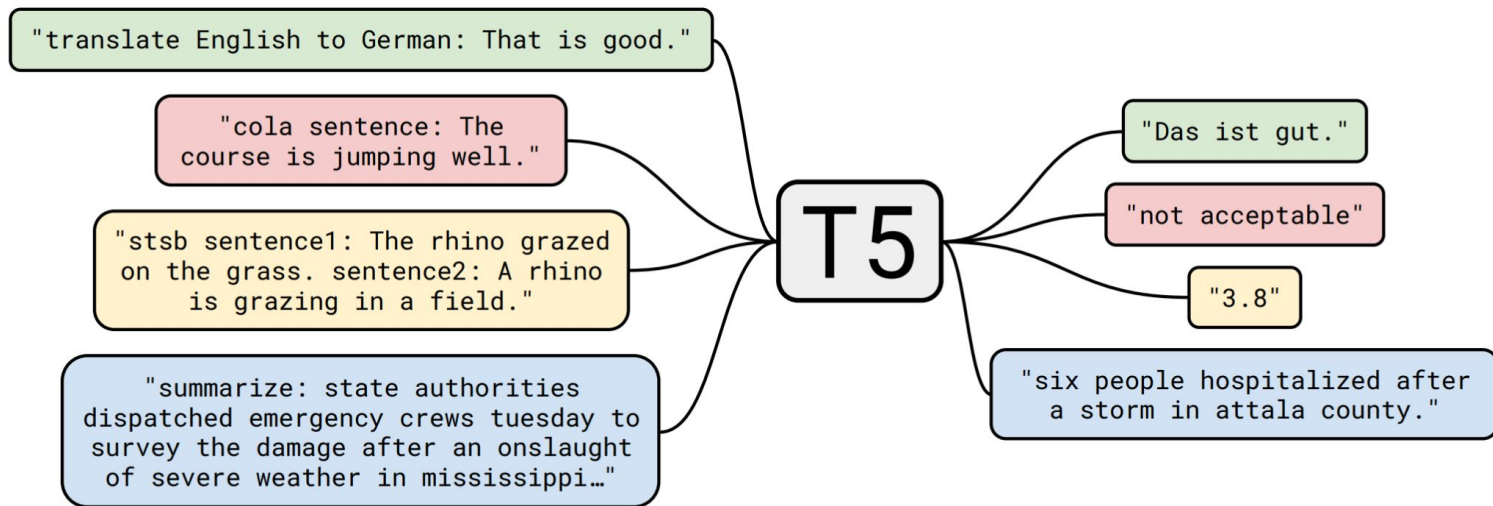
Thank you <X> me to your party <Y> week.

Targets

<X> for inviting <Y> last <Z>

T5 Fine-Tuning

- Cast every task as a language generation problem
 - Even classification!



Benchmark Results

- Outperformed best encoder-only models

Model	GLUE Average	CoLA Matthew's	SST-2 Accuracy	MRPC F1	MRPC Accuracy	STS-B Pearson	STS-B Spearman
Previous best	89.4 ^a	69.2 ^b	97.1 ^a	93.6^b	91.5^b	92.7 ^b	92.3 ^b
T5-Small	77.4	41.0	91.8	89.7	86.6	85.6	85.0
T5-Base	82.7	51.1	95.2	90.7	87.5	89.4	88.6
T5-Large	86.4	61.2	96.3	92.4	89.9	89.9	89.2
T5-3B	88.5	67.1	97.4	92.5	90.0	90.6	89.8
T5-11B	90.3	71.6	97.5	92.8	90.4	93.1	92.8

Model	QQP F1	QQP Accuracy	MNLI-m Accuracy	MNLI-mm Accuracy	QNLI Accuracy	RTE Accuracy	WNLI Accuracy
Previous best	74.8 ^c	90.7^b	91.3 ^a	91.0 ^a	99.2^a	89.2 ^a	91.8 ^a
T5-Small	70.0	88.0	82.4	82.3	90.3	69.9	69.2
T5-Base	72.6	89.4	87.1	86.2	93.7	80.1	78.8
T5-Large	73.9	89.9	89.9	89.6	94.8	87.2	85.6
T5-3B	74.4	89.7	91.4	91.2	96.3	91.1	89.7
T5-11B	75.1	90.6	92.2	91.9	96.9	92.8	94.5

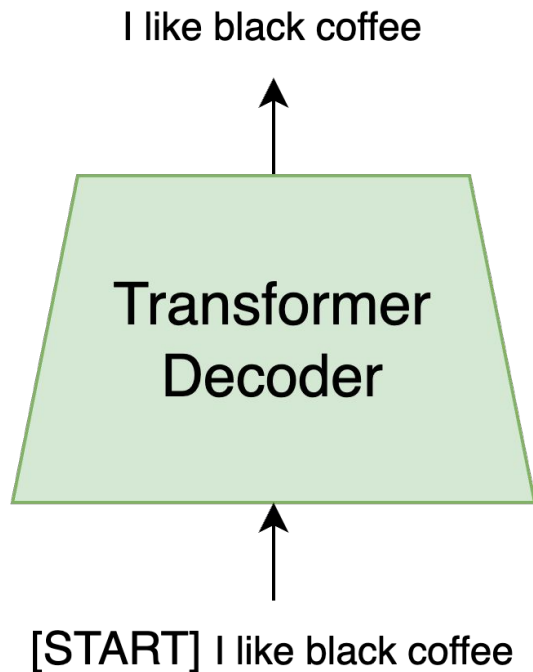
T5 Ablations- Dataset Size

- More data helps!

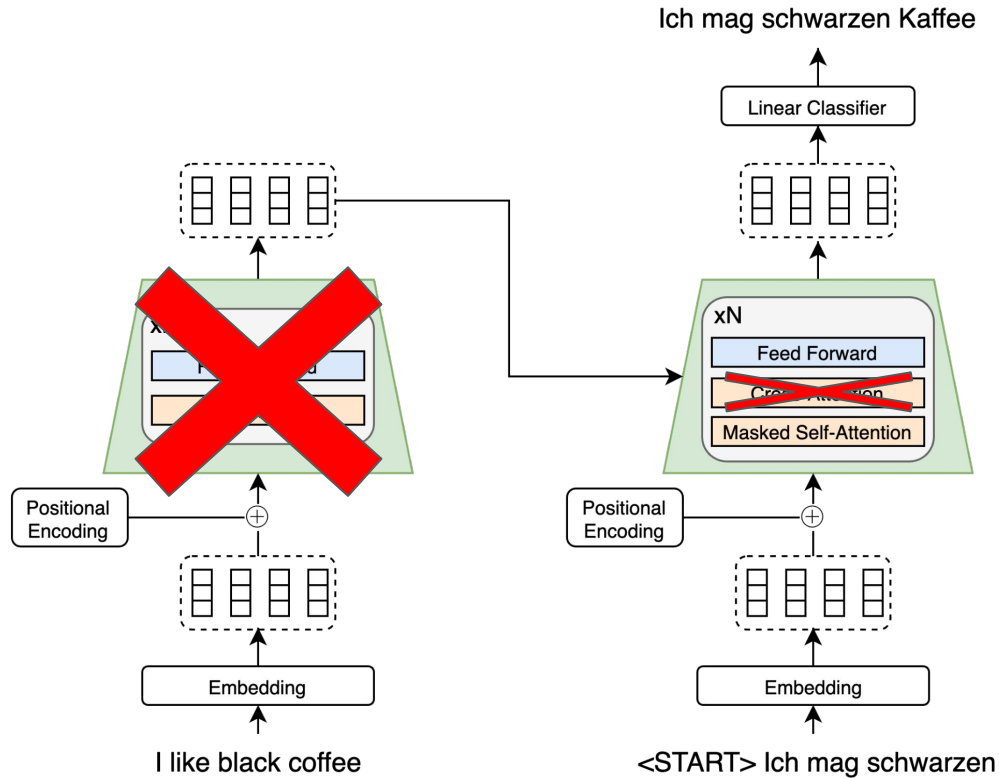
Number of tokens	Repeats	GLUE	CNNDM	SQuAD	SGLUE	EnDe	EnFr	EnRo
★ Full data set	0	83.28	19.24	80.88	71.36	26.98	39.82	27.65
2^{29}	64	82.87	19.19	80.97	72.03	26.83	39.74	27.63
2^{27}	256	82.62	19.20	79.78	69.97	27.02	39.71	27.33
2^{25}	1,024	79.55	18.57	76.27	64.76	26.38	39.56	26.80
2^{23}	4,096	76.34	18.33	70.92	59.29	26.37	38.84	25.81

GPT

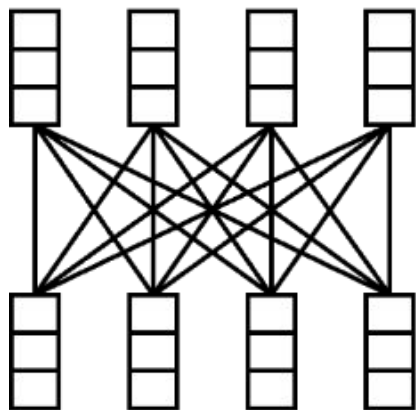
- Only use the transformer decoder
 - Train on next word prediction
 - “The capital of France is _”
- Easy to apply in streaming settings
 - E.g. chatbots



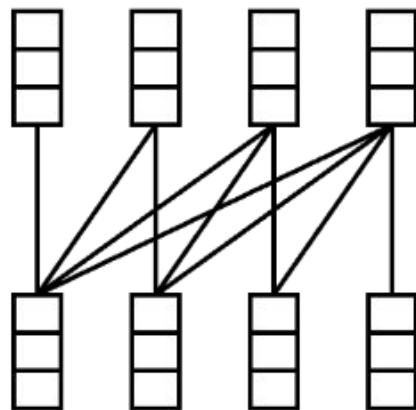
Decoder-Only Architecture



Review: Masked Self-Attention



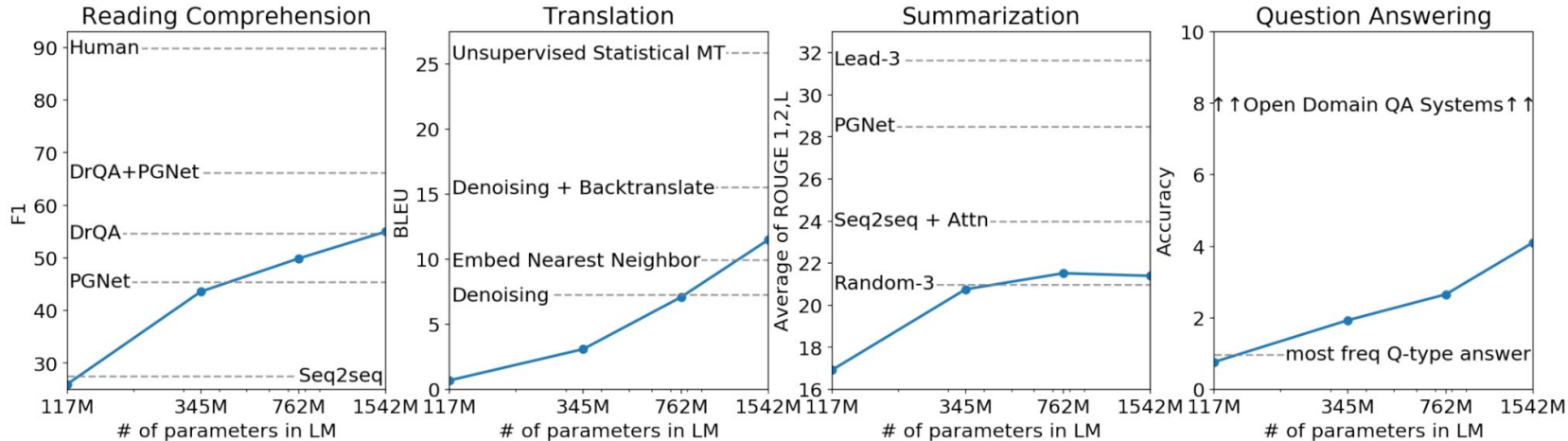
Self-Attention



Masked Self-Attention

GPT-2

- Train self-supervised auto-regressive LMs on web text
 - Next-word prediction
 - Up to 1.5 billion parameters
- Observed non-trivial zero-shot performance



“Language Models are Unsupervised Multitask Learners”

GPT-2 Zero-Shot Capabilities

- Question-answering without any fine-tuning
 - Formatting is important!

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

GPT-2

- Naturally occurring translation demonstrations in the pre-training corpus

”I’m not the cleverest man in the world, but like they say in French: **Je ne suis pas un imbecile** [I’m not a fool].

In a now-deleted post from Aug. 16, Soheil Eid, Tory candidate in the riding of Joliette, wrote in French: **”Mentez mentez, il en restera toujours quelque chose,”** which translates as, **”Lie lie and something will always remain.”**

“I hate the word ‘**perfume**,’” Burr says. ‘It’s somewhat better in French: ‘**parfum**.’

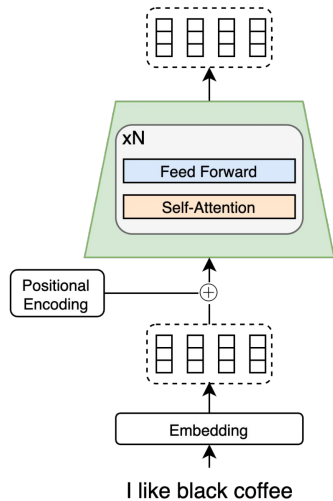
If listened carefully at 29:55, a conversation can be heard between two guys in French: **“-Comment on fait pour aller de l’autre coté? -Quel autre coté?”**, which means **“- How do you get to the other side? - What side?”**.

If this sounds like a bit of a stretch, consider this question in French: **As-tu aller au cinéma?**, or **Did you go to the movies?**, which literally translates as **Have-you to go to movies/theater?**

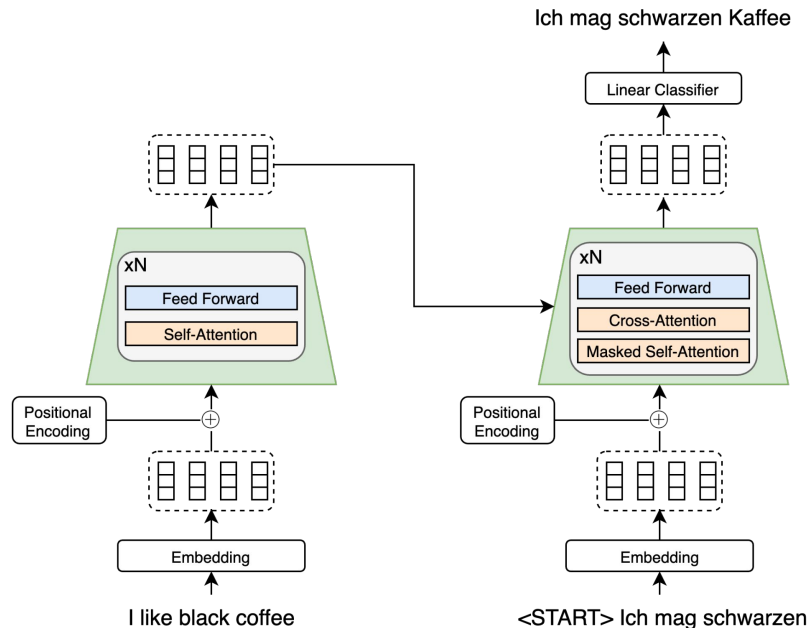
“Brevet Sans Garantie Du Gouvernement”, translated to English: **“Patented without government warranty”**.

Table 1. Examples of naturally occurring demonstrations of English to French and French to English translation found throughout the WebText training set.

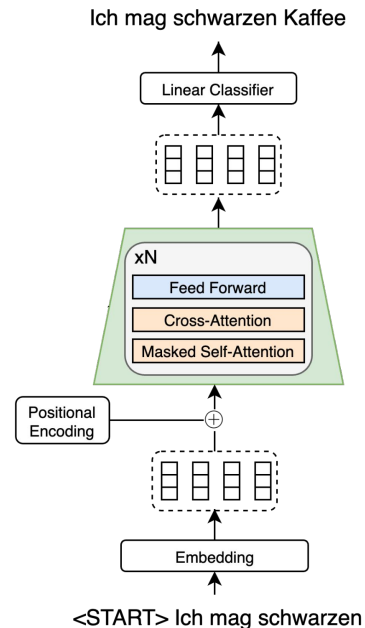
BERT



T5/BART

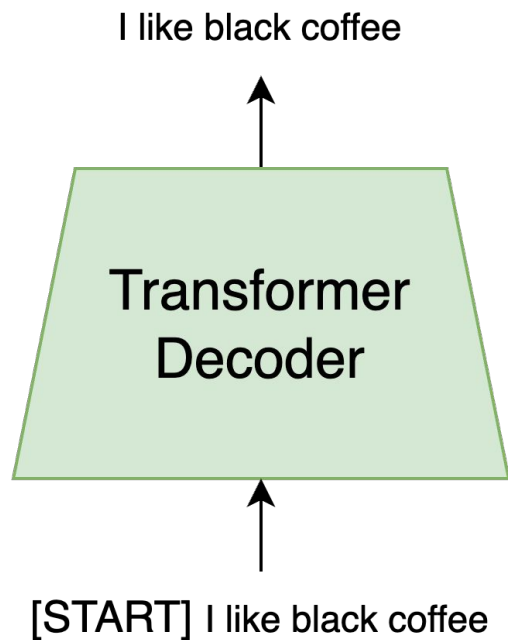
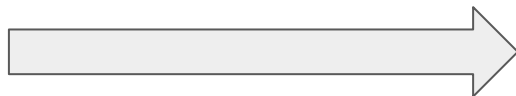
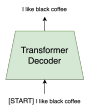


GPT



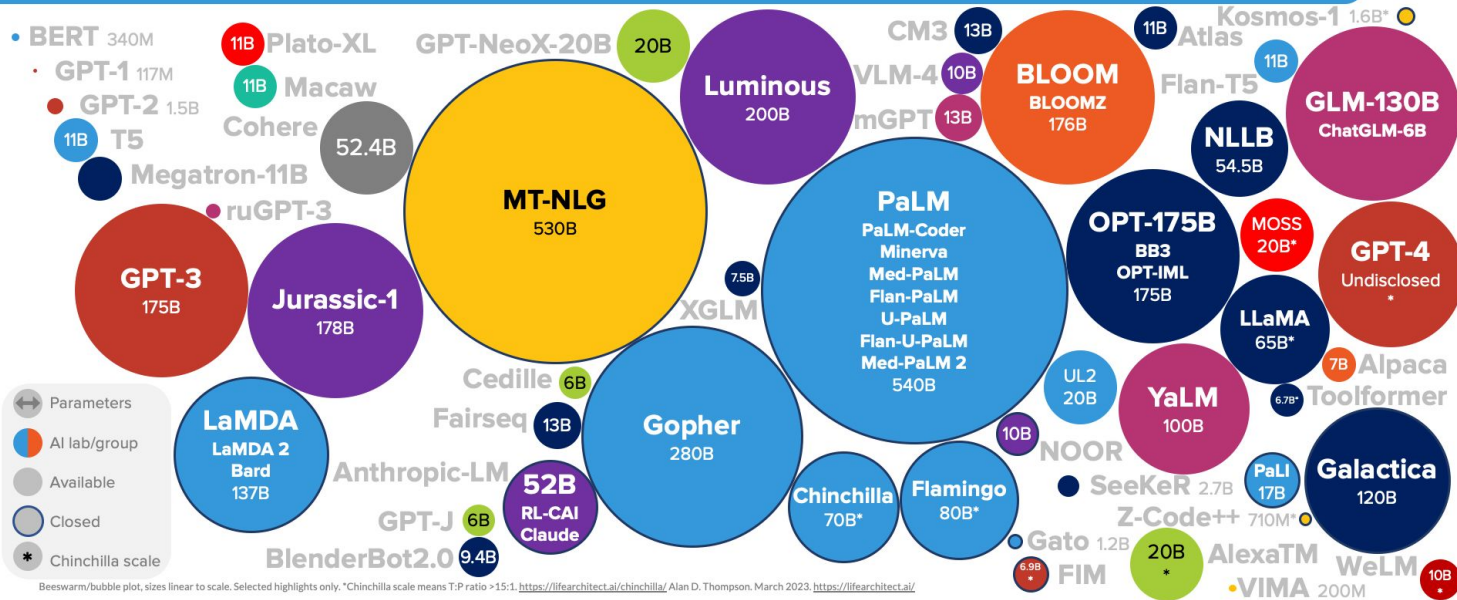
GPT-3

- Train a bigger version of GPT-2 on more data



GPT-3

LANGUAGE MODEL SIZES TO MAR/2023



Beeswarm/bubble plot, sizes linear to scale. Selected highlights only. *Chinchilla scale means T:P ratio >15:1. <https://liferichitect.ai/chinchilla/> Alan D. Thompson, March 2023. <https://liferichitect.ai/>



GPT-3

Model Name	n_{params}	n_{layers}	d_{model}	n_{heads}	d_{head}	Batch Size	Learning Rate
GPT-3 Small	125M	12	768	12	64	0.5M	6.0×10^{-4}
GPT-3 Medium	350M	24	1024	16	64	0.5M	3.0×10^{-4}
GPT-3 Large	760M	24	1536	16	96	0.5M	2.5×10^{-4}
GPT-3 XL	1.3B	24	2048	24	128	1M	2.0×10^{-4}
GPT-3 2.7B	2.7B	32	2560	32	80	1M	1.6×10^{-4}
GPT-3 6.7B	6.7B	32	4096	32	128	2M	1.2×10^{-4}
GPT-3 13B	13.0B	40	5140	40	128	2M	1.0×10^{-4}
GPT-3 175B or “GPT-3”	175.0B	96	12288	96	128	3.2M	0.6×10^{-4}

Table 2.1: Sizes, architectures, and learning hyper-parameters (batch size in tokens and learning rate) of the models which we trained. All models were trained for a total of 300 billion tokens.

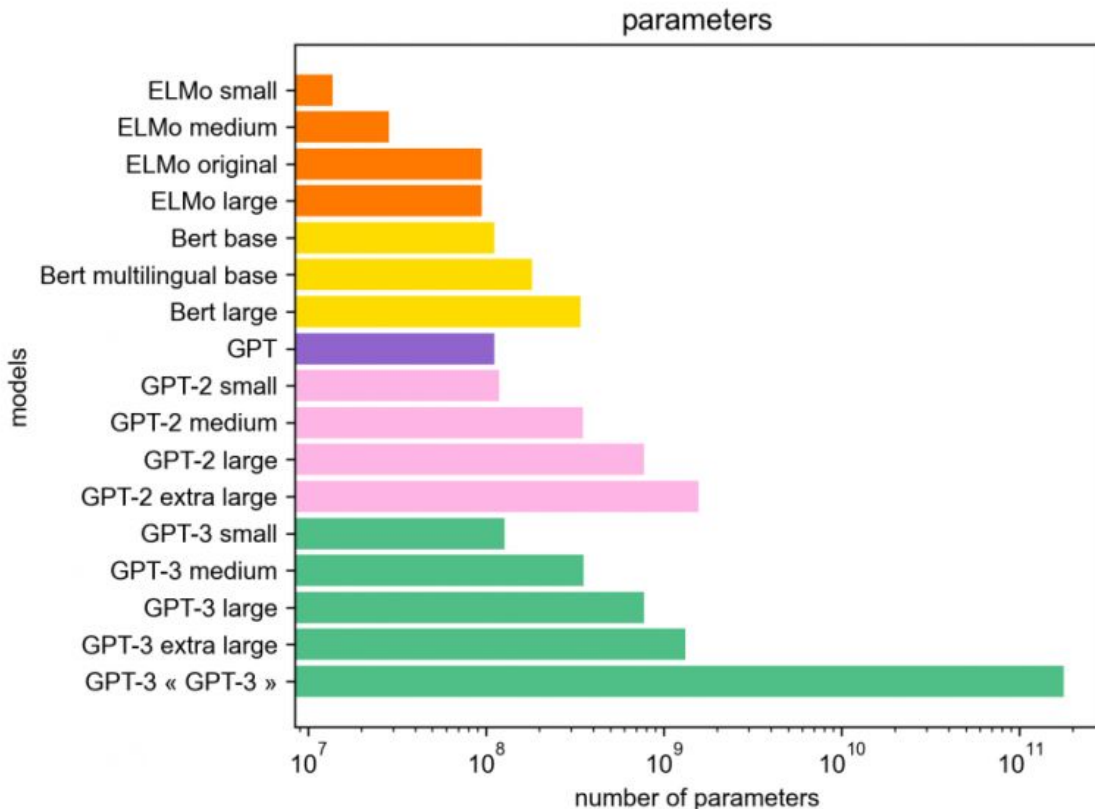
Discuss

- How many parameters in a single feedforward layer of GPT-3
 - An MLP with:
 - Input dimension of 12,228
 - One hidden layer with dimension $4 \times 12,228$
 - Output dimension of 12,228

Model Name	n_{params}	n_{layers}	d_{model}	n_{heads}	d_{head}	Batch Size	Learning Rate
GPT-3 Small	125M	12	768	12	64	0.5M	6.0×10^{-4}
GPT-3 Medium	350M	24	1024	16	64	0.5M	3.0×10^{-4}
GPT-3 Large	760M	24	1536	16	96	0.5M	2.5×10^{-4}
GPT-3 XL	1.3B	24	2048	24	128	1M	2.0×10^{-4}
GPT-3 2.7B	2.7B	32	2560	32	80	1M	1.6×10^{-4}
GPT-3 6.7B	6.7B	32	4096	32	128	2M	1.2×10^{-4}
GPT-3 13B	13.0B	40	5140	40	128	2M	1.0×10^{-4}
GPT-3 175B or “GPT-3”	175.0B	96	12288	96	128	3.2M	0.6×10^{-4}

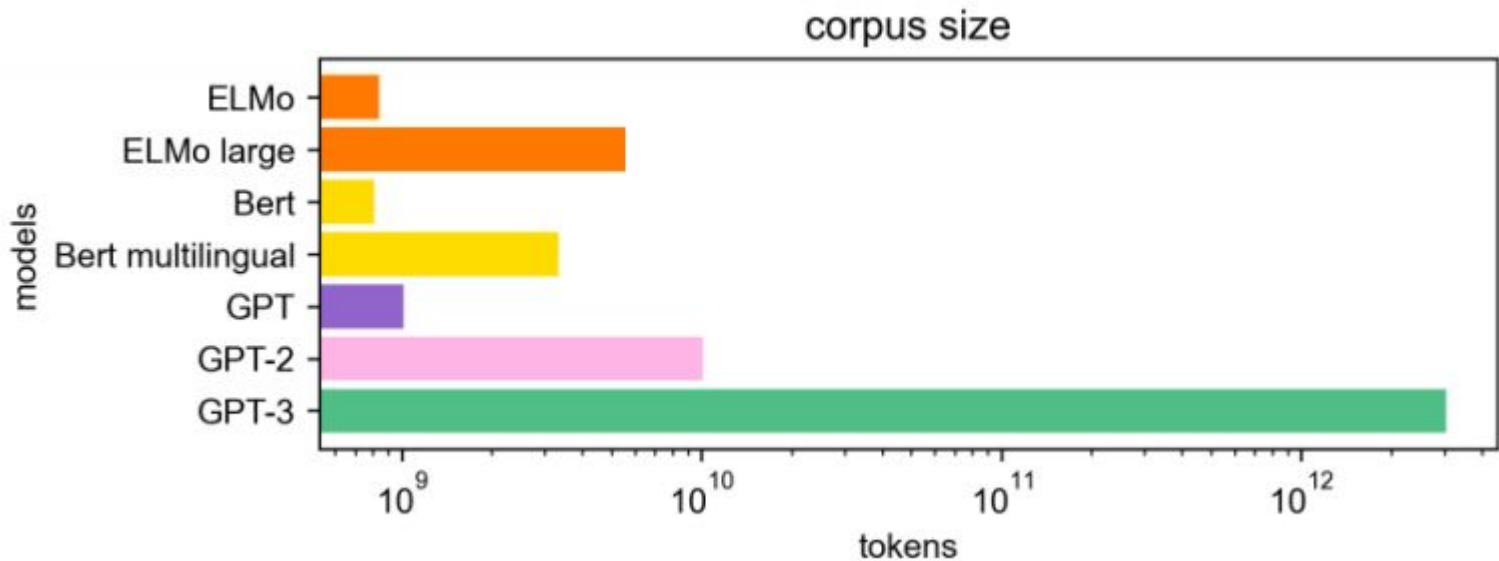
Table 2.1: Sizes, architectures, and learning hyper-parameters (batch size in tokens and learning rate) of the models which we trained. All models were trained for a total of 300 billion tokens.

Scaling Model Size



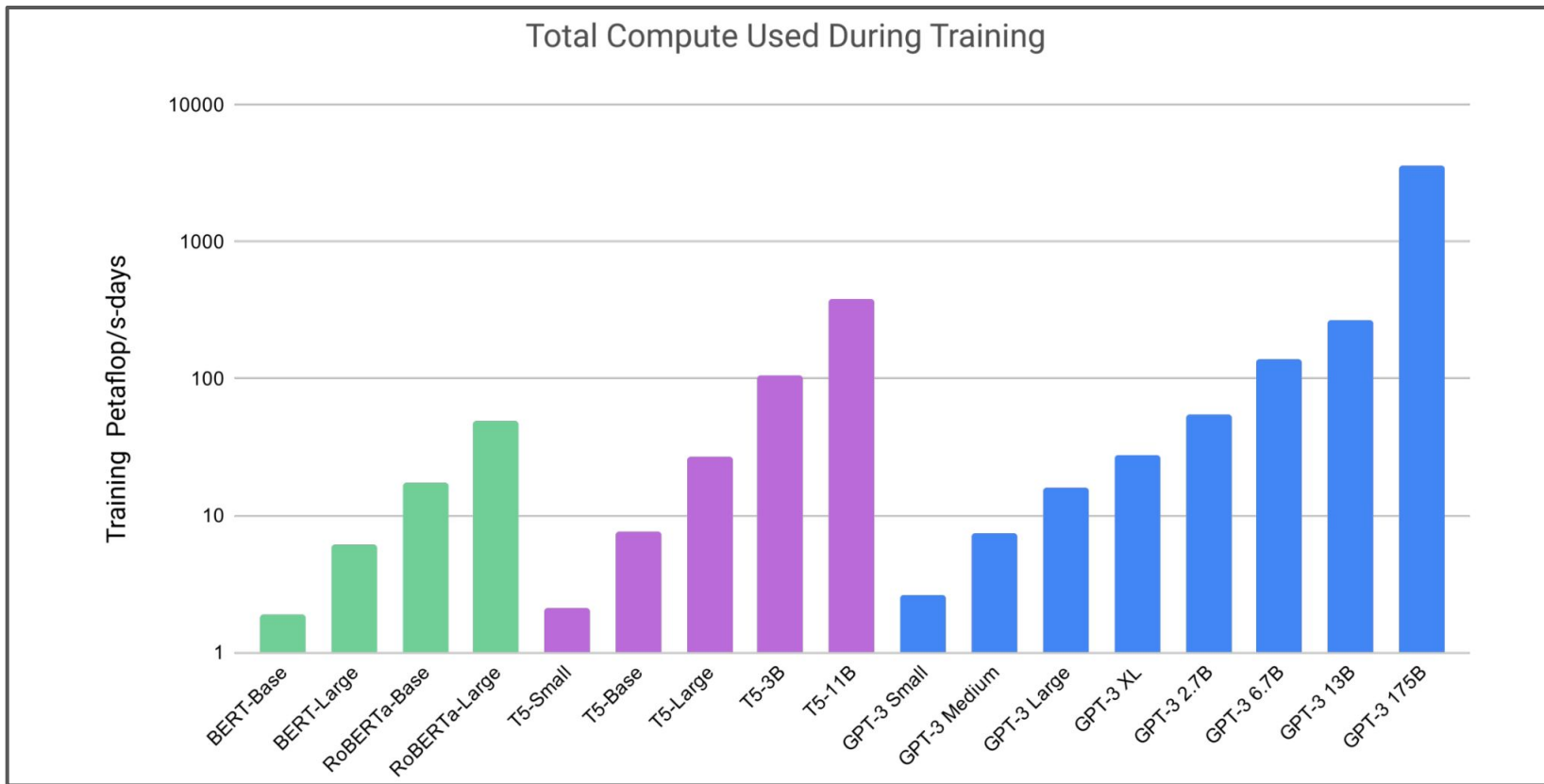
Source: <https://hellofuture.orange.com/en/the-gpt-3-language-model-revolution-or-evolution/>

Scaling Training Data



Source: <https://hellofuture.orange.com/en/the-gpt-3-language-model-revolution-or-evolution/>

Scaling Training Compute



Effect of Scaling on Validation Loss

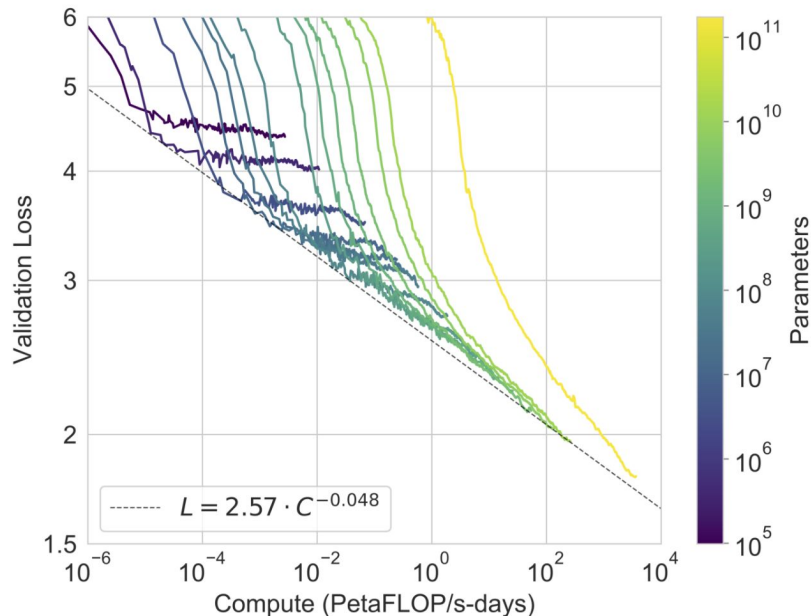


Figure 3.1: Smooth scaling of performance with compute. Performance (measured in terms of cross-entropy validation loss) follows a power-law trend with the amount of compute used for training. The power-law behavior observed in [KMH⁺20] continues for an additional two orders of magnitude with only small deviations from the predicted curve. For this figure, we exclude embedding parameters from compute and parameter counts.

Power Laws

$$C = C_0 N D$$

$$L = \frac{A}{N^\alpha} + \frac{B}{D^\beta} + L_0$$

- Performance improves predictably with increased compute, data, and parameters
 - Can actually fit power laws!
 - Predict performance before training!

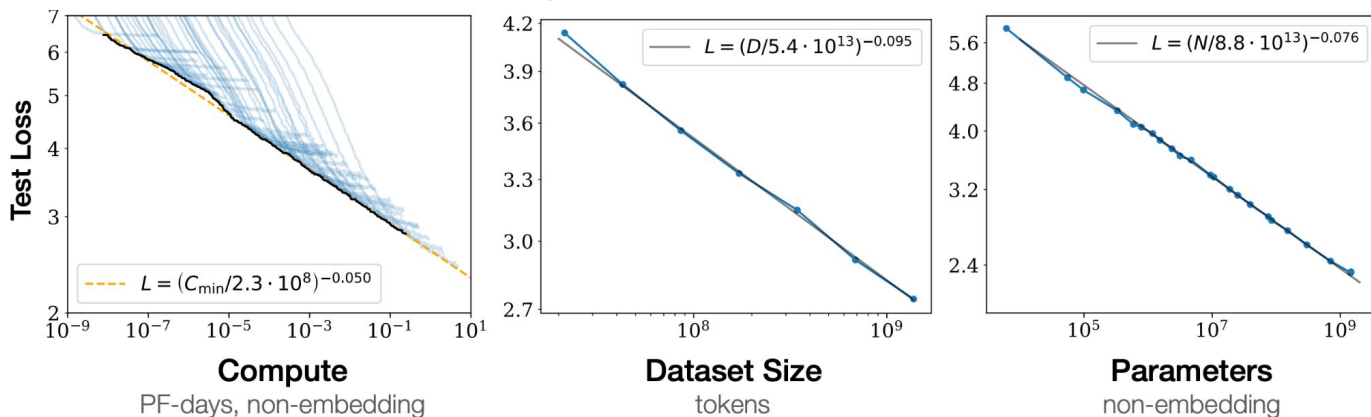


Figure 1 Language modeling performance improves smoothly as we increase the model size, dataset size, and amount of compute² used for training. For optimal performance all three factors must be scaled up in tandem. Empirical performance has a power-law relationship with each individual factor when not bottlenecked by the other two.

GPT-3 Evaluation

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.

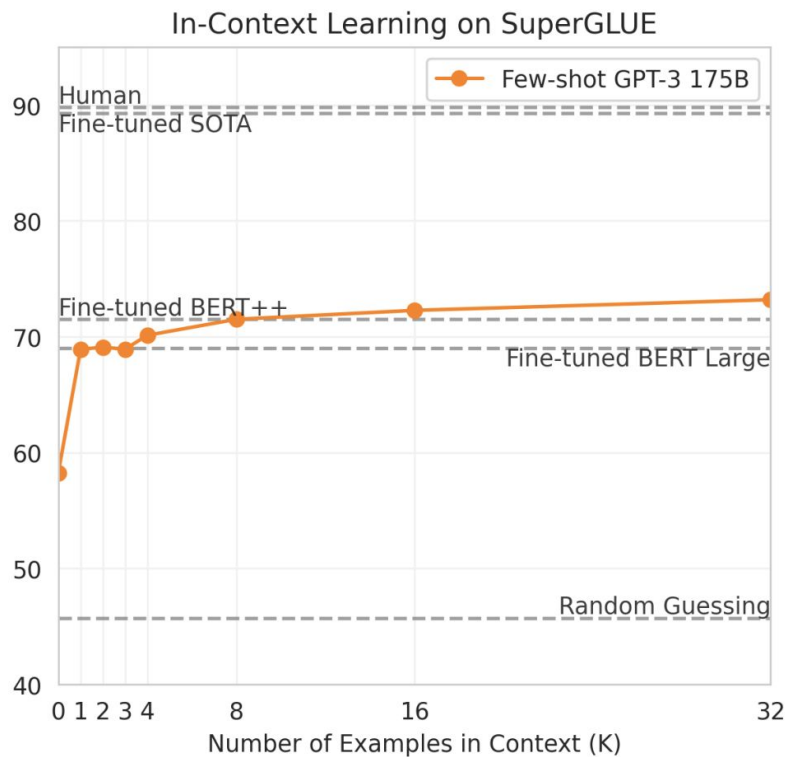
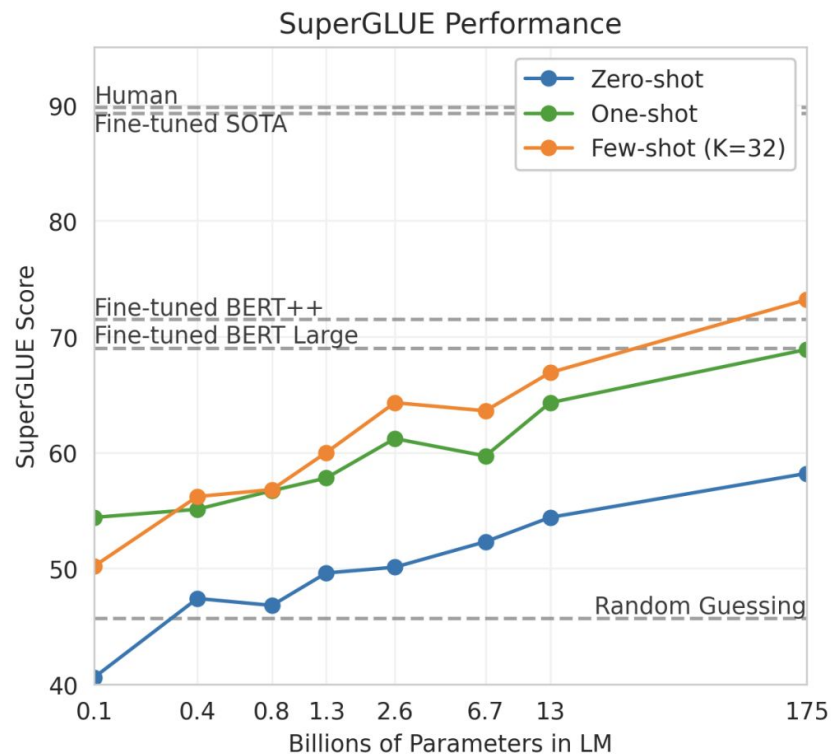
```
1 Translate English to French: ← task description
2 sea otter => loutre de mer ← examples
3 peppermint => menthe poivrée ←
4 plush girafe => girafe peluche ←
5 cheese => ..... ← prompt
```

Fine-tuning

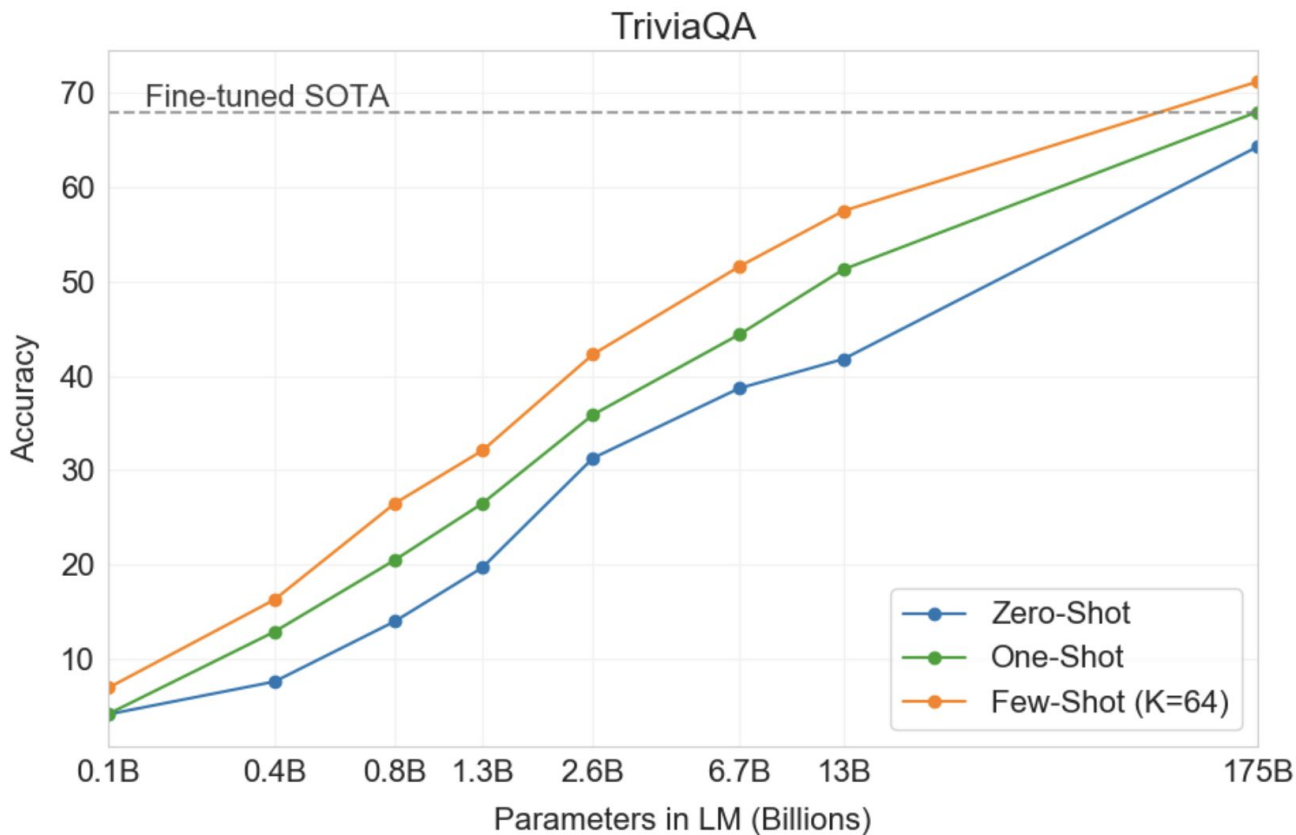
The model is trained via repeated gradient updates using a large corpus of example tasks.



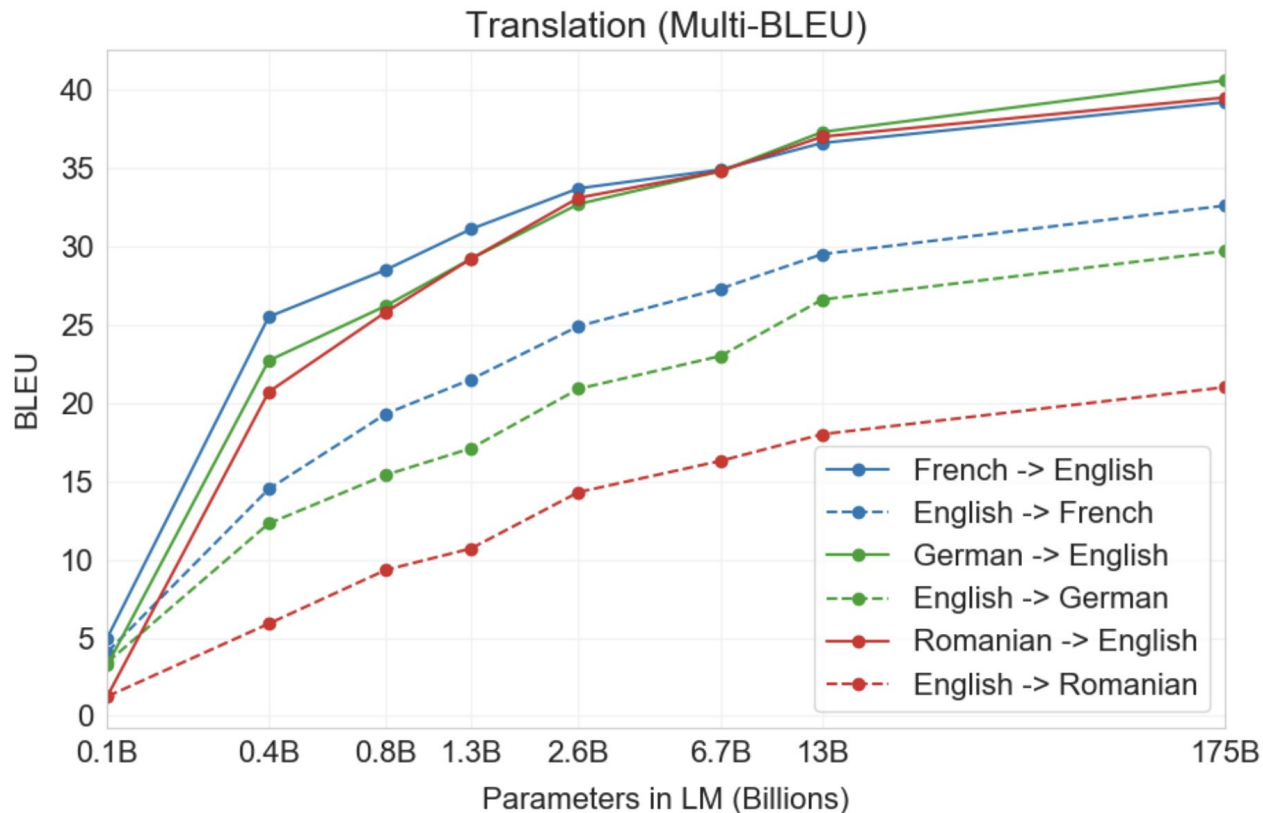
GPT-3 Evaluation



GPT-3 Evaluation



GPT-3 Evaluation



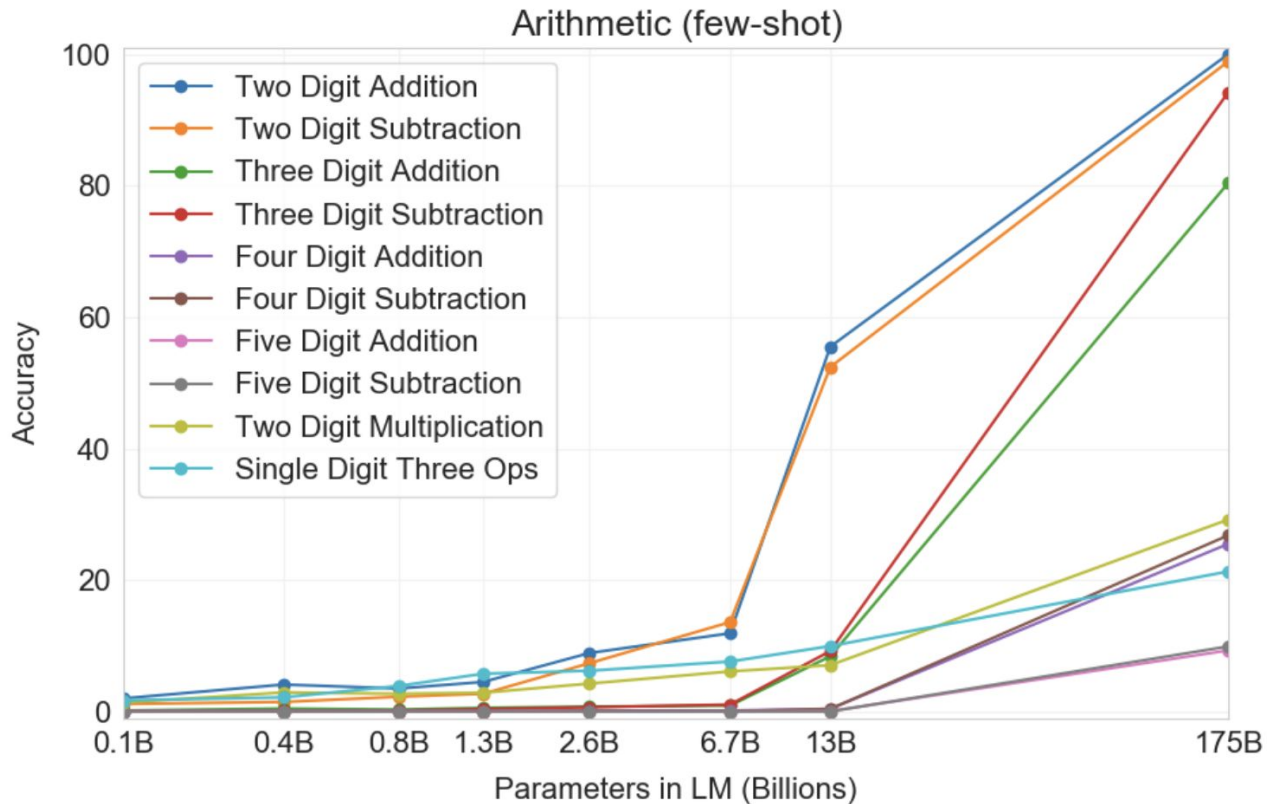
GPT-3 News Articles

- Humans struggle to identify GPT-3 generated news articles

	Mean accuracy	95% Confidence Interval (low, hi)	t compared to control (p -value)	“I don’t know” assignments
Control (deliberately bad model)	86%	83%–90%	-	3.6 %
GPT-3 Small	76%	72%–80%	3.9 ($2e-4$)	4.9%
GPT-3 Medium	61%	58%–65%	10.3 ($7e-21$)	6.0%
GPT-3 Large	68%	64%–72%	7.3 ($3e-11$)	8.7%
GPT-3 XL	62%	59%–65%	10.7 ($1e-19$)	7.5%
GPT-3 2.7B	62%	58%–65%	10.4 ($5e-19$)	7.1%
GPT-3 6.7B	60%	56%–63%	11.2 ($3e-21$)	6.2%
GPT-3 13B	55%	52%–58%	15.3 ($1e-32$)	7.1%
GPT-3 175B	52%	49%–54%	16.9 ($1e-34$)	7.8%

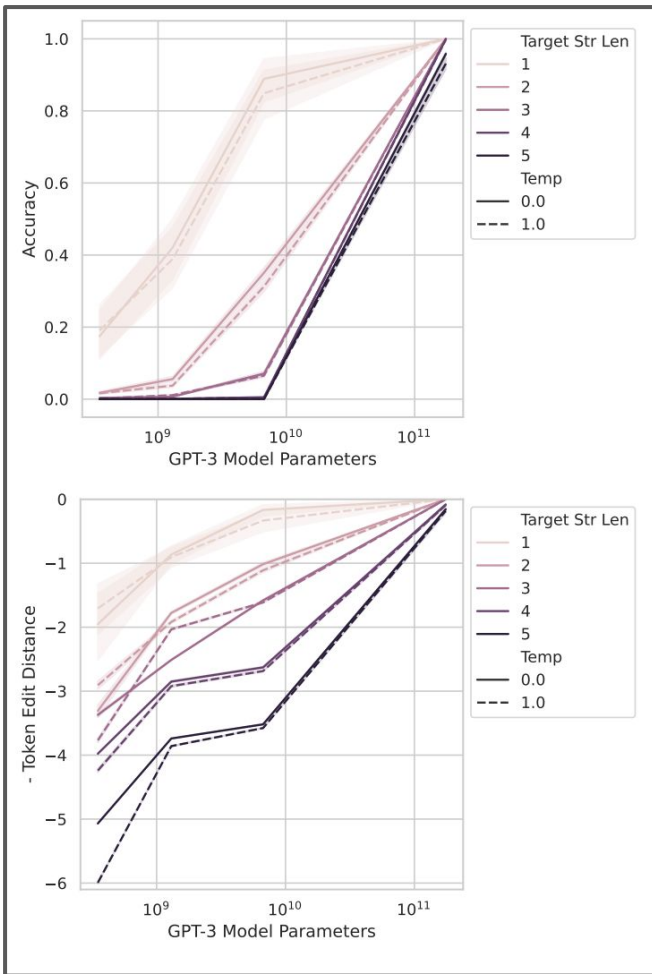
Table 3.11: Human accuracy in identifying whether short (~200 word) news articles are model generated.

Emergent Capabilities



Is Emergence a Mirage?

- Look at four digit addition under different metrics
 - Emergence can be an artefact of the evaluation metric
- Addition capabilities improve smoothly when using a more granular metric
 - Exact match accuracy -> Token edit distance



Very Versatile

Explaining a Joke

I will explain these jokes:

(1) The problem with kleptomaniacs is that they always take things literally.

Explanation: This joke is wordplay. Someone who "takes things literally" is someone who doesn't fully understand social cues and context, which is a negative trait. But the definition of kleptomania is someone who literally takes things.

(2) Always borrow money from a pessimist. They'll never expect it back.

Explanation: Most people expect you to pay them back when you borrow money, however a pessimist is someone who always assumes the worst, so if you borrow money from them, they will expect that you won't pay them back anyways.

Input: I was supposed to start writing the paper at 5:00 PM. But then I started playing with this cool new language model for 10 minutes. 10 minutes later, it's suddenly 9:30 PM!

Model Output: This joke is about how time flies when you're having fun. The person was supposed to start writing a paper at 5:00 PM, but they started playing with a language model instead. They thought they were only playing with it for 10 minutes, but it was actually 4.5 hours.

Input: I think I understand why me and my partner are such a good match. I get excited when I smell food and become super alert when the doorbell rings. My girlfriend Shelley becomes ecstatic when she sees a person who she just saw 5 hours ago.

Model Output: This joke is about dogs. Dogs get excited when they smell food and when the doorbell rings. The girlfriend is excited to see someone she just saw 5 hours ago, which is a trait of dogs.

Any problems with auto-regressive LMs?

Prompt:

What is the purpose of the list C in the code below?

```
def binomial_coefficient(n, r):  
    C = [0 for i in range(r + 1)];  
    C[0] = 1;  
    for i in range(1, n + 1):  
        j = min(i, r);  
        while j > 0:  
            C[j] += C[j - 1];  
            j -= 1;  
    return C[r]
```

Any problems with auto-regressive LMs?

Prompt:

What is the purpose of the list C in the code below?

```
def binomial_coefficient(n, r):  
    C = [0 for i in range(r + 1)];  
    C[0] = 1;  
    for i in range(1, n + 1):  
        j = min(i, r);  
        while j > 0:  
            C[j] += C[j - 1];  
            j -= 1;  
    return C[r]
```

GPT-3 175B completion:

- A. to store the value of C[0]
 - B. to store the value of C[1]
 - C. to store the value of C[i]
 - D. to store the value of C[i - 1]
-

Any problems with auto-regressive LMs?

- Just trained on next-word prediction
 - Won't follow instructions!

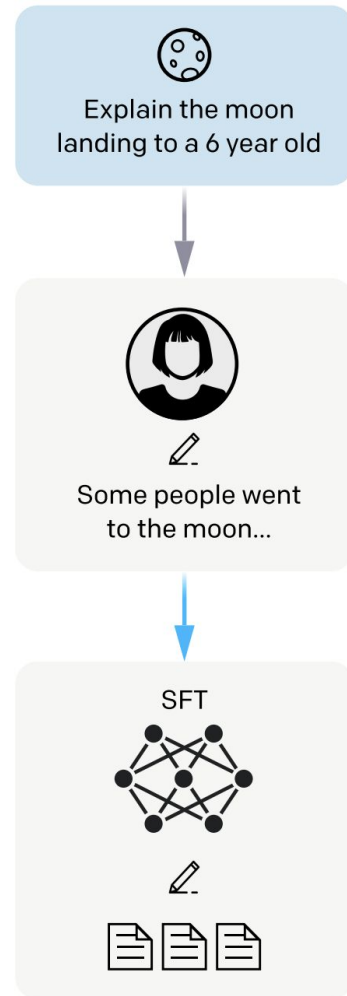
Instruction Tuning

- Collect a small dataset of instruction following demonstrations
 - Fine-tune the pre-trained LM

A prompt is sampled from our prompt dataset.

A labeler demonstrates the desired output behavior.

This data is used to fine-tune GPT-3 with supervised learning.



Instruction Tuning

- Aligns LM with user intent!

Prompt:

What is the purpose of the list C in the code below?

```
def binomial_coefficient(n, r):  
    C = [0 for i in range(r + 1)];  
    C[0] = 1;  
    for i in range(1, n + 1):  
        j = min(i, r);  
        while j > 0:  
            C[j] += C[j - 1];  
            j -= 1;  
    return C[r]
```

GPT-3 175B completion:

- A. to store the value of C[0]
- B. to store the value of C[1]
- C. to store the value of C[i]
- D. to store the value of C[i - 1]

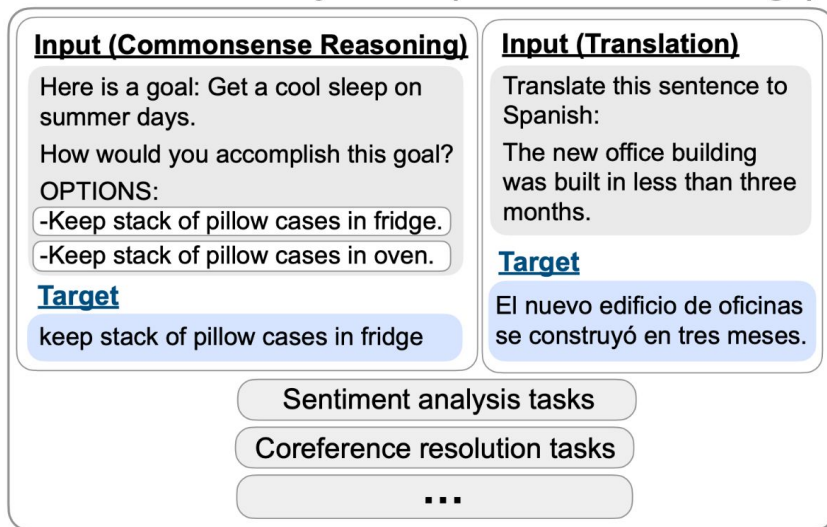
InstructGPT 175B completion:

The list C in this code is used to store the values of the binomial coefficient as the function iterates through the values of n and r. It is used to calculate the value of the binomial coefficient for a given value of n and r, and stores the result in the final return value of the function.

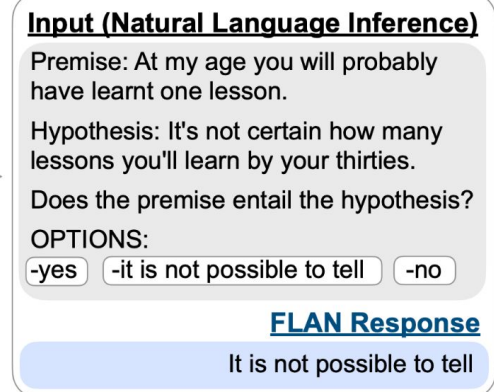
Instruction Tuning

- Model learns to generalize to unseen instructions at inference-time

Finetune on many tasks (“instruction-tuning”)



Inference on unseen task type



Prompt Engineering

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. ✅

Chain of Thought Prompting

- Break down problem into subparts
 - Learns to do it with few-shot in-context learning

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.

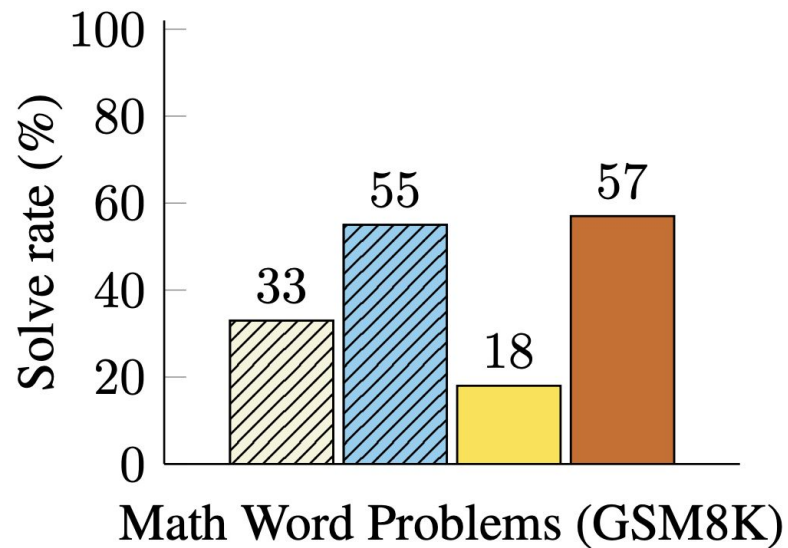
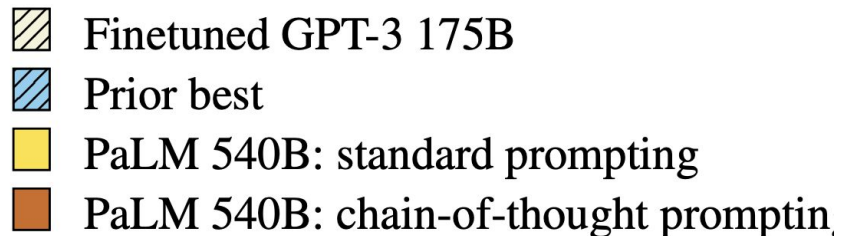
Coin Flip (state tracking)

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.

Chain of Thought Prompting

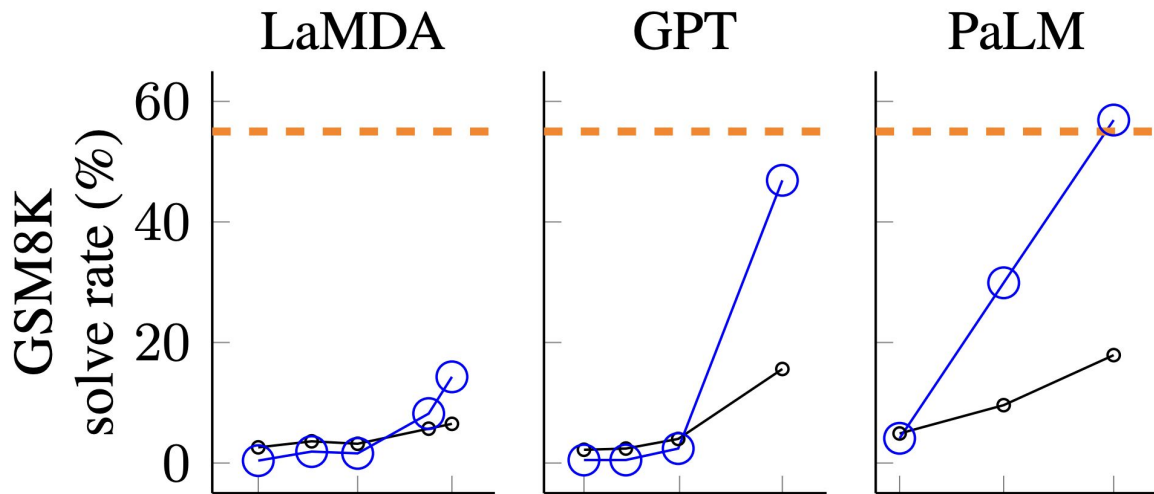
- Can significantly improve performance



Chain of Thought Prompting

- No benefit at small scales

- Standard prompting
- Chain-of-thought prompting
- - - Prior supervised best



Sensitivity to the Prompt

- Performance depends a lot on the prompt!

Table 4: Robustness study against template measured on the MultiArith dataset with text-davinci-002. (*1) This template is used in Ahn et al. [2022] where a language model is prompted to generate step-by-step actions given a high-level instruction for controlling robotic actions. (*2) This template is used in Reynolds and McDonnell [2021] but is not quantitatively evaluated.

No.	Category	Template	Accuracy
1	instructive	Let's think step by step.	78.7
2		First, (*1)	77.3
3		Let's think about this logically.	74.5
4		Let's solve this problem by splitting it into steps. (*2)	72.2
5		Let's be realistic and think step by step.	70.8
6		Let's think like a detective step by step.	70.3
7		Let's think	57.5
8		Before we dive into the answer,	55.7
9		The answer is after the proof.	45.7
10	misleading	Don't think. Just feel.	18.8
11		Let's think step by step but reach an incorrect answer.	18.7
12		Let's count the number of "a" in the question.	16.7
13		By using the fact that the earth is round,	9.3
14	irrelevant	By the way, I found a good restaurant nearby.	17.5
15		AbraKadabra!	15.5
16		It's a beautiful day.	13.1
-		(Zero-shot)	17.7

Toolformer

- LLMs can be designed to use tools like calculators, QA systems and calendars to produce better results
- LLM are trained to generate and issue queries to external tools

The New England Journal of Medicine is a registered trademark of [QA("Who is the publisher of The New England Journal of Medicine?") → Massachusetts Medical Society] the MMS.

Out of 1400 participants, 400 (or [Calculator(400 / 1400) → 0.29] 29%) passed the test.

The name derives from "la tortuga", the Spanish word for [MT("tortuga") → turtle] turtle.

The Brown Act is California's law [WikiSearch("Brown Act") → The Ralph M. Brown Act is an act of the California State Legislature that guarantees the public's right to attend and participate in meetings of local legislative bodies.] that requires legislative bodies, like city councils, to hold their meetings open to the public.

Limitations: Hallucinations

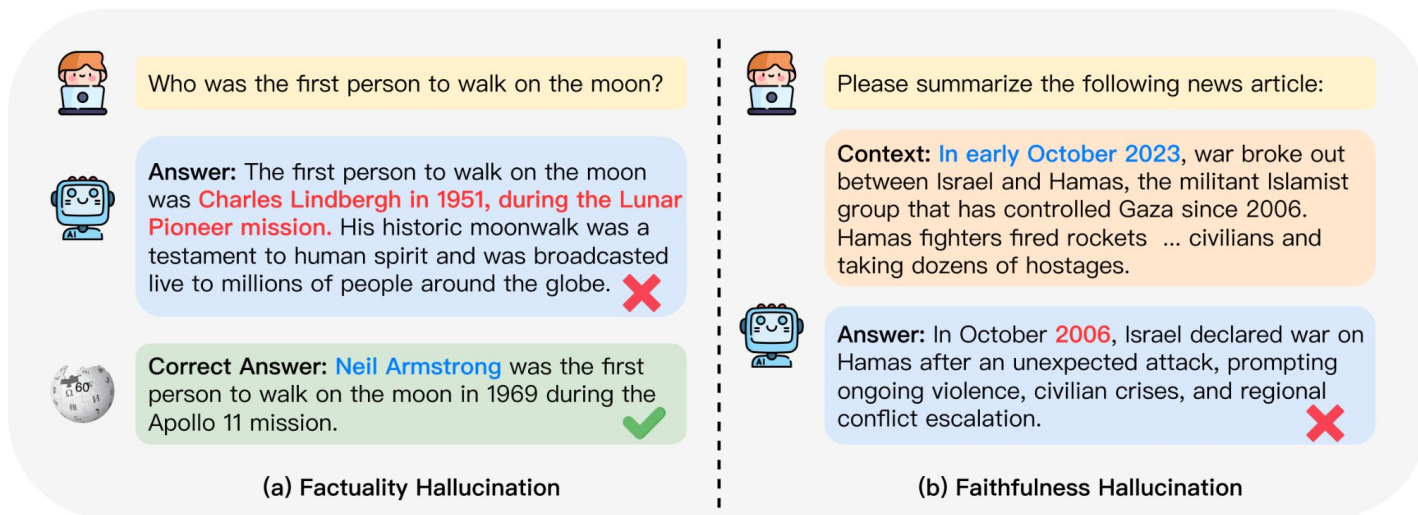
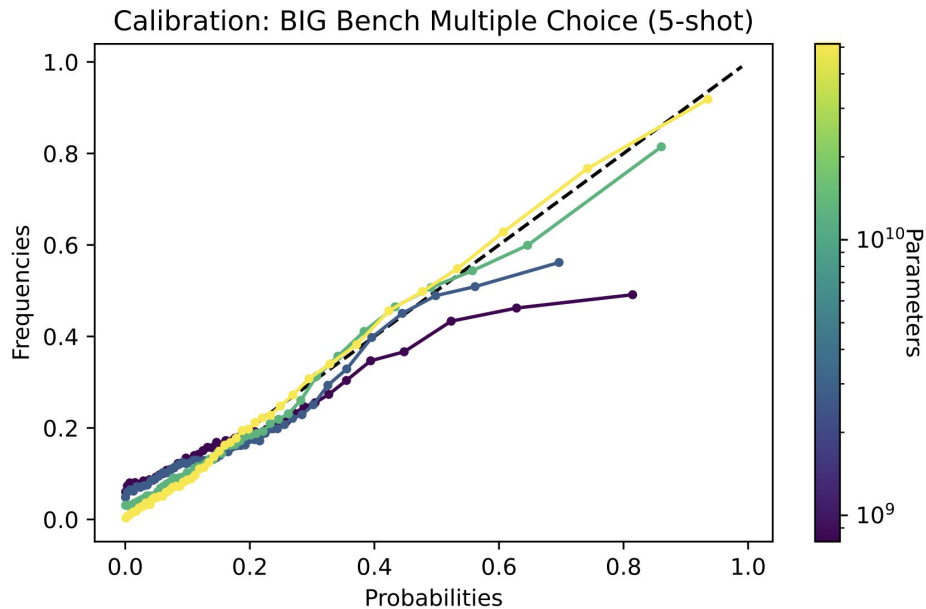


Figure 1: An intuitive example of LLM hallucination.

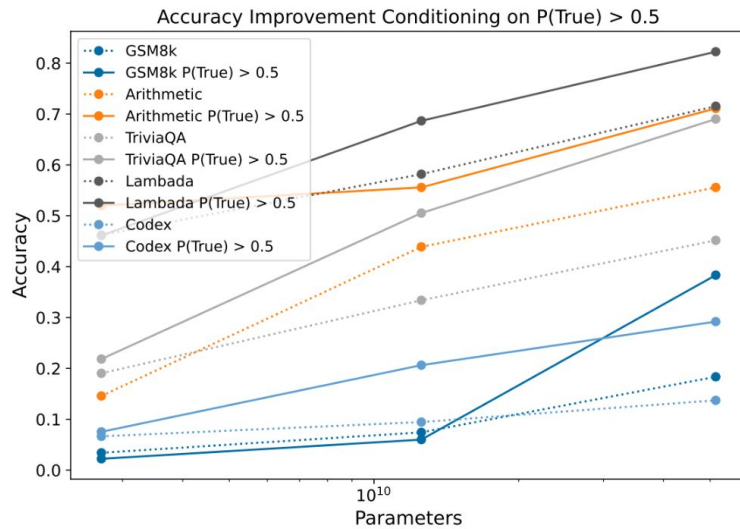
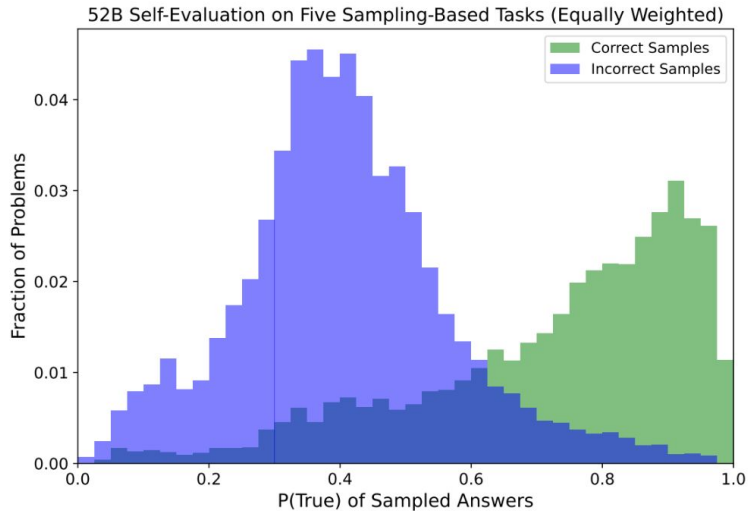
How to Incorporate Uncertainty

- Language Models (Mostly) Know What They Know
 - Kadavath et al. (2022)
- LMs are pretty well calibrated!
 - Multiple choice evaluation



How to Incorporate Uncertainty

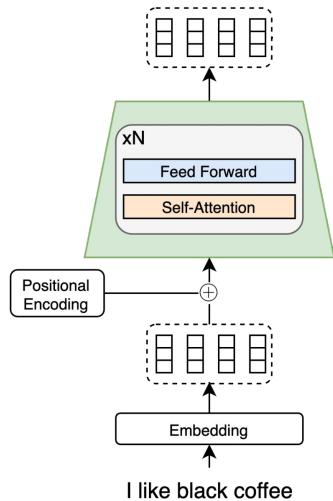
- Language Models (Mostly) Know What They Know
 - Kadavath et al. (2022)
- Can ask the LM to evaluate the likelihood of their own answers



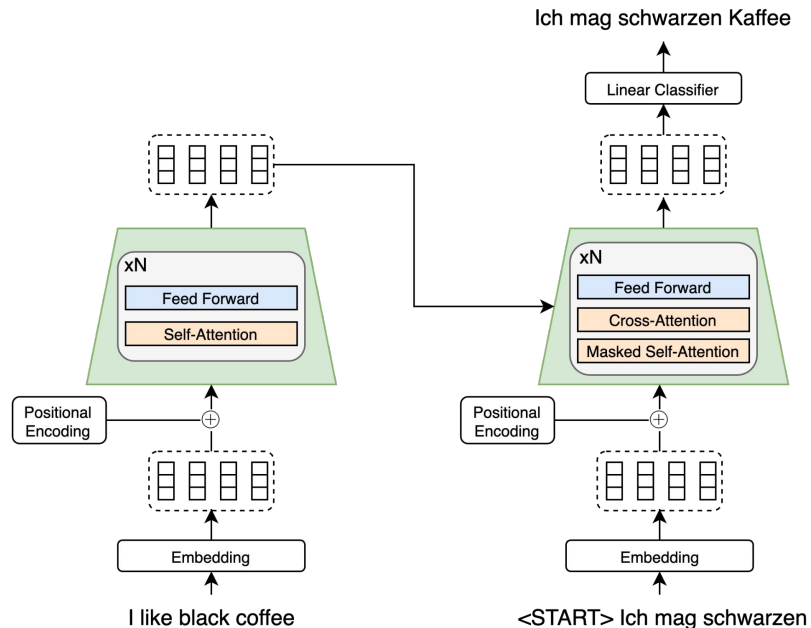
Significant Energy Consumption

Consumption	CO₂e (lbs)
Air travel, 1 passenger, NY↔SF	1984
Human life, avg, 1 year	11,023
American life, avg, 1 year	36,156
Car, avg incl. fuel, 1 lifetime	126,000
Training one model (GPU)	
NLP pipeline (parsing, SRL)	39
w/ tuning & experimentation	78,468
Transformer (big)	192
w/ neural architecture search	626,155

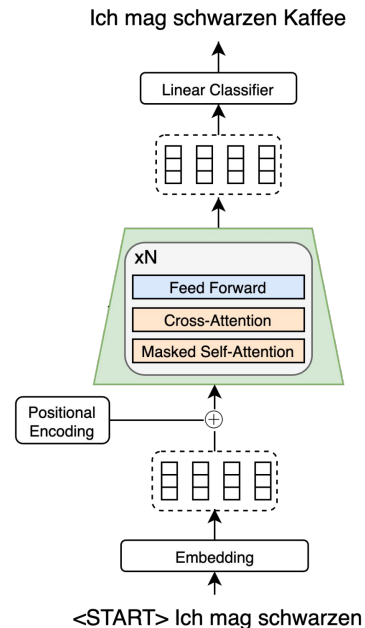
BERT



T5/BART



GPT



Recap

- Different variants of language models
 - Encoder only, encoder-decoder, decoder only
- Performance improves with data and model size
 - Can actually fit power laws and predict performance before training!
- LLMs learn how to do a variety of tasks during pre-training
- A second phase is required to give pre-trained LMs instruction following capabilities
 - Instruction tuning
- Performance can depend on the prompt
- LMs often hallucinate!