GPT-3

Qumeng Sun

Supervised by Michaela Vystrcilová

20/06/2023

UNIVERSITÄT
GÖTTINGEN

Language Models are Few-Shot Learners

2 Approach

Our basic pre-training approach, including model, data, and training, is similar to the process described in [RWC+19], with relatively straightforward scaling up of the model size, dataset size and diversity, and length of training. Our use of in-context learning is also similar to [RWC+19], but in this work we systematically explore different settings for learning within the context. Therefore, we start this section by explicitly defining and contrasting the different settings that we will be evaluating GPT-3 on or could in principle evaluate GPT-3 on. These settings can be seen as lying on a spectrum of how much task-specific data they tend to rely on. Specifically, we can identify at least four points on this spectrum (see Figure 2.1 for an illustration):

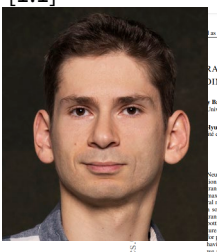
In the Next 25 mins

- Background Med Kit
  - The history of GPTs
  - The architecture of GPTs
- Features
  - Large
  - Meta-Learning
- Performance
- Shortcomings

- Definitely not all parts of the paper (41 pages without appendix)

Background Medical Kit

[1.1]



[1.2]

Attention 2014 [1.4]

Université de Montréal [1.3]

PLM 2003 [1.5]

Transformer 2017 [1.6]

Google AI

Decoder 2018.1 [1.7]

GPT 2018.6 [1.9~11]

GPT-2 2019

GPT-3 2020

Google Brain

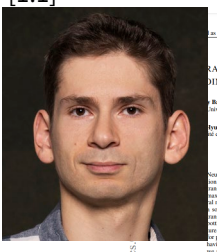
BERT 2018.10 [1.8]

Adapted from Mu Li (2021)

[1.1] D.Bahdanau's blog https://rizar.github.io/
[1.2] Photo by KOKI-KIKO, from C2 Montréal
[1.3] UdeM https://crown.edu.mm/lecturer/universite-de-montreal/

Background Medical Kit

[1.1]



[1.2]

[1.3]

Université de Montréal

[1.4]

Attention 2014

[1.5]

PLM 2003

[1.6]

Transformer 2017

[1.7]

Decoder 2018.1

[1.8]

BERT 2018.10

[1.9~11]

GPT 2018.6

GPT-2 2019

GPT-3 2020

OpenAI

Microsoft

Google AI

Google Brain

Adapted from Mu Li (2021)

[1.1] D.Bahdanau's blog https://rizar.github.io/
[1.2] Photo by KOKI-KIKO, from C2 Montréal
[1.3] UdeM https://crown.edu.mm/lecturer/universite-de-montreal/

Transformer

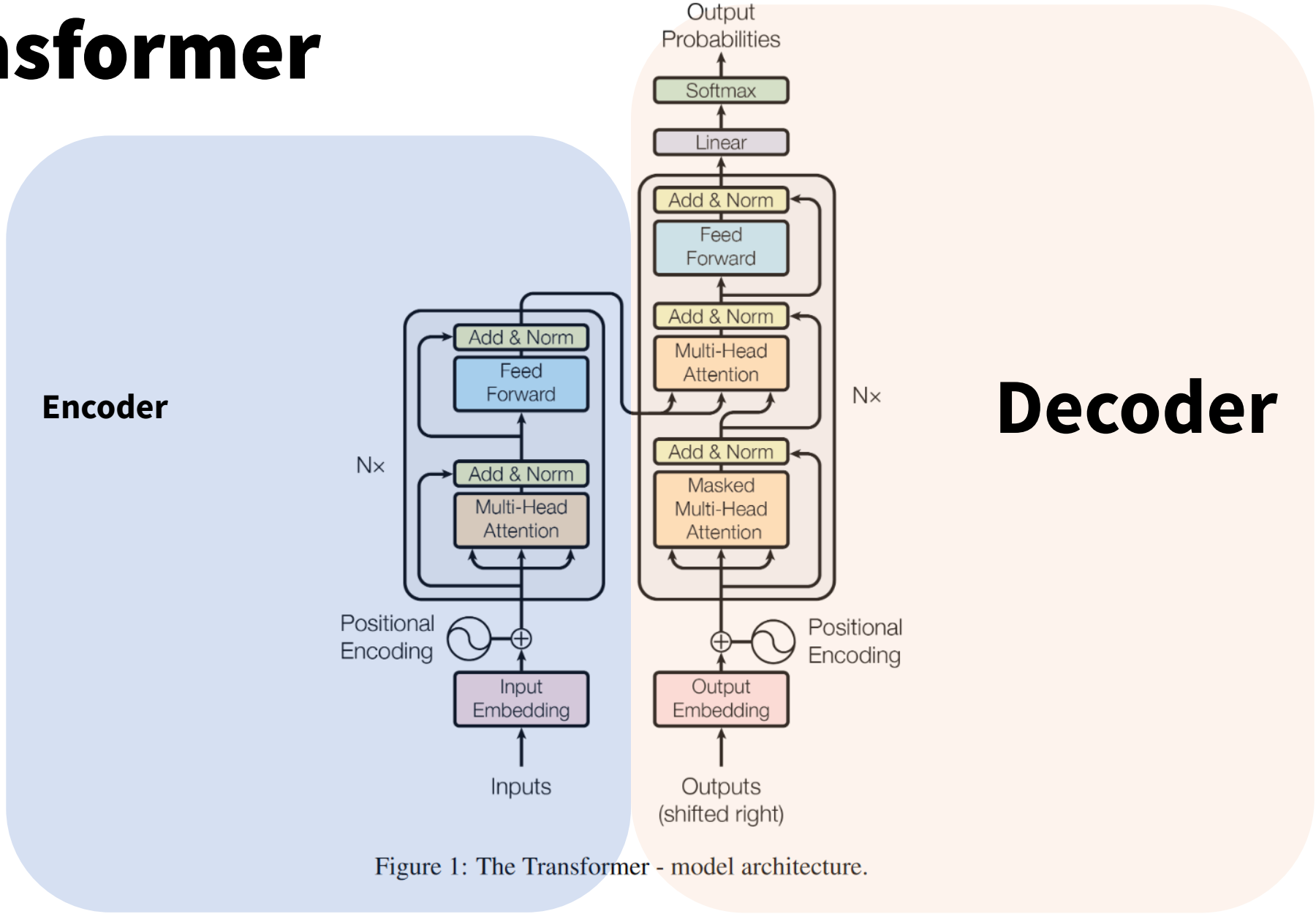
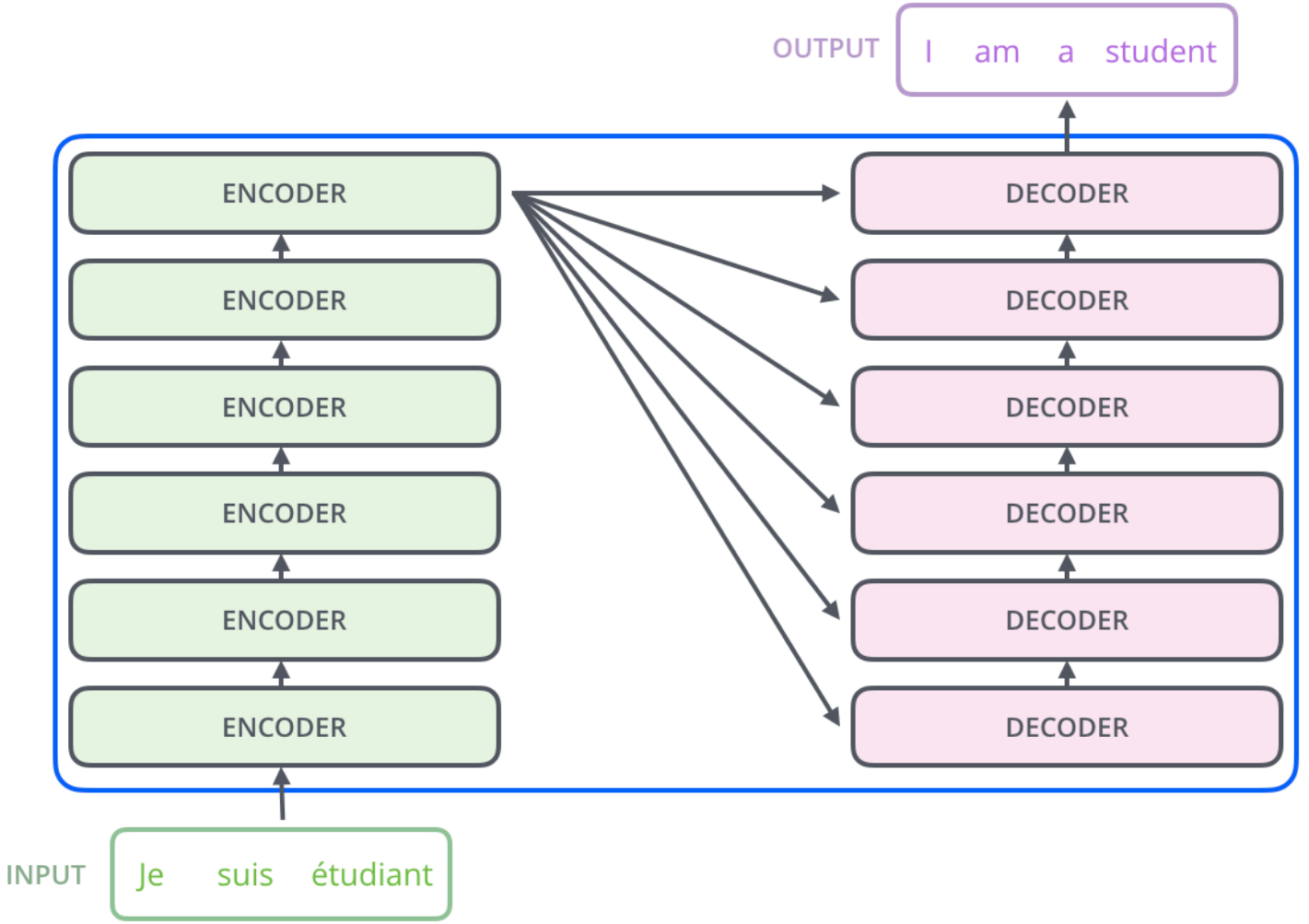


Figure 1: The Transformer - model architecture.

Attention Is All You Need, 2017

Hidden slide

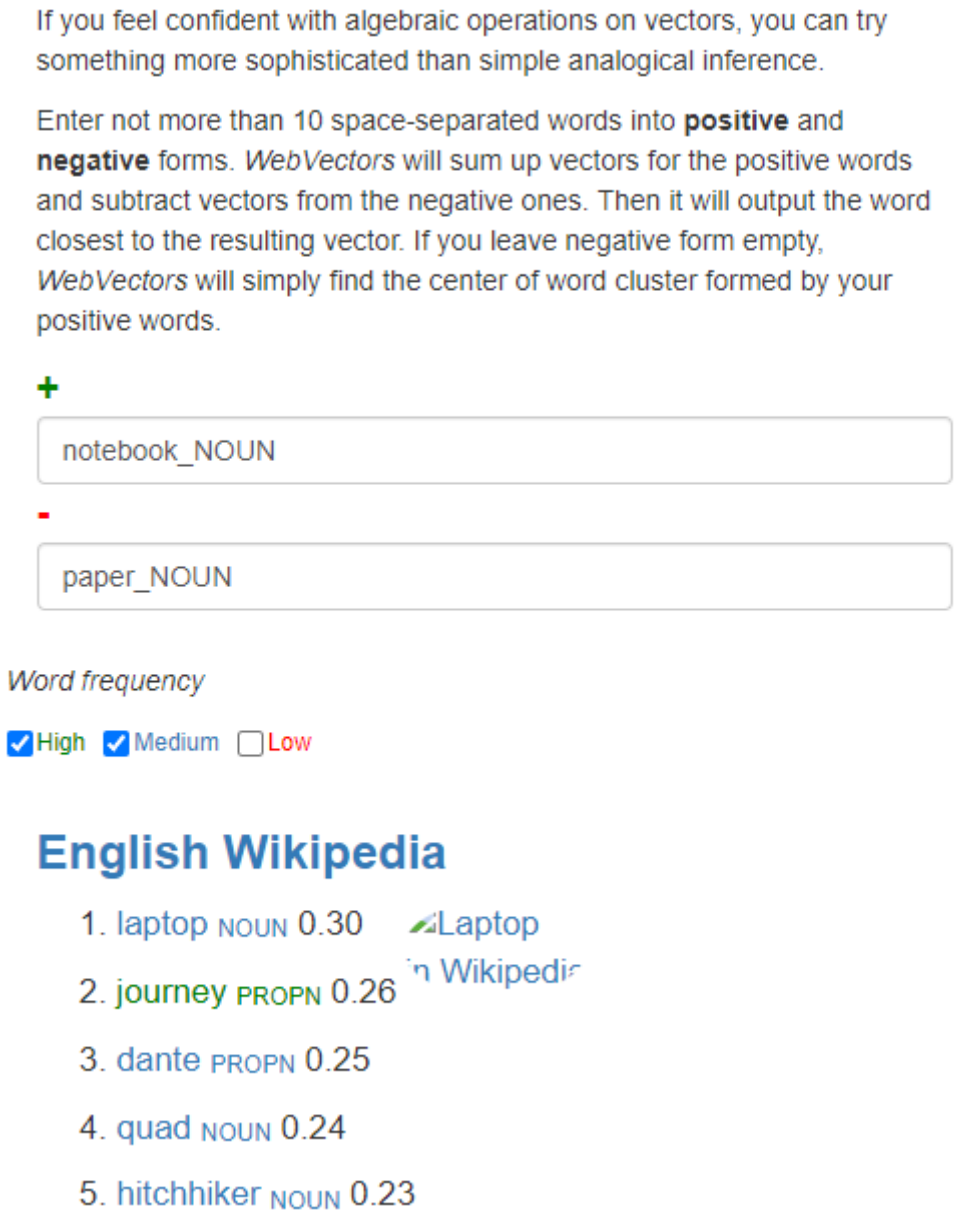


Credit : Jay Alammar, The Illustrated Transformer

Hidden slide

Word embeddings[2]

- Word2Vec
- King - Man + Woman = Queen



http://vectors.nlpl.eu/explore/embeddings/en/calculator/

GPT-1

Pre-training

Fine-tuning

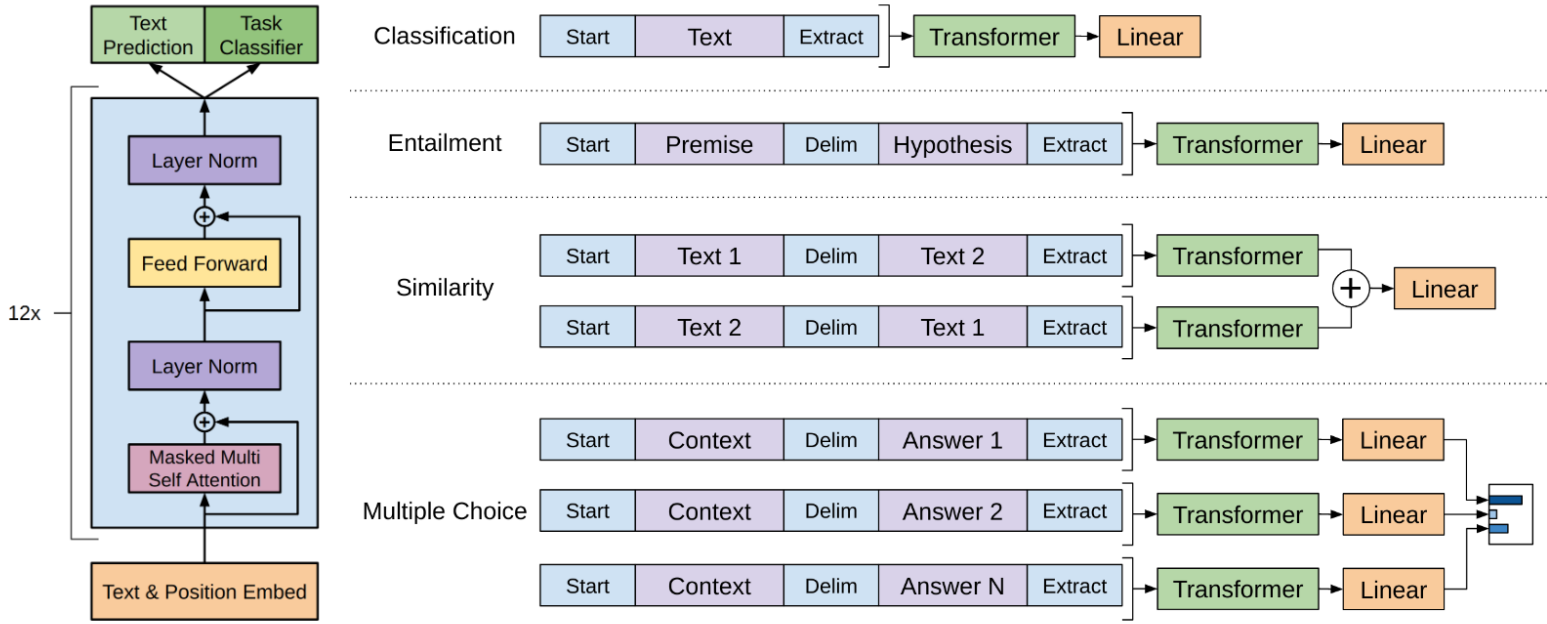


Figure 1: **(left)** Transformer architecture and training objectives used in this work. **(right)** Input transformations for fine-tuning on different tasks. We convert all structured inputs into token sequences to be processed by our pre-trained model, followed by a linear+softmax layer.

Start token <s>,

Extract token <e>,

The delimiter token is ($),

+ added element-wise,

Normalized via a SoftMax layer

Adapted from Maziar Raissi(2021), and Alec Radford et al.(2018)

Hidden slide

GPT-1 (Unsupervised Pre-Training)

$\mathcal{U} = \{u_1, \ldots, u_n\}$ → an unsupervised corpus of words

$$L_1(\mathcal{U}) = \sum_i \log P(u_i | u_{i-k}, \ldots, u_{i-1}; \Theta)$$

→ a standard language modeling objective to maximize the likelihood(MLE)

$U = (u_{-k}, \ldots, u_{-1})$ → Word Embeddings

$h_0 = UW_e + W_p$ → Position embeddings

→ Token embeddings

$h_l = \texttt{transformer\_block}(h_{l-1}) \forall i \in [1, n]$

$P(u) = \texttt{softmax}(h_n W_e^T)$

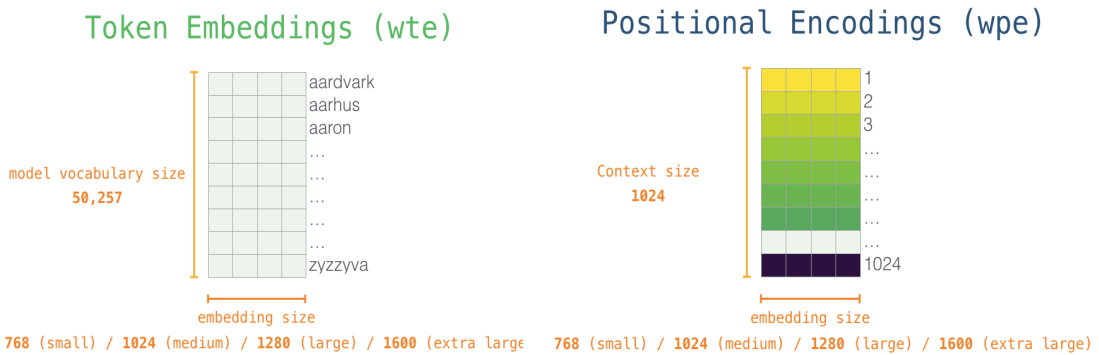


Figure Weight Matrix[2.1]

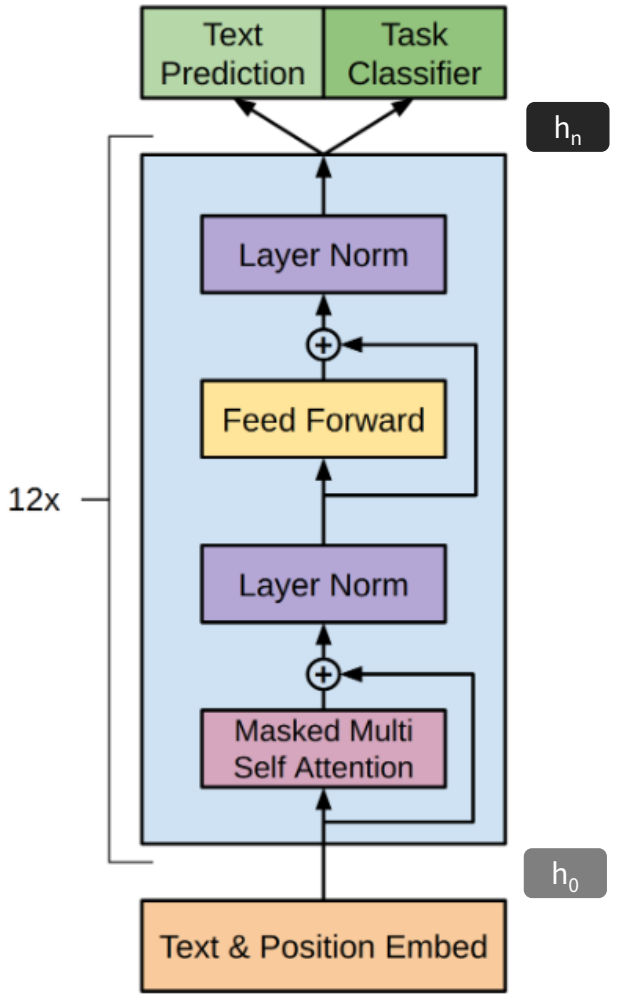


Figure Pre-Training[1.9]

[2.1] Jay Alammar. (2021)
Adapted from Maziar Raissi(2021), and Alec Radford et al.(2018)

Hidden slide

GPT-1 (Supervised Fine-Tuning)

$\mathcal{C}$ → labeled dataset $x^1, \ldots, x^m$ → sequence of input tokens

h_l^m → The final transformer block's activation

$$P(y|x^1, \ldots, x^m) = \texttt{softmax}(h_l^m W_y).$$

→ The added linear output layer, to get probability distribution.

$$L_2(\mathcal{C}) = \sum_{(x,y)} \log P(y|x^1, \ldots, x^m).$$

→ Objective to maximize

$$L_3(\mathcal{C}) = L_2(\mathcal{C}) + \lambda * L_1(\mathcal{C})$$

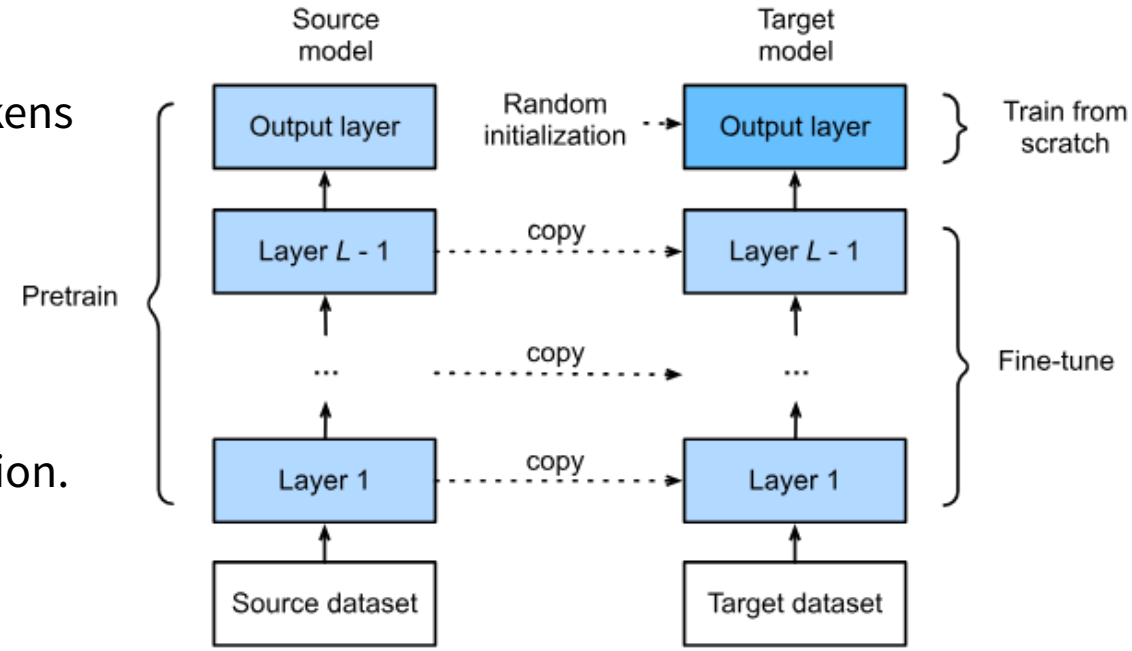


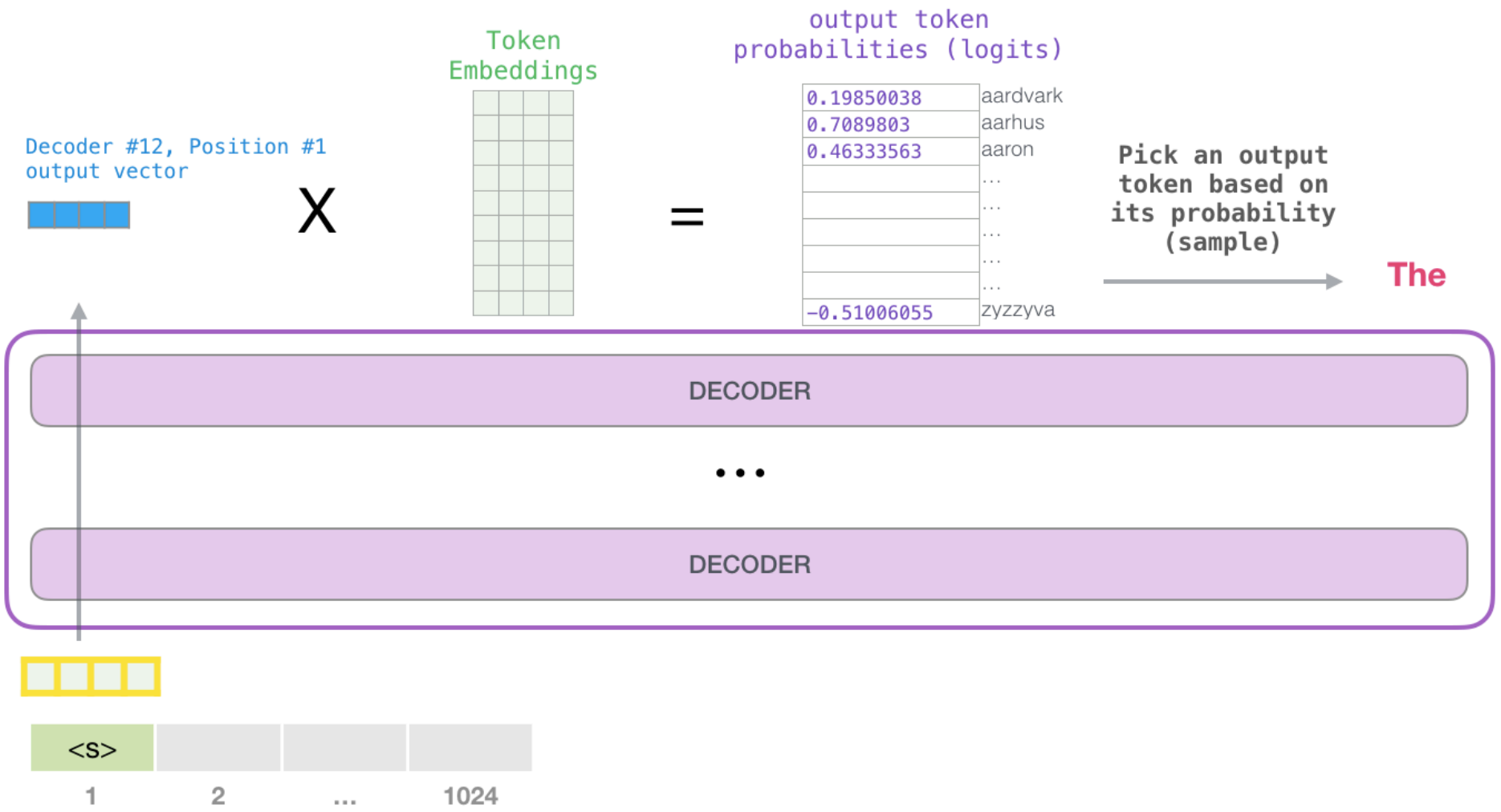
Figure Fine-tuning [3]

An auxiliary objective to the fine-tuning helped learning by
(a) improving the generalization of the supervised model and
(b) accelerating convergence.

[3] Zhang et al.(2021)

Adapted from Maziar Raissi(2021), and Alec Radford et al.(2018)

Hidden slide



Credit : Jay Alammar, The Illustrated Transformer

GPT-2

$(x_1, x_{2,\ldots}, x_n)$ → set of examples(sentences)

$x = (s_1, s_2, \cdots, s_n)$ → variable-length sequence of symbols

$p(x) = \prod_{i=1}^{n} p(s_n | s_{1,\ldots}, s_{n-1})$

$p(s_{n-k}, \ldots, s_n | s_1, \ldots, s_{n-k-1})$

$p(output \mid input, task)$

Example:

p("Good Morning" | "Guten Morgen", translation to English)

Input: (translate from German to English, "Guten Morgen")

[x] Fine-tuning
[x] Linear Layers for different tasks

In GPT-1, the task conditioning at the **architectural level**(or outer loop) cannot change after training(or fine-tuning)

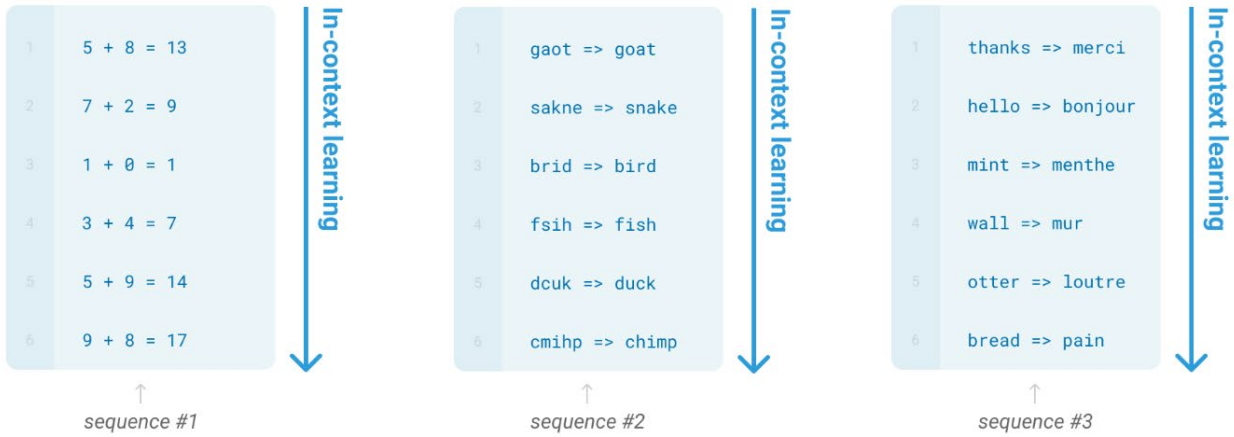
In GPT-2, we perform task conditioning at the **algorithmic level**(or inner loop) called **Meta-Learning**[5]

Meta-Learning = Inner Loop + Outer Loop

Gradient update when pre-training

Learning via SGD during unsupervised pre-training

In-context learning



Credit : The paper

GPT

$\mathcal{U} = \{u_1, \ldots, u_n\}$ → an unsupervised corpus of words

$$L_1(\mathcal{U}) = \sum_i \log P(u_i | u_{i-k}, \ldots, u_{i-1}; \Theta)$$

→ a standard language modeling objective to maximize the likelihood(MLE)

$U = (u_{-k}, \ldots, u_{-1})$ → Word Embeddings

$h_0 = UW_e + W_p$ → Position embeddings

→ Token embeddings

$h_l = \texttt{transformer\_block}(h_{l-1}) \forall i \in [1, n]$

$P(u) = \texttt{softmax}(h_n W_e^T)$

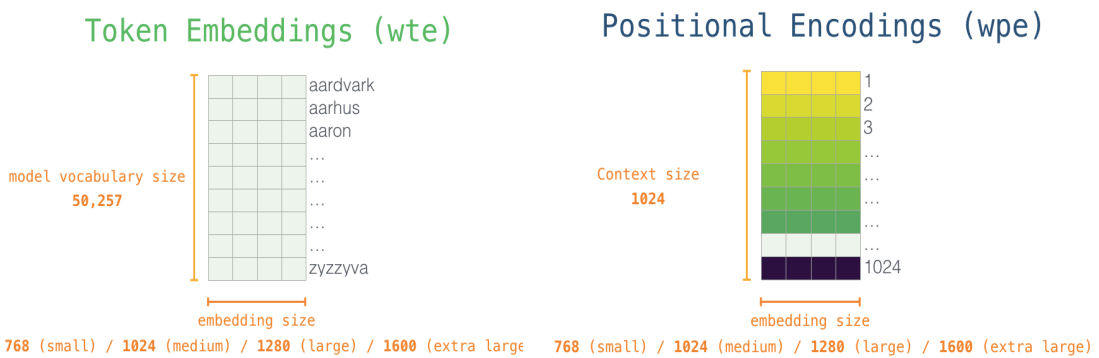


Figure Weight Matrix[2.1]

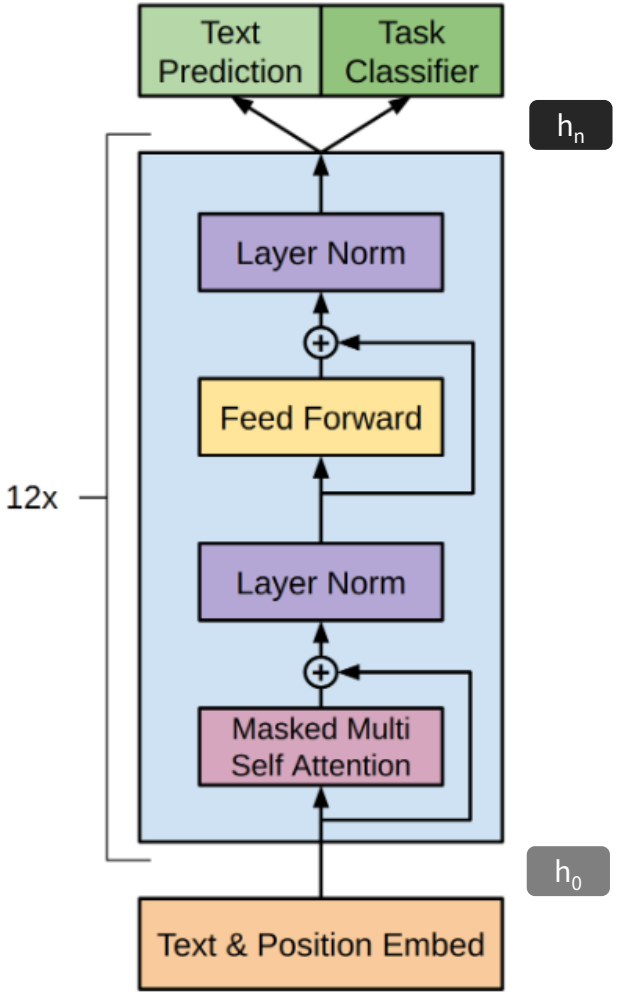


Figure Pre-Training[1.9]

[2.1] Jay Alammar. (2021)
Adapted from Maziar Raissi(2021), and Alec Radford et al.(2018)

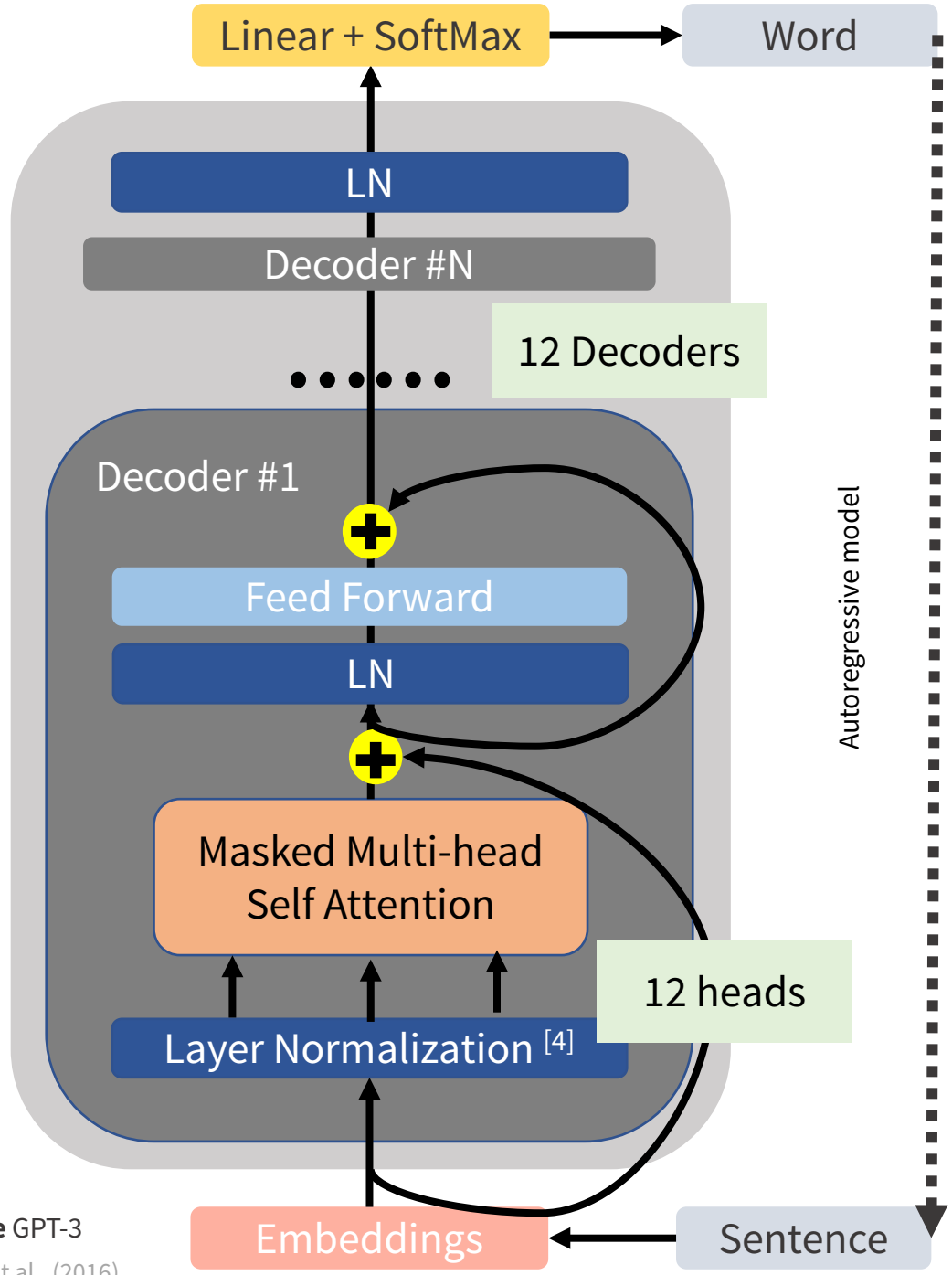


Figure GPT-3

[4] He et al., (2016)

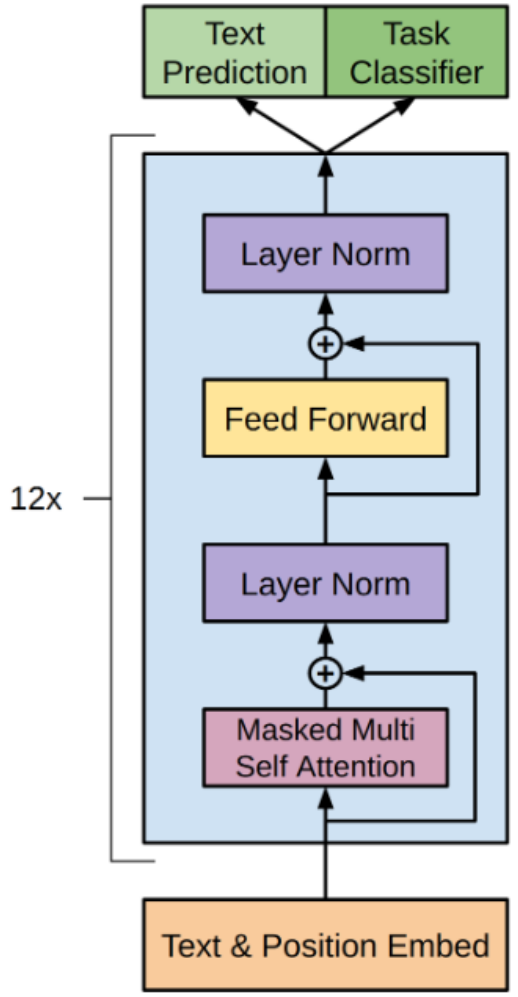
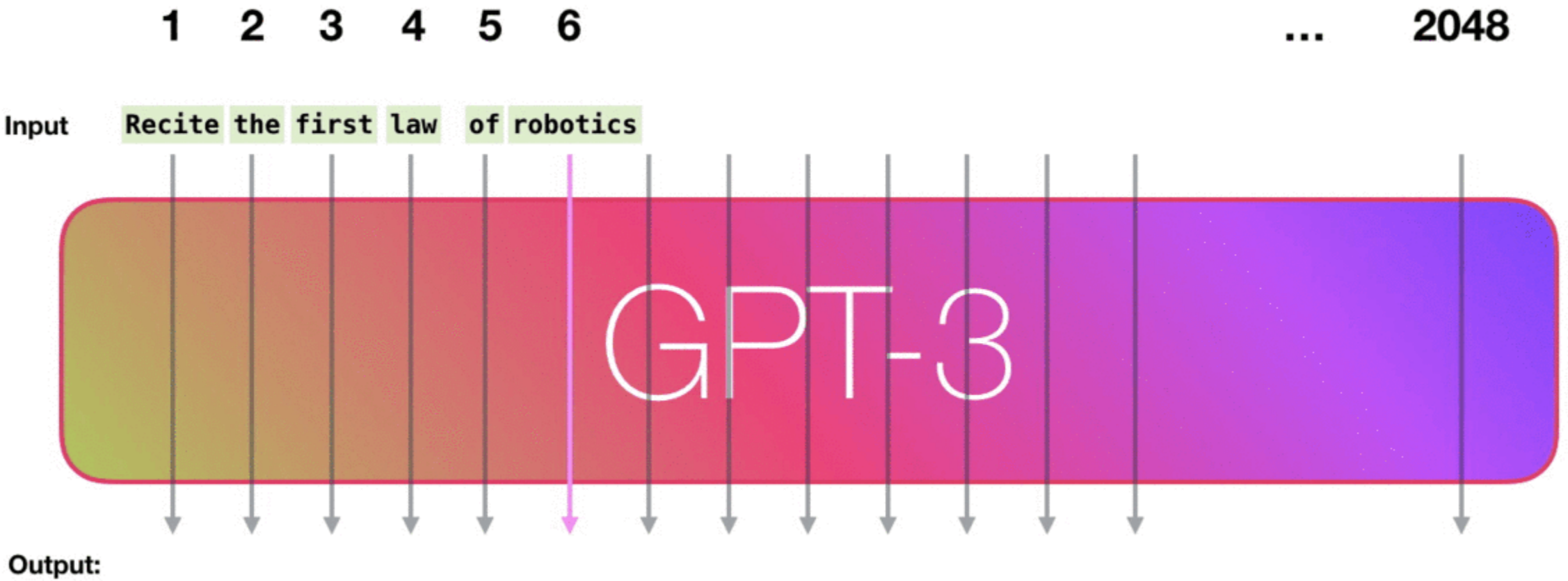


Figure GPT-1[1,9]

GPT-3

Changes from GPT-1

- Layer normalization(LN) was moved to the input of each sub-block, similar to a pre-activation residual network.
- An additional LN was added after the final self-attention block.



Credit: Jay Alammar(2020). How GPT3 Works - Visualizations and Animations

Larger without Fine-tuning

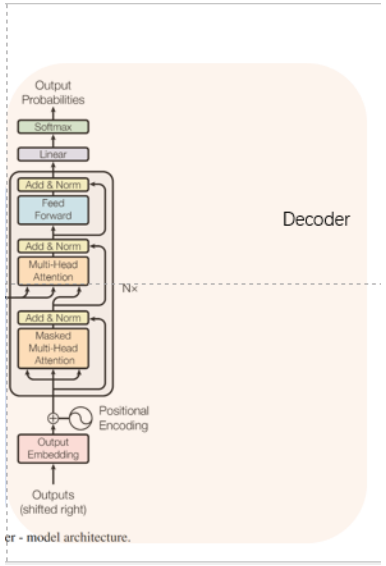
Much larger with Few-shot

Google AI

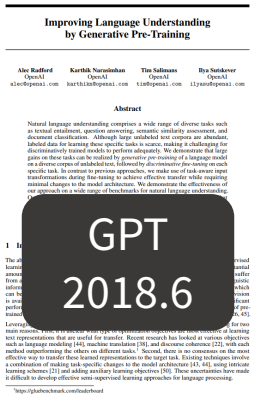
GENERATING WIKIPEDIA BY SUMMARIZING LONG SEQUENCES

Peter J. Liu*, Mohammad Saleh,
Etienne Pot, Ben Goodrich, Ryan Sepassi, Łukasz Kaiser, Noam Shazeer
Google Brain
Mountain View, CA
{peterjliu,msaleh,epot,bgoodrich,rsepassi,lukaszkaiser,noam}@google.com

Decoder 2018.1 [1.7]



er - model architecture.


Improving Language Understanding by Generative Pre-Training

Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever
OpenAI

Abstract

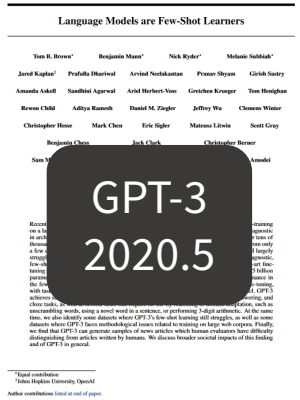
GPT 2018.6 [1.9~11]


Language Models are Unsupervised Multitask Learners

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever

Abstract

GPT-2 2019.2


Language Models are Few-Shot Learners

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, Dario Amodei

arXiv:2005.14165v4 [cs.CL] 22 Jul 2020

GPT-3 2020.5

Google Brain


BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding

Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova
Google AI Language
{jacobdevlin,mingweichang,kentonl,kristout}@google.com

Abstract

arXiv:1810.04805v2 [cs.CL] 24 May 2019

BERT 2018.10 [1.8]

GPT-3

Motivation:

1. From a practical perspective, the need for a large dataset of labeled examples for every new task **limits the applicability** of language models.
2. Attempts to fine-tune in response to specific tasks and data sets are suspect of **bragging**.
3. The way **humans** process information differs from fine-tuning.

Goal: Training a fine-tuning-free model that performs well on multiple tasks

Research Gap: The current model that has yet to be fine-tuned cannot challenge the results of the fine-tuned model.

Method: LLM, Prompt/in-context learning/meta-learning

Features

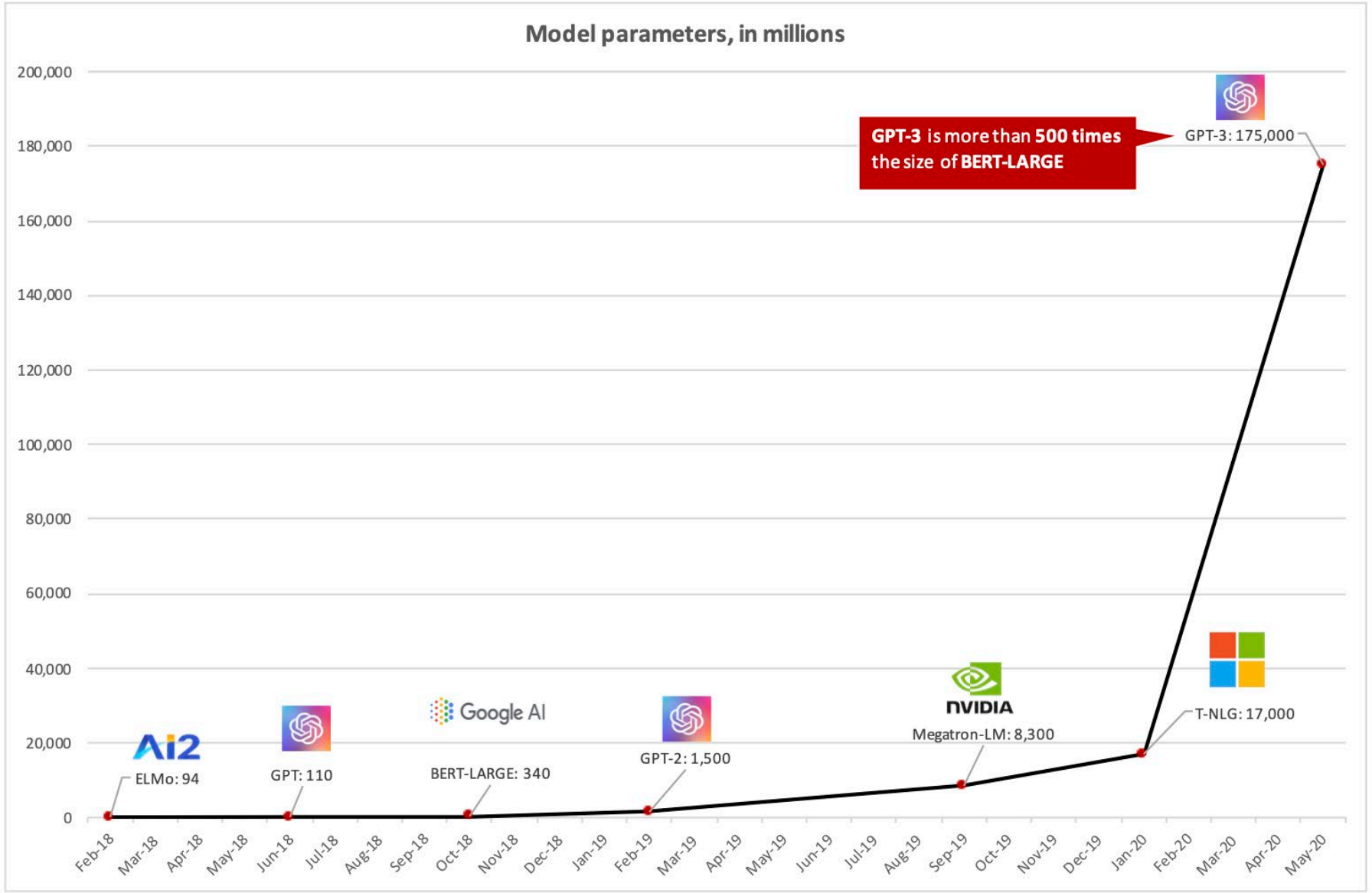
1. Large
2. One Model for Multiple Tasks

Features

1. Large
   - Scaling laws for neural language models(Kaplan et al., 2020)
   - 8 models. The biggest one is called GPT-3
2. One Model for Multiple Tasks

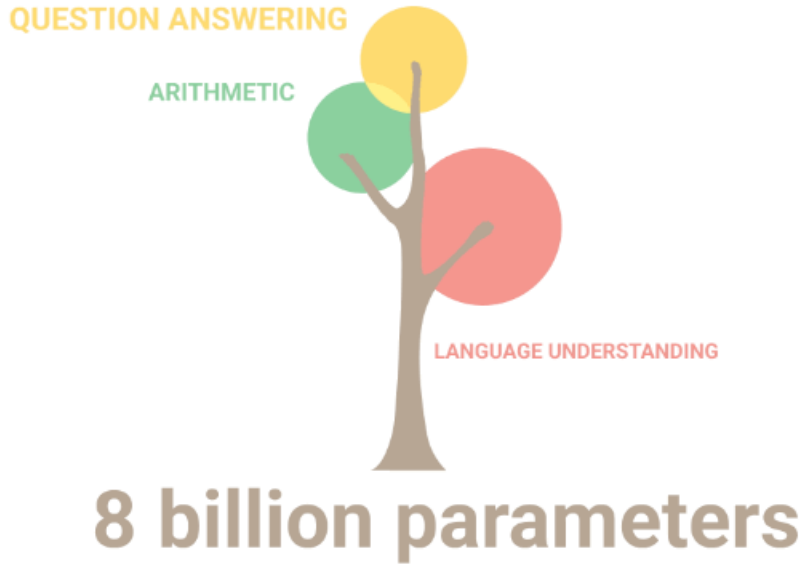
Model Name	n_{params}	n_{layers}	d_{model}	n_{heads}	d_{head}	Batch Size	Learning Rate
GPT-3 Small	125M	12	768	12	64	0.5M	6.0×10^{-4}
GPT-3 Medium	350M	24	1024	16	64	0.5M	3.0×10^{-4}
GPT-3 Large	760M	24	1536	16	96	0.5M	2.5×10^{-4}
GPT-3 XL	1.3B	24	2048	24	128	1M	2.0×10^{-4}
GPT-3 2.7B	2.7B	32	2560	32	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.



Maximilian Ahrens, 2020

https://twitter.com/_MaxAhrens_/status/1285228257252646914

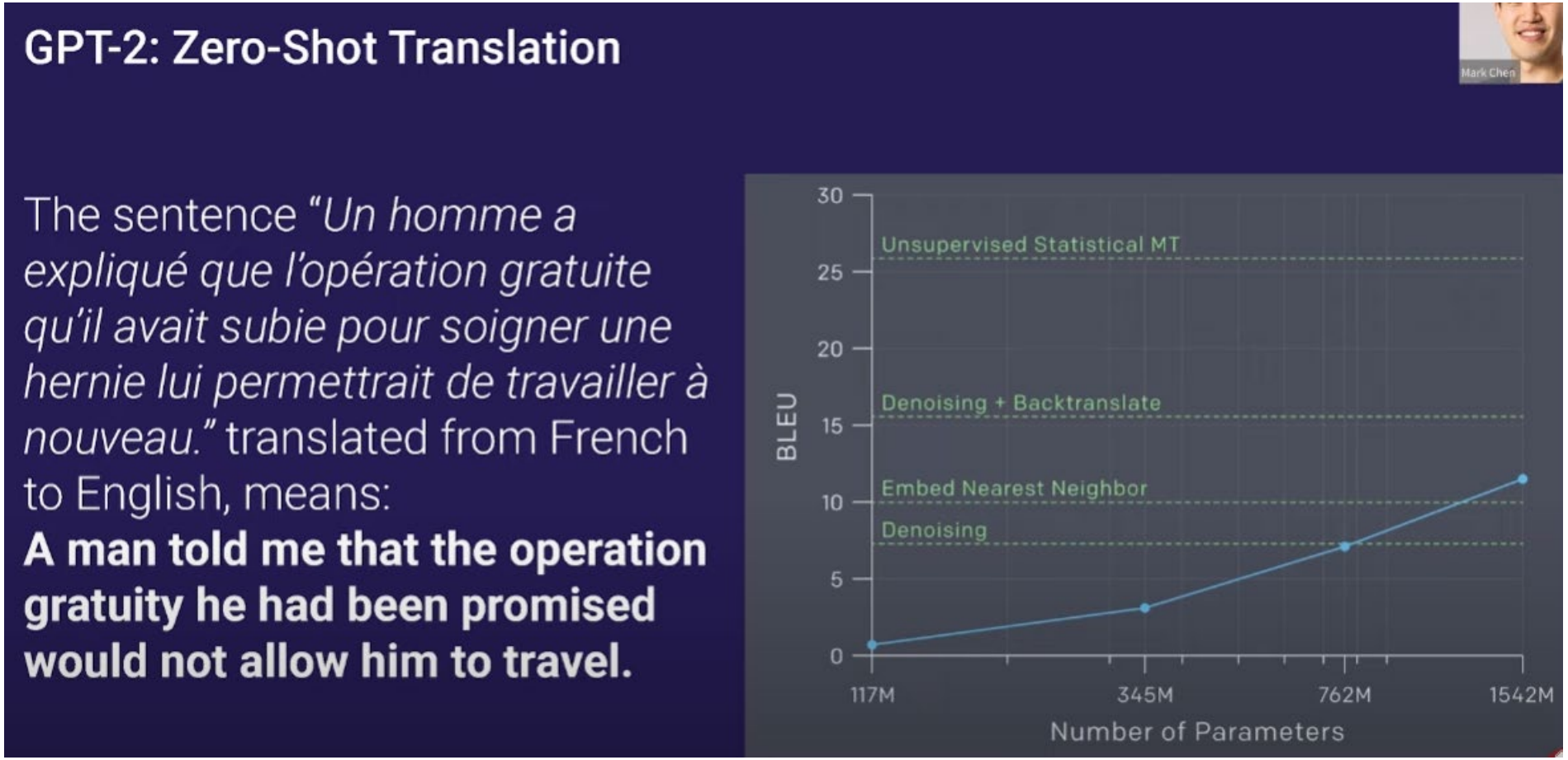


540B PaLM, https://ai.googleblog.com/2022/04/pathways-language-model-palm-scaling-to.html

Features

1. Large
   - Scaling laws for neural language models(Kaplan et al., 2020)
   - 8 models. The biggest one is called GPT-3
2. One Model for Multiple Tasks
   - Train a single model to do thousands or millions of things(Jeff Dean, 2021)
     - Natural language can be recognized without the need to build special input structures.
   - Fine-tuning
     - When faced with different tasks, GPT-3 performs **in-context learning** to adapt to different tasks, and it **does not do any fine-tuning or gradient updating** in the process.
     - GPT can be fine-tuned.
   - Few-shot

Zero-Shot GPT-2



Mark Chen. (2022), https://www.youtube.com/watch?v=qGkzHFllWDY

Few-shot

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
```

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

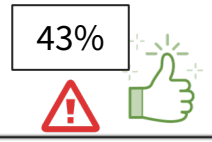
Fine-tuning

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

Credit: The paper

Performance

The Next Word/Sentence Task



Setting	LAMBADA (acc)	LAMBADA (ppl)	StoryCloze (acc)	HellaSwag (acc)
SOTA	68.0[a]	8.63[b]	**91.8**[c]	**85.6**[d]
GPT-3 Zero-Shot	**76.2**	**3.00**	83.2	78.9
GPT-3 One-Shot	**72.5**	**3.35**	84.7	78.1
GPT-3 Few-Shot	**86.4**	**1.92**	87.7	79.3

Table 3.2: Performance on cloze and completion tasks. GPT-3 significantly improves SOTA on LAMBADA while achieving respectable performance on two difficult completion prediction datasets. [a][Tur20] [b][RWC+19] [c][LDL19] [d][LCH+20]

- In the **next word prediction task**, GPT-3 performs well, outperforming the current SOTA.

- LAMBADA: predicting the last word in a paragraph
- StoryCloze: picking a good ending for a story
- HellaSwag: choosing the best ending for a five-sentence story

Data Contamination Alert
Some data is overlapped in the training and test sets

Hidden slide

LAMBADA

(1) *Context:* "Yes, I thought I was going to lose the baby." "I was scared too," he stated, sincerity flooding his eyes. "You were ?" "Yes, of course. Why do you even ask?" "This baby wasn't exactly planned for."
Target sentence: "Do you honestly think that I would want you to have a _____ ?"
Target word: miscarriage

(2) *Context:* "Why?" "I would have thought you'd find him rather dry," she said. "I don't know about that," said Gabriel. "He was a great craftsman," said Heather. "That he was," said Flannery.
Target sentence: "And Polish, to boot," said _____.
Target word: Gabriel

(3) *Context:* Preston had been the last person to wear those chains, and I knew what I'd see and feel if they were slipped onto my skin-the Reaper's unending hatred of me. I'd felt enough of that emotion already in the amphitheater. I didn't want to feel anymore. "Don't put those on me," I whispered. "Please."
Target sentence: Sergei looked at me, surprised by my low, raspy please, but he put down the _____.
Target word: chains

(4) *Context:* They tuned, discussed for a moment, then struck up a lively jig. Everyone joined in, turning the courtyard into an even more chaotic scene, people now dancing in circles, swinging and spinning in circles, everyone making up their own dance steps. I felt my feet tapping, my body wanting to move.
Target sentence: Aside from writing, I 've always loved _____.
Target word: dancing

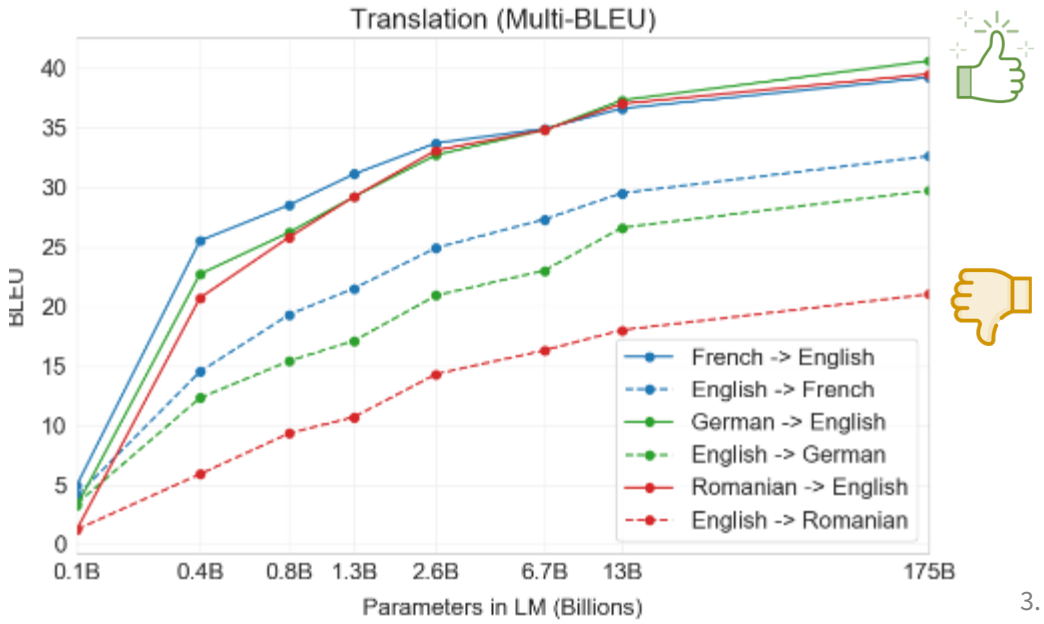
(5) *Context:* He shook his head, took a step back and held his hands up as he tried to smile without losing a cigarette. "Yes you can," Julia said in a reassuring voice. "I 've already focused on my friend. You just have to click the shutter, on top, here."
Target sentence: He nodded sheepishly, through his cigarette away and took the _____.
Target word: camera

(6) *Context:* In my palm is a clear stone, and inside it is a small ivory statuette. A guardian angel. "Figured if you're going to be out at night getting hit by cars, you might as well have some backup." I look at him, feeling stunned. Like this is some sort of sign.
Target sentence: But as I stare at Harlin, his mouth curved in a confident grin, I don't care about _____.
Target word: signs

Translation

20%

Setting	En→Fr	Fr→En	En→De	De→En	En→Ro	Ro→En
SOTA (Supervised)	**45.6**[a]	35.0 [b]	**41.2**[c]	40.2[d]	**38.5**[e]	**39.9**[e]
XLM [LC19]	33.4	33.3	26.4	34.3	33.3	31.8
MASS [STQ+19]	37.5	34.9	28.3	35.2	35.2	33.1
mBART [LGG+20]	-	-	29.8	34.0	35.0	30.5
GPT-3 Zero-Shot	25.2	21.2	24.6	27.2	14.1	19.9
GPT-3 One-Shot	28.3	33.7	26.2	30.4	20.6	38.6
GPT-3 Few-Shot	32.6	39.2	29.7	40.6	21.0	39.5

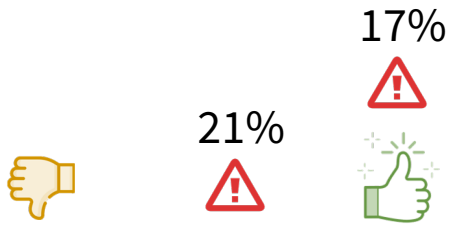


- GPT-3 achieves SOTA level in all translation tasks where **the target language is English**.
- Performance is better when our output is close to the training data.
- **The frequency** with which the desired output appears **in the training** set is critical.

News Article Generation

	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%

QA



Setting	NaturalQS	WebQS	TriviaQA
RAG (Fine-tuned, Open-Domain) [LPP+20]	**44.5**	**45.5**	**68.0**
T5-11B+SSM (Fine-tuned, Closed-Book) [RRS20]	36.6	44.7	60.5
T5-11B (Fine-tuned, Closed-Book)	34.5	37.4	50.1
GPT-3 Zero-Shot	14.6	14.4	64.3
GPT-3 One-Shot	23.0	25.3	**68.0**
GPT-3 Few-Shot	29.9	41.5	**71.2**

Table 3.3: Results on three Open-Domain QA tasks. GPT-3 is shown in the few-, one-, and zero-shot settings, as compared to prior SOTA results for closed book and open domain settings. TriviaQA few-shot result is evaluated on the wiki split test server.

Natural QS

Question: what color was john wilkes booth's hair
Wikipedia Page: John_Wilkes_Booth
Long answer: Some critics called Booth "the handsomest man in America" and a "natural genius", and noted his having an "astonishing memory"; others were mixed in their estimation of his acting. He stood 5 feet 8 inches (1.73 m) tall, had jet-black hair , and was lean and athletic. Noted Civil War reporter George Alfred Townsend described him as a "muscular, perfect man" with "curling hair, like a Corinthian capital".

Short answer: jet-black

Context →	Q: Who played tess on touched by an angel? A:
Target Completion →	Delloreese Patricia Early (July 6, 1931 { November 19, 2017), known professionally as Della Reese

Figure G.24: Formatted dataset example for Natural Questions

- GPT-3 performs poorly on **long text generation**
  - NQS needs long answers.
  - TQS needs only short answers.
- GPT-3 needs a **detailed prompt**.
  - NQS has short questions
  - TQA has long questions (avg. 14)

Trivia QA

Question: The Dodecanese Campaign of WWII that was an attempt by the Allied forces to capture islands in the Aegean Sea was the inspiration for which acclaimed 1961 commando film?
Answer: The Guns of Navarone
Excerpt: The Dodecanese Campaign of World War II was an attempt by Allied forces to capture the Italian-held Dodecanese islands in the Aegean Sea following the surrender of Italy in September 1943, and use them as bases against the German-controlled Balkans. The failed campaign, and in particular the Battle of Leros, inspired the 1957 novel **The Guns of Navarone** and the successful 1961 movie of the same name.

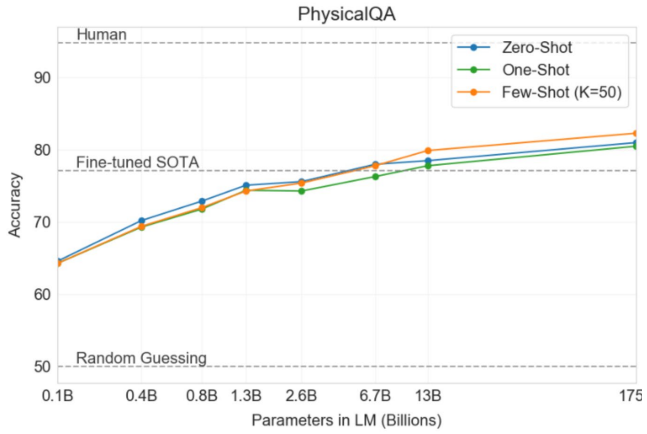
Context →	Q: 'Nude Descending A Staircase' is perhaps the most famous painting by which 20th century artist? A:
Target Completion →	MARCEL DUCHAMP
Target Completion →	r mutt
Target Completion →	duchamp
Target Completion →	marcel duchamp
Target Completion →	R.Mutt
Target Completion →	Marcel duChamp
Target Completion →	Henri-Robert-Marcel Duchamp
Target Completion →	Marcel du Champ
Target Completion →	henri robert marcel duchamp
Target Completion →	Duchampian
Target Completion →	Duchamp
Target Completion →	duchampian
Target Completion →	marcel du champ
Target Completion →	Marcel Duchamp
Target Completion →	MARCEL DUCHAMP

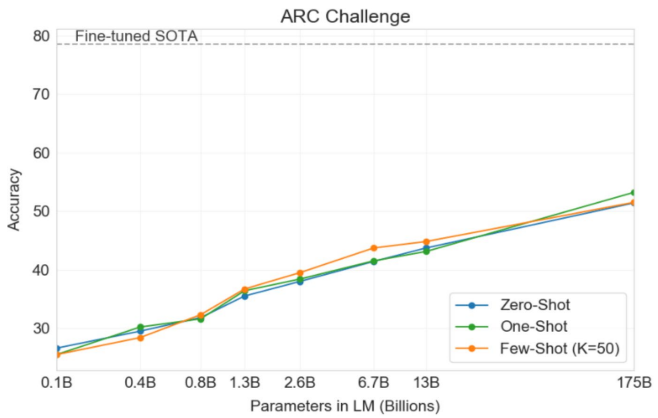
Figure G.34: Formatted dataset example for TriviaQA. TriviaQA allows for multiple valid completions.

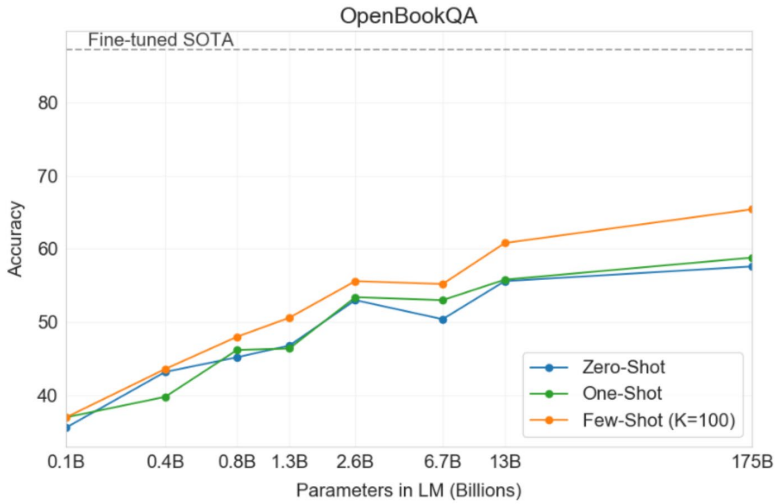
Abstract Logical Thinking

Reasoning

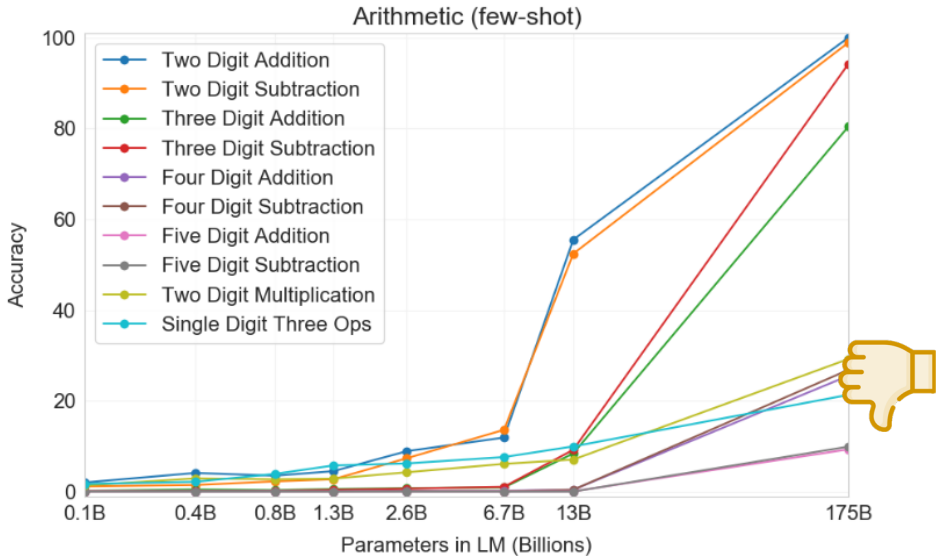
29%







Arithmetic



- The higher the **frequency**, the better the results

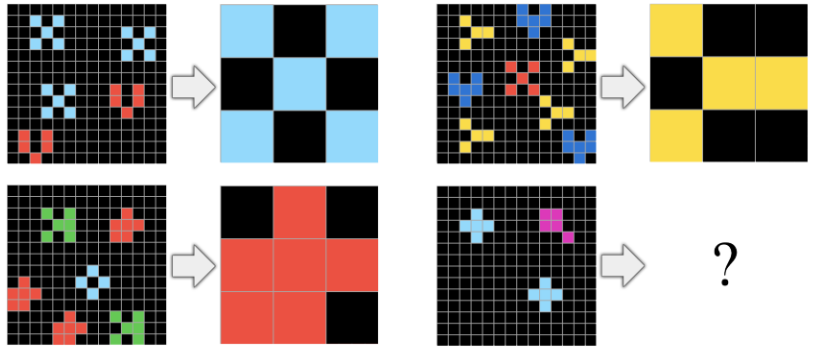


Figure 10: A task where the implicit goal is to count unique objects and select the object that appears the most times (the actual task has more demonstration pairs than these three).

Credit: François Chollet, ARC

Question:
Which of these would let the most heat travel through?
A) a new pair of jeans.
B) a steel spoon in a cafeteria.
C) a cotton candy at a store.
D) a calvin klein cotton hat.

Science Fact:
Metal is a thermal conductor.

Common Knowledge:
Steel is made of metal.
Heat travels through a thermal conductor.

Figure 1: An example for a question with a given set of choices and supporting facts.

Credit: Todor Mihaylov et al., Open Book Question Dataset

- Inability to perform complex reasoning
  - While GPT-3 can mimic complex reasoning to some extent, it still **has difficulty understanding complex problems** or performing tasks that require **deeper understanding.**
- GPT-3 **does not have access to the physical world**.
  - No access even to audio and video, so they lack understanding of the natural world

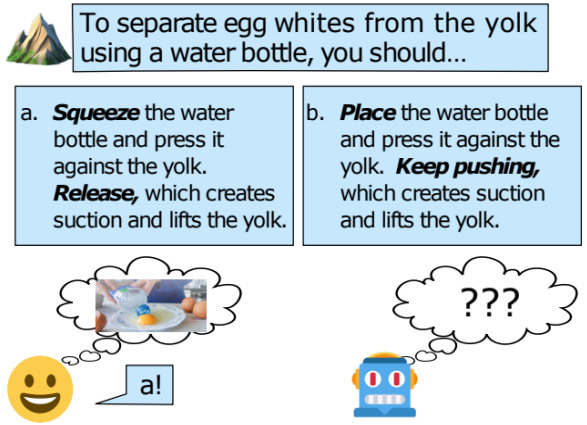


Figure 1: **PIQA** : Given a physical **goal** expressed in natural language, like 'to separate egg whites...,' a model must choose the most sensible **solution**. Our dataset tests the ability of natural language understanding models to link text to a robust intuitive-physics model of the world. Here, humans easily pick answer **a)** because separating the egg requires *pulling* the yolk out, while machines are easily fooled.

Credit: PIQA

Conversational Reading Comprehension

Setting	CoQA	DROP	QuAC	SQuADv2	RACE-h	RACE-m
Fine-tuned SOTA	**90.7**[a]	**89.1**[b]	**74.4**[c]	**93.0**[d]	**90.0**[e]	**93.1**[e]
GPT-3 Zero-Shot	81.5	23.6	41.5	59.5	45.5	58.4
GPT-3 One-Shot	84.0	34.3	43.3	65.4	45.9	57.4
GPT-3 Few-Shot	85.0	36.5	44.3	69.8	46.8	58.1
	64%	93%	99%	94%	44%	

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

Figure G.20: Formatted dataset example for DROP

Context →	TITLE: William Perry (American football) - Professional career PARAGRAPH: In 1985, he was selected in the first round of the 1985 NFL Draft by the Chicago Bears; he had been hand-picked by coach Mike Ditka. However, defensive coordinator Buddy Ryan, who had a highly acrimonious relationship with Ditka, called Perry a "wasted draft-pick". Perry soon became a pawn in the political power struggle between Ditka and Ryan. Perry's "Refrigerator" nickname followed him into the NFL and he quickly became a favorite of the Chicago Bears fans. Teammates called him "Biscuit," as in "one biscuit shy of 350 pounds." While Ryan refused to play Perry, Ditka decided to use Perry as a fullback when the team was near the opponents' goal line or in fourth and short situations, either as a ball carrier or a lead blocker for star running back Walter Payton. Ditka stated the inspiration for using Perry as a fullback came to him during five-yard sprint exercises. During his rookie season, Perry rushed for two touchdowns and caught a pass for one. Perry even had the opportunity to run the ball during Super Bowl XX, as a nod to his popularity and contributions to the team's success. The first time he got the ball, he was tackled for a one-yard loss while attempting to throw his first NFL pass on a halfback option play. The second time he got the ball, he scored a touchdown (running over Patriots linebacker Larry McGrew in the process). About halfway through his rookie season, Ryan finally began to play Perry, who soon proved that he was a capable defensive lineman. His Super Bowl ring size is the largest of any professional football player in the history of the event. His ring size is 25, while the ring size for the average adult male is between 10 and 12. Perry went on to play for ten years in the NFL, retiring after the 1994 season. In his ten years as a pro, he regularly struggled with his weight, which hampered his performance at times. He played in 138 games, recording 29.5 sacks and five fumble recoveries, which he returned for a total of 71 yards. In his offensive career he ran five yards for two touchdowns, and had one reception for another touchdown. Perry later attempted a comeback, playing an unremarkable 1996 season with the London Monarchs of the World League of American Football (later NFL Europa). Q: what team did he play for? A:
Target Completion →	the Chicago Bears

Figure G.25: Formatted dataset example for QuAC

- QuAC is a **conversational** dataset
  - Requires a model to answer questions in a conversational context.
  - This requires the model to understand the context, including the previous questions and answers.
- GPT-3 is **unidirectional** and can only be read from left to right.
  - This means it cannot reread the text in front of it or see it behind it.

Comparisons

	SuperGLUE Average	BoolQ Accuracy	CB Accuracy	CB F1	COPA Accuracy	RTE Accuracy
Fine-tuned SOTA	**89.0**	**91.0**	**96.9**	**93.9**	**94.8**	**92.5**
Fine-tuned BERT-Large	69.0	77.4	83.6	75.7	70.6	71.7
GPT-3 Few-Shot	71.8	76.4	75.6	52.0	92.0	69.0

	WiC Accuracy	WSC Accuracy	MultiRC Accuracy	MultiRC F1a	ReCoRD Accuracy	ReCoRD F1
Fine-tuned SOTA	**76.1**	**93.8**	**62.3**	**88.2**	**92.5**	**93.3**
Fine-tuned BERT-Large	69.6	64.6	24.1	70.0	71.3	72.0
GPT-3 Few-Shot	49.4	80.1	30.5	75.4	90.2	91.1

COPA and WiC

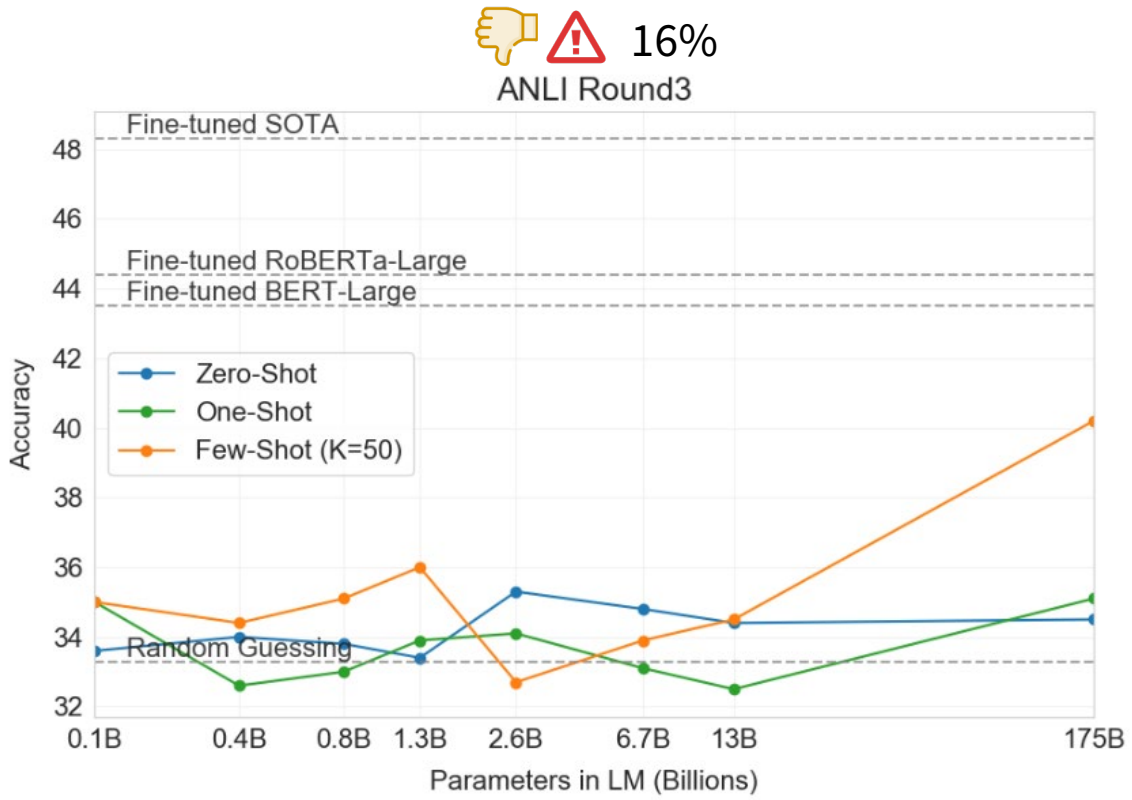
Context →	My body cast a shadow over the grass because
Correct Answer →	the sun was rising.
Incorrect Answer →	the grass was cut.

Figure G.5: Formatted dataset example for COPA

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

Figure G.32: Formatted dataset example for WiC

Comparisons



- WIC
  - involves comparing the use of a word in two sentences)
- ANLI
  - involves comparing two sentences to see if one implies the other

- Understanding the **contextual** information and **comparing** the two sentences is necessary to generate **accurate short** answers.

Hidden slide

Context →	anli 3: anli 3: We shut the loophole which has American workers actually subsidizing the loss of their own job. They just passed an expansion of that loophole in the last few days: $43 billion of giveaways, including favors to the oil and gas industry and the people importing ceiling fans from China. Question: The loophole is now gone True, False, or Neither?
Correct Answer →	False
Incorrect Answer →	True
Incorrect Answer →	Neither

Figure G.10: Formatted dataset example for ANLI R3

Hidden slide

Context →	READING COMPREHENSION ANSWER KEY While this process moved along, diplomacy continued its rounds. Direct pressure on the Taliban had proved unsuccessful. As one NSC staff note put it, "Under the Taliban, Afghanistan is not so much a state sponsor of terrorism as it is a state sponsored by terrorists." In early 2000, the United States began a high-level effort to persuade Pakistan to use its influence over the Taliban. In January 2000, Assistant Secretary of State Karl Inderfurth and the State Department's counterterrorism coordinator, Michael Sheehan, met with General Musharraf in Islamabad, dangling before him the possibility of a presidential visit in March as a reward for Pakistani cooperation. Such a visit was coveted by Musharraf, partly as a sign of his government's legitimacy. He told the two envoys that he would meet with Mullah Omar and press him on Bin Laden. They left, however, reporting to Washington that Pakistan was unlikely in fact to do anything," given what it sees as the benefits of Taliban control of Afghanistan." President Clinton was scheduled to travel to India. The State Department felt that he should not visit India without also visiting Pakistan. The Secret Service and the CIA, however, warned in the strongest terms that visiting Pakistan would risk the President's life. Counterterrorism officials also argued that Pakistan had not done enough to merit a presidential visit. But President Clinton insisted on including Pakistan in the itinerary for his trip to South Asia. His one-day stopover on March 25, 2000, was the first time a U.S. president had been there since 1969. At his meeting with Musharraf and others, President Clinton concentrated on tensions between Pakistan and India and the dangers of nuclear proliferation, but also discussed Bin Laden. President Clinton told us that when he pulled Musharraf aside for a brief, one-on-one meeting, he pleaded with the general for help regarding Bin Laden." I offered him the moon when I went to see him, in terms of better relations with the United States, if he'd help us get Bin Laden and deal with another issue or two." The U.S. effort continued. Who did The State Department feel should visit both India and Pakistan?
Correct Answer →	- [False] Bin Laden
Incorrect Answer →	- [True] Bin Laden

Figure G.15: Formatted dataset example for MultiRC. There are three levels within MultiRC: (1) the passage, (2) the questions, and (3) the answers. During evaluation, accuracy is determined at the per-question level, with a question being considered correct if and only if all the answers within the question are labeled correctly. For this reason, we use K to refer to the number of **questions** shown within the context.

Hidden slide

SAT Analogie

Context →	lull is to trust as
Correct Answer →	cajole is to compliance
Incorrect Answer →	balk is to fortitude
Incorrect Answer →	betray is to loyalty
Incorrect Answer →	hinder is to destination
Incorrect Answer →	soothe is to passion

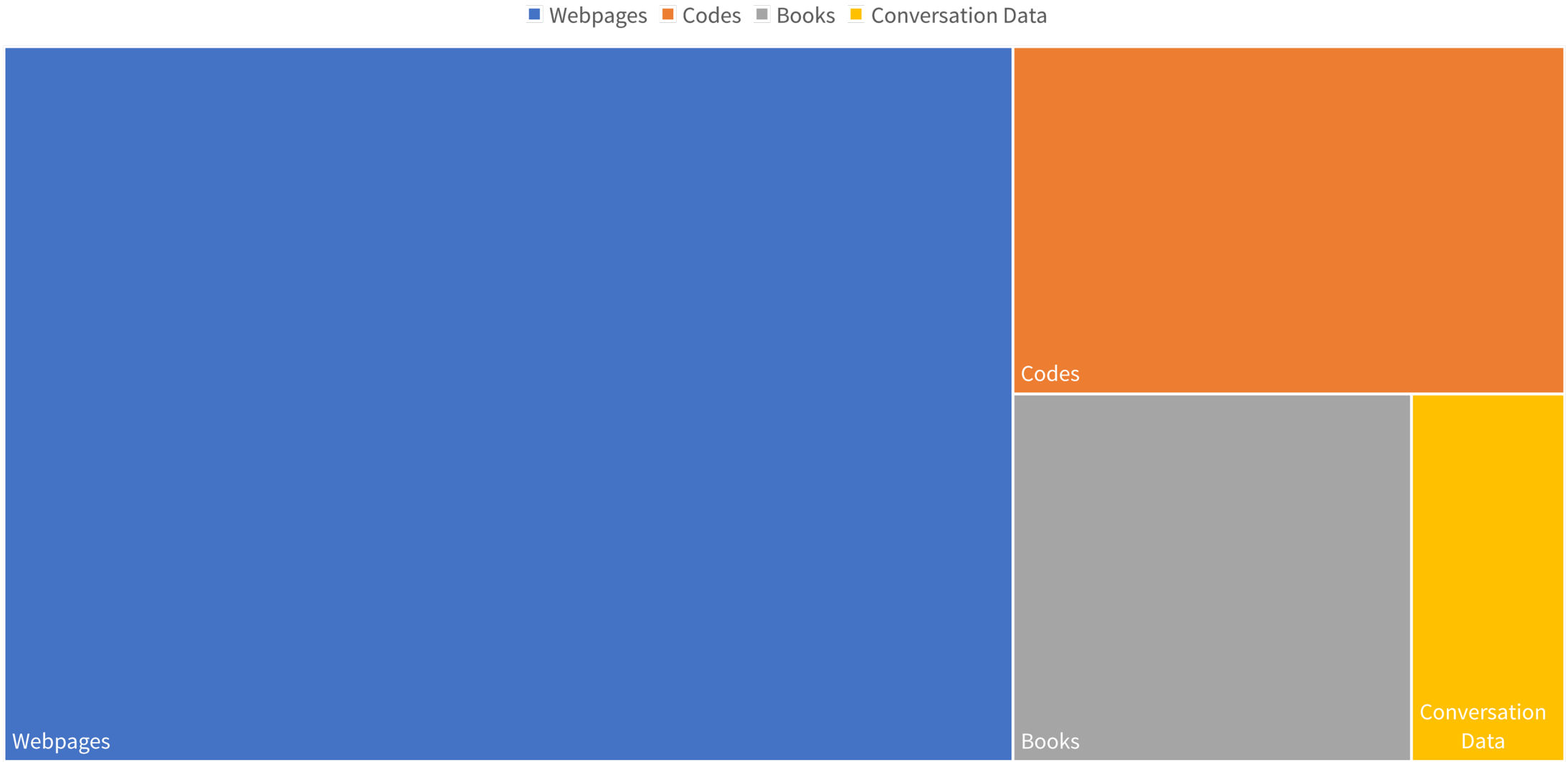
Figure G.12: Formatted dataset example for SAT Analogies

Data

Dataset	Quantity (tokens)	Weight in training mix	Epochs elapsed when training for 300B tokens
Common Crawl (filtered)	410 billion	60%	0.44
WebText2	19 billion	22%	2.9
Books1	12 billion	8%	1.9
Books2	55 billion	8%	0.43
Wikipedia	3 billion	3%	3.4

- Main dataset: (a subset of) Common Crawl 570GB
  - From 2016 to 2019, 41 shards.
  - A binary classifier is trained to predict WebText vs. Common Crawl.
  - If the classifier thinks it is more similar to WebText, it samples (retains) the document with a higher probability.
- Other datasets
  - To make the data higher quality, four other (higher quality) datasets were also put into the training data.

Ratios of various data sources in the pre-training data for



Adapted from W.Zhao et al., A Survey of Large Language Models, 2023

Hidden slide

Data

Dataset	Quantity (tokens)	Weight in training mix	Epochs elapsed when training for 300B tokens
Common Crawl (filtered)	410 billion	60%	0.44
WebText2	19 billion	22%	2.9
Books1	12 billion	8%	1.9
Books2	55 billion	8%	0.43
Wikipedia	3 billion	3%	3.4

The paper, CS342

Hidden slide

Data

Dataset	Quantity (tokens)	Weight in training mix	Epochs elapsed when training for 300B tokens
Common Crawl (filtered)	410 billion	60%	0.44
WebText2	19 billion	22%	2.9
Books1	12 billion	8%	1.9
Books2	55 billion	8%	0.43
Wikipedia	3 billion	3%	3.4

- Main dataset: (a subset of) Common Crawl
  - From 2016 to 2019, 41 shards.
  - A binary classifier is trained to predict WebText vs. Common Crawl.
  - If the classifier thinks it is more similar to WebText, it samples (retains) the document with a higher probability.
- Fuzzily deduplicated
  - Resulting in a 10% reduction in data volume
- Other datasets
  - To make the data higher quality, four other (higher quality) datasets were also put into the training data.
- Data contamination
  - 13 grams = data overlap
- Same data for all models

Hidden slide

Data Contamination

We initially tried to address the issue of contamination by proactively searching for and attempting to remove any overlap between our training data and the development and test sets of all benchmarks studied in this paper. Unfortunately, a bug resulted in only partial removal of all detected overlaps from the training data. Due to the cost of training, it wasn't feasible to retrain the model. To address this, we investigate in detail how the remaining detected overlap impacts results.

Hidden slide

Data Contamination

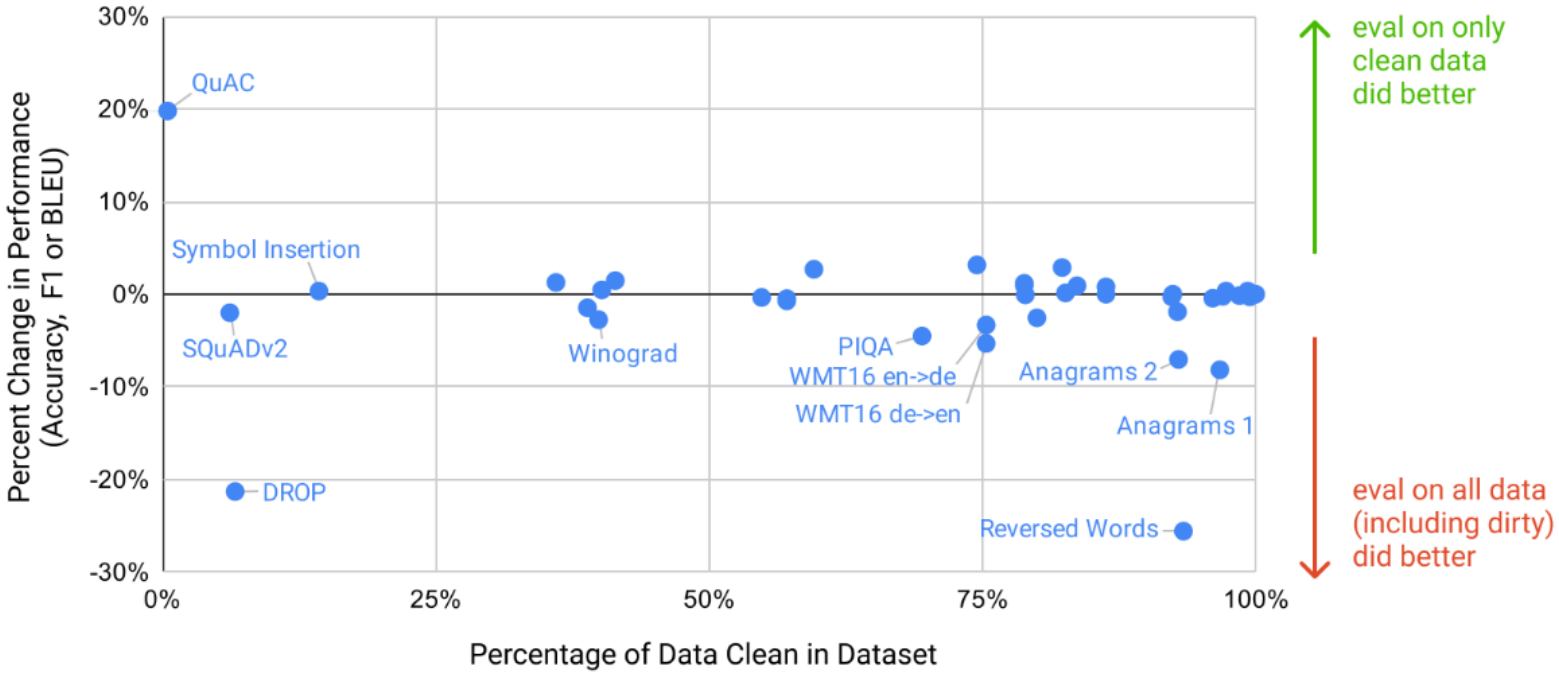
- Select data sets that are not exposed to the Internet(Penn Tree Bank)
- Clean version of datasets
- A whole section to analyze the impact of data contamination
  - The authors do not consider the impact to be significant

Hidden slide

Clean version of datasets

- Reverse the data-cleaning process
  - Clean the test set instead of cleaning the training set
- Strict standards are used
  - 13-gram = overlap

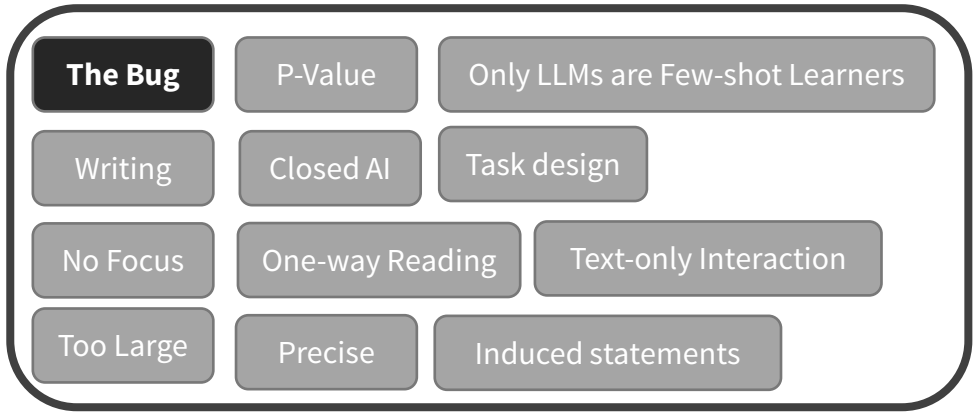
Hidden slide



Hidden slide

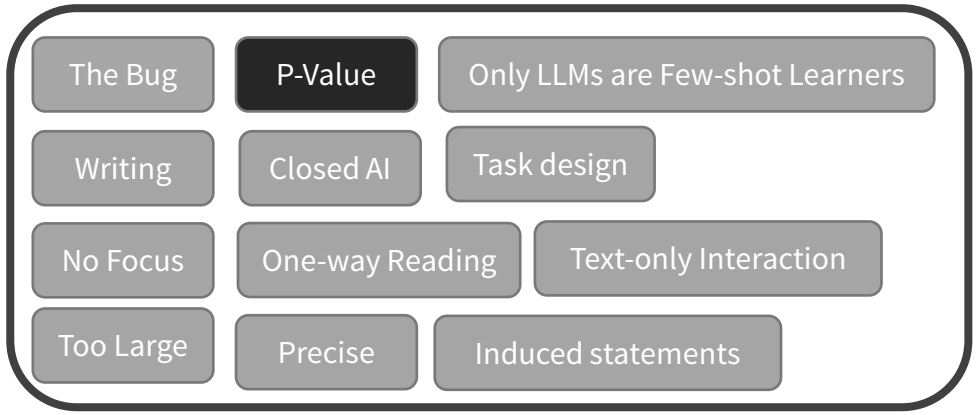
- DROP
  - **-21% Difference**
  - The training data only contains questions, not answers.
  - A more likely explanation for the performance drop is that the remaining 6% of filtered examples come from a slightly different distribution than the dirty ones.
- RW
  - **-26% Difference**
  - The rigorous overlap analysis resulted in the inclusion of much of the underlying information as overlap as well, which is critical for problem solving.
- LAMBADA
  - **0% Difference**

Shortcomings



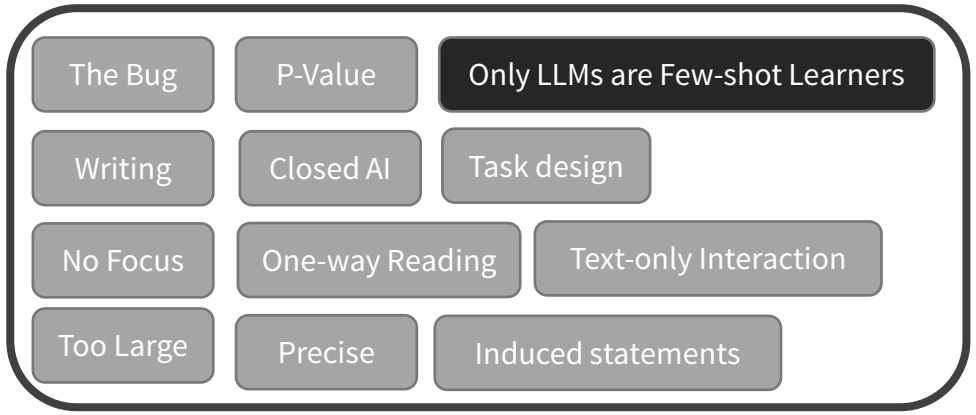
- The bug.
  - This was an irreversible mistake that left 90% of the downstream tasks marked as contaminated.
  - They could have had many fewer pages.
  - Only two datasets are not affected. Due to the data amount, 1% matters.
- The p-value.
- Only large models are few-shot learners.

Shortcomings



- The bug.
  - This was an irreversible mistake that left 90% of the downstream tasks marked as contaminated.
  - They could have had many fewer pages.
  - Only two datasets are not affected. Due to the data amount, 1% matters.
- The p-value.
  - They use the p-value
- Only large models are few-shot learners.

Shortcomings



- The bug.
  - This was an irreversible mistake which left 90% of the downstream tasks marked as contaminated.
  - They could have had many fewer pages.
  - Only two datasets are not affected. Due to the data amount, 1% matters.
- The p-value.
  - They use the p-value
- Only large models are few-shot learners.
  - The authors do not specifically emphasize this point, even though seven images in the paper prove it.

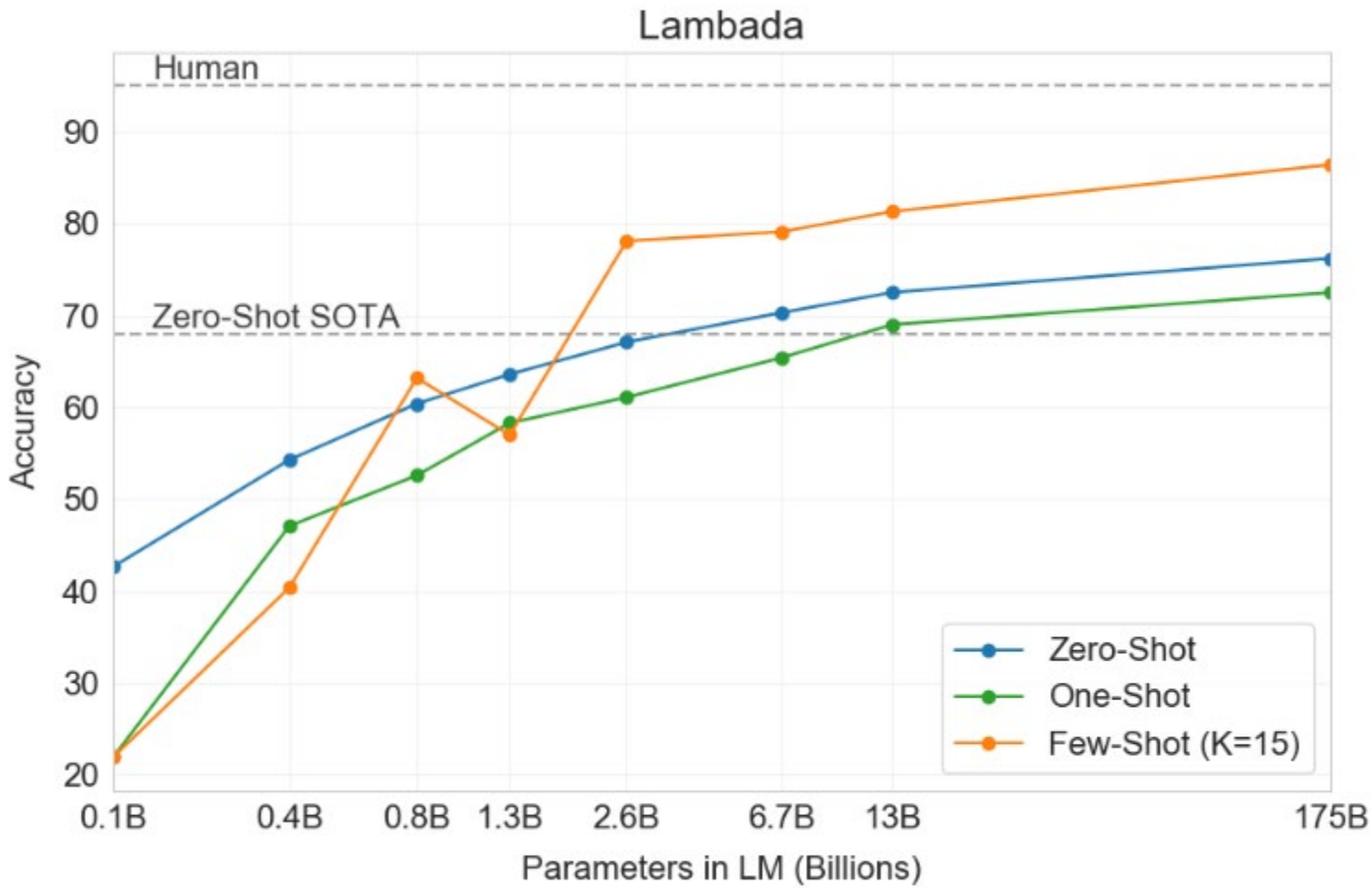
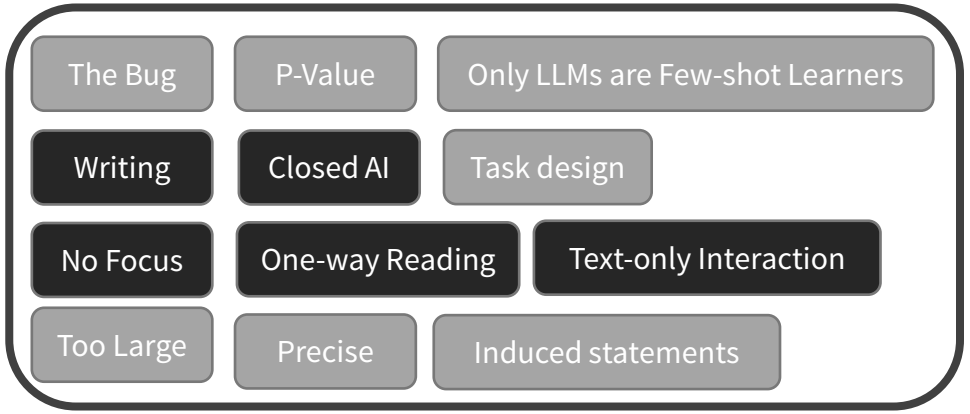


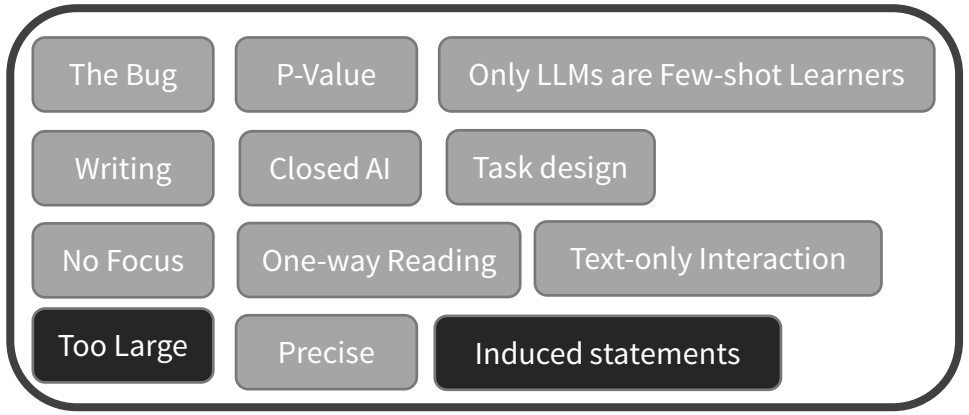
Figure 1.2/3.2/ 3.5/3.6/3.7/3.8/3.12/…

Shortcomings



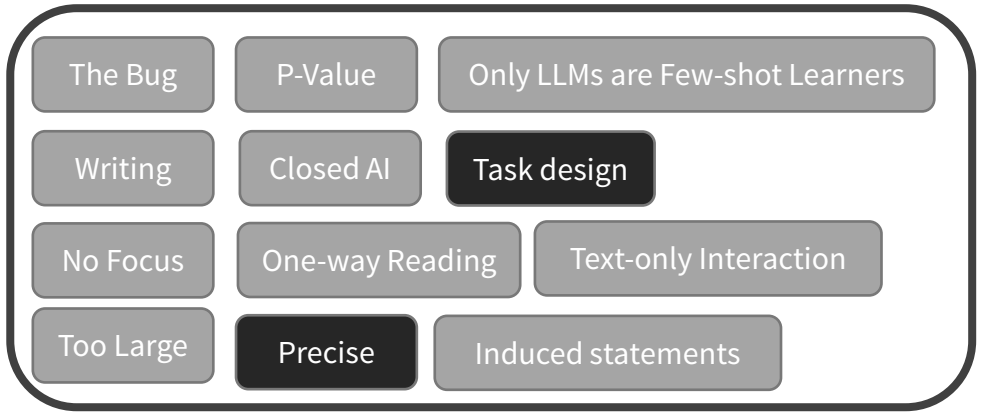
- Writing style is not suitable for beginners.
  - They did not discuss their architecture in detail, with no figures or parameter numbers.
  - They wrote a paper of 41 pages and use two sentences to describe the model itself.
  - They assume that the audience already read their old papers rather than do a brief introduction to the method they are using.
- Closed AI
- Learning lacks focus
  - We learn math and foreign languages from textbooks, which we know are credible and well-designed, so the points in them are essential and easy to learn.
  - Instruct-GPT solves this problem in part through human feedback
- Communicate only through text, not video or audio.
- Cannot retain contextual information in conversations.

Shortcomings



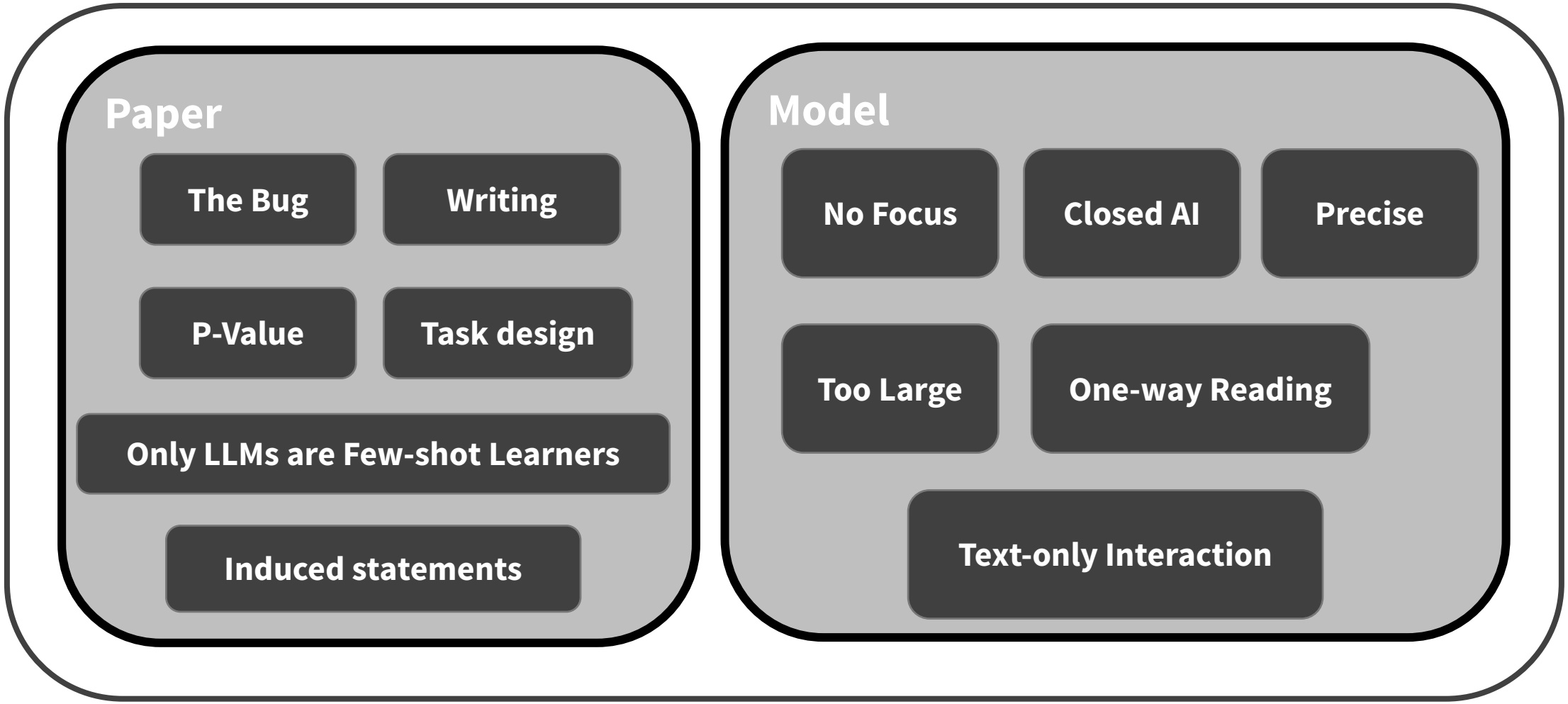
- Too large to convince me.
  - Although the GPT-3 should be smaller than or equal to the whole dataset, it is still too early to say the model understands the natural languages due to the performance on multiple tasks. (GPT-2 XL has a 40GB dataset and only takes up 6.45GB of hard drive space. So GPT-3 should be around 650GB)
- When they talk about the data contamination of PIQA, they claim this dataset was published after they trained the model.
  - This implies that the model's performance will not be affected, which is untrue. Reduces the credibility of the article.

Shortcomings



- Tasks can be better designed to test the understanding of the model
- Difficult to generate precise answers
  - Intelligent models should be able to generate at least the original content (original text) in the training data, not the vague content. If only vague content can be generated proves their complete lack of comprehension.
- Still a far cry from the way humans think
  - Humans don't need the entire Internet to learn a skill
  - The human brain has a meager power consumption.

Shortcomings



Future progress

- A more balanced and efficient **dataset**.
  - The Common Crawl over-records statements from young users in developed countries.
  - More than half of the users in the Reddit forums are young men.
  - 8.8%-15% of Wikipedia users are female.
- **Clean** the datasets more carefully.
  - K. Lee et al.(2021) propose a method to efficiently remove duplicates from large databases.
- Other forms of **interaction**
  - RLHF introduced human feedback.
  - Video, audio.

Conclusion

- GPT-3 **exceeds SOTA** on several tasks.
  - Language Modeling [PTB, LAMBADA]
  - Translation from other languages to English
  - News article generation
- GPT-3 has a significant performance gap in different tasks.
  - The GPT-3 does not perform well in reading comprehension, especially when there is a comparison (of two words or sentences).
  - GPT-3 cannot think abstractly and logically. [ARC, Arithmetic]
  - GPT-3 performance is related to whether the required output frequently appears online. [Translation]
  - GPT-3 performs poorly on long text generation. [QA]
- Only **Large** LMs are few-shot learners
  - On smaller models, Few-Shot does not have a significant effect.

Takeaways

- **Decoder** from Transformer
  - The GPT series all use Only-Decoder as the infrastructure, but they did not first propose it.
- Transfer**(Meta) Learning** from CV
  - GPT-3 demonstrates that as long as the model is large enough, we can directly instruct the model to perform tasks **using natural language without fine-tuning**.
- **Money** from Microsoft
  - The GPT-3 has 170B params, which used to be the world's largest and most expensive model.
- **Mistakes** from Unicorn
  - People(even those who work for OpenAI) do make mistakes(for multiple times). Avoid it with the right process/workflow.

Sorry, OpenAI… But you should do better

gpt-2

Code and models from the paper "Language Models are Unsupervised Multitask Learners".

You can read about GPT-2 and its staged release in our original blog post, 6 month follow-up post, and final post.

We have also released a dataset for researchers to study their behaviors.

* *Note that our original parameter counts were wrong due to an error (in our previous blog posts and paper). Thus you may have seen small referred to as 117M and medium referred to as 345M.*

Reference

[1.4] Bengio, Y., Ducharme, R., Vincent, P., & Janvin, C. (2003). A Neural Probabilistic Language Model. *J. Mach. Learn. Res., 3, 1137–1155.*

[1.5] Bahdanau, D., Cho, K., & Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

[1.6] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, & Illia Polosukhin (2017). Attention Is All You Need. *CoRR, abs/1706.03762.*

[1.7] Peter J. Liu, Mohammad Saleh, Etienne Pot, Ben Goodrich, Ryan Sepassi, Lukasz Kaiser, & Noam Shazeer (2018). Generating Wikipedia by Summarizing Long Sequences. *CoRR, abs/1801.10198.*

[1.8] Jacob Devlin, Ming-Wei Chang, Kenton Lee, & Kristina Toutanova (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *CoRR, abs/1810.04805.*

[1.9] Radford, A., & Narasimhan, K. (2018). Improving Language Understanding by Generative Pre-Training.

[1.10] Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language Models are Unsupervised Multitask Learners.

[1.11] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, & Dario Amodei (2020). Language Models are Few-Shot Learners. *CoRR, abs/2005.14165.*

[2] F. Huber, “King -Man +Woman = King ?,” *Medium*, Jul. 15, 2019. https://blog.esciencecenter.nl/king-man-woman-king-9a7fd2935a85 (accessed Jun. 16, 2023).

Reference

Li, M. (2022). *GPT， GPT-2， GPT-3.* www.youtube.com. https://www.youtube.com/watch?v=t70Bl3w7bxY&t

CS324 2023 Percy Liang et al. https://stanford-cs324.github.io/winter2022

I. Magar and R. Schwartz, “Data Contamination: From Memorization to Exploitation,” arXiv.org, Mar. 15, 2022. https://arxiv.org/abs/2203.08242 (accessed Jun. 13, 2023).

R. L. Wasserstein and N. A. Lazar, “The ASA Statement on p-Values: Context, Process, and Purpose,” The American Statistician, vol. 70, no. 2, pp. 129–133, Apr. 2016, doi: https://doi.org/10.1080/00031305.2016.1154108.

P. Pitak-Arnnop, K. Dhanuthai, A. Hemprich, and N. C. Pausch, “Misleading p-value:do you recognise it?,” European journal of dentistry, vol. 4, no. 3, pp. 356–8, 2010, Accessed: Feb. 09, 2023. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2897873/

Alammar, J. (2018, 2019, 2020). *The Illustrated Transformer/GPT-2/GPT-3*. Jalammar.github.io.

Raissi, M. (2021). Open Problems in Applied Deep Learning*. arXiv preprint arXiv:2301.11316*.

Karpathy, A. (2023, May 19). CS25 I Stanford Seminar - Transformers United 2023: Introduction to Transformers w/ Andrej Karpathy. Retrieved June 19, 2023, from YouTube, website: https://www.youtube.com/watch?v=XfpMkf4rD6E

Dean, J. (2021, October 28). Introducing Pathways: A next-generation AI architecture. Retrieved from Google website: https://blog.google/technology/ai/introducing-pathways-next-generation-ai-architecture/

[3] Zhang, A., Lipton, Z., Li, M., & Smola, A. (2021). Dive into Deep Learning*. arXiv preprint arXiv:2106.11342*.

[4] He, K., Zhang, X., Ren, S., and Sun, J. Identity mappings in deep residual networks. In European conference on computer vision, pp. 630–645. Springer, 2016.

[5] Chelsea Finn, Pieter Abbeel, & Sergey Levine. (2017). Model-Agnostic Meta-Learning for Fast Adaptation of Deep Networks.

[6] Katherine Lee, Daphne Ippolito, Andrew Nystrom, Chiyuan Zhang, Douglas Eck, Chris Callison-Burch, & Nicholas Carlini (2021). Deduplicating Training Data Makes Language Models Better. CoRR, abs/2107.06499.

I'm wearing a bulletproof vest today.

Question, please!

It's okay to criticize

Hidden slide

Clarification – What I didn’t cover

- Limitations part of the paper.
- Broader Impacts of the paper.
  - Misuse of LMs
  - Fairness, Bias, and Representation
  - Energy Usage
- Tasks
  - Word scrambling and Manipulation Tasks
    - Some require letter-level manipulation capabilities, and the data is primarily contaminated.
    - Most of them do not require the model's understanding of the language and can be solved by a simple similarity judgment.

Please note that this page of slides is prepared to avoid misunderstandings after this slide is online and will not be shown in the presentation.

Hidden slide

Clarification – What I oversimplified

- Tokens are not equal to words. -> Check “Word Piece”
  - Shawshank => “WP” => Shaw+sh+ank
- Transformer GPT-3 was used is called the Sparse Transformer. It is slightly different from the decoder-only transformer.
  - Rewon Child, Scott Gray, Alec Radford, and Ilya Sutskever. Generating long sequences with sparse transformers, 2019.

Please note that this page of slides is prepared to avoid misunderstandings after this slide is online and will not be shown in the presentation.