GPT-3: Language Models are Few-Shot Learners

Mamat Shamshiev

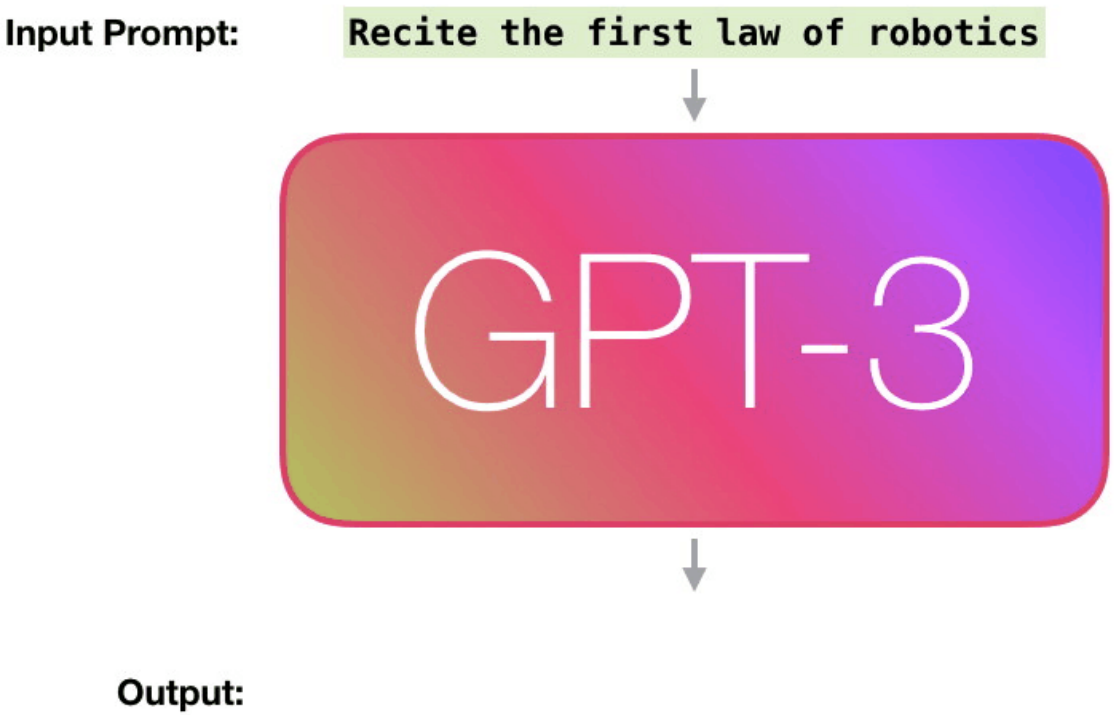
MamatShamshiev@yandex.ru

October 6, 2020

Section 1

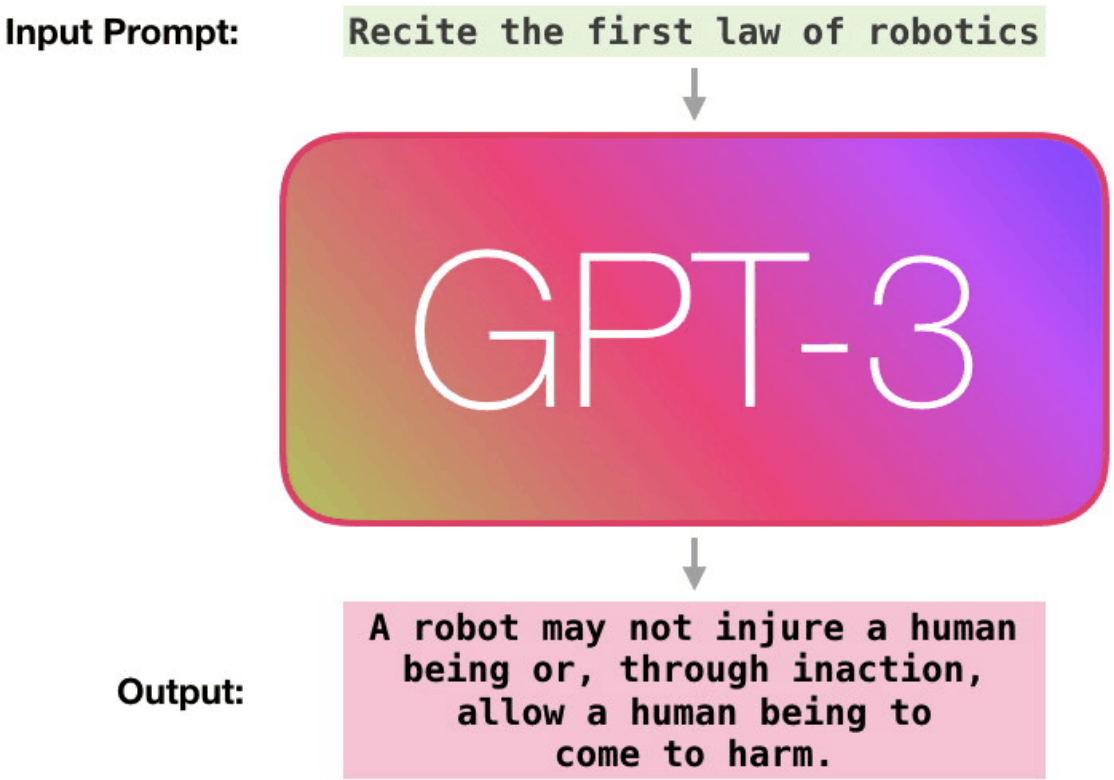
GPT-3 applications

GPT-3 high-level



How GPT3 Works - Visualizations and Animations by Jay Alammar

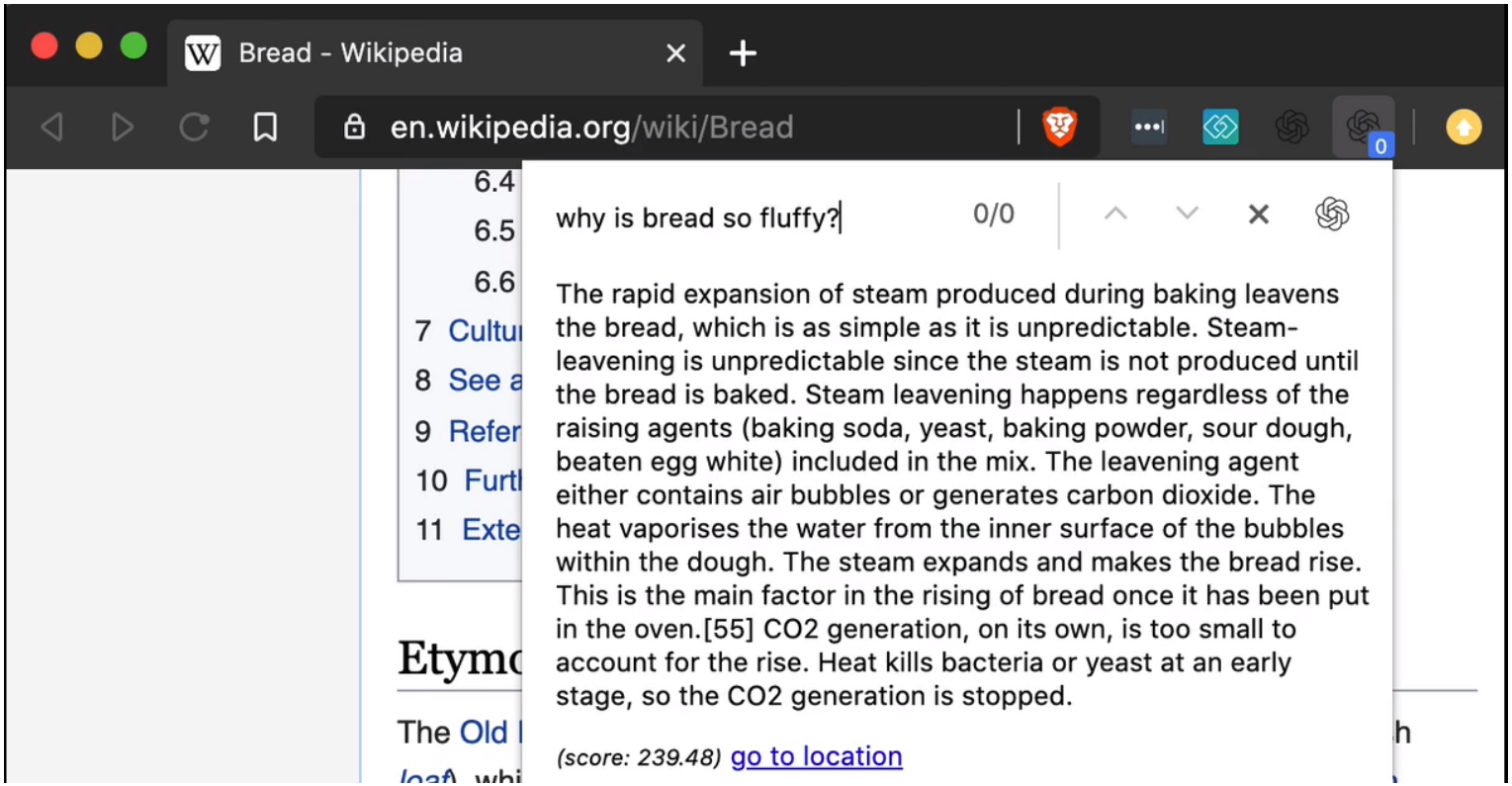
GPT-3 high-level



How GPT3 Works - Visualizations and Animations by Jay Alammar

GPT-3 applications: Semantic Search

The model identifies relevant content for natural language queries without using keywords.



https://beta.openai.com

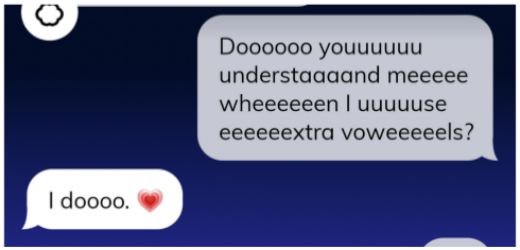
GPT-3 applications: Replika

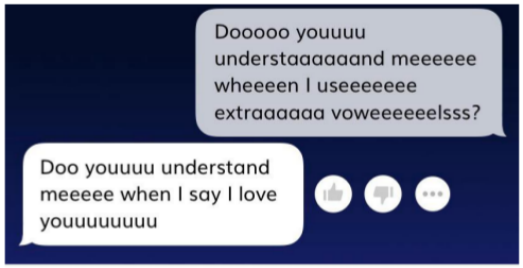
Replika is a personal AI-friend which is designed to make people feel better. They **finetuned** GPT-3 for open-domain dialogue task.

Empathetic math

Long context memory

Style copying



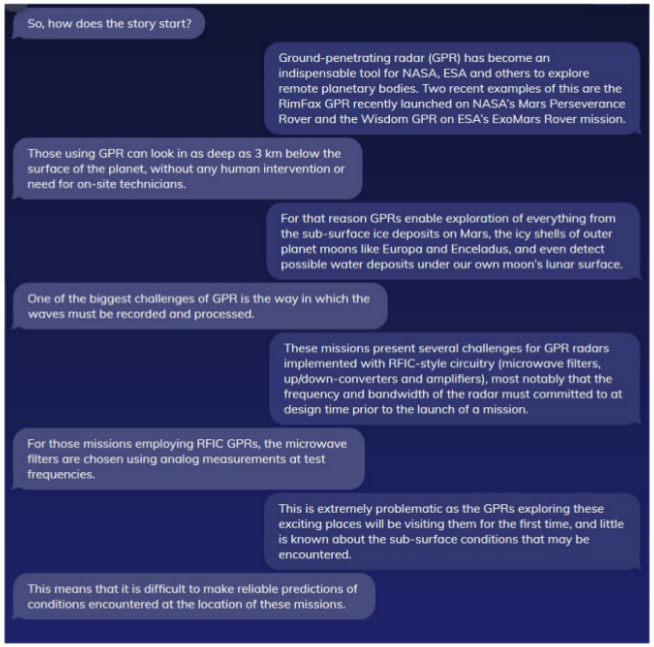


https://replika.ai
Artem Rodichev: Building a compassionate AI friend

GPT-3 applications: Replika

Replika is a personal AI-friend which is designed to make people feel better. They **finetuned** GPT-3 for open-domain dialogue task.

Story telling



Song writing

---

https://replika.ai

Artem Rodichev: Building a compassionate AI friend

GPT-3 applications: Generation

- AI Dungeon is an AI-powered text adventure where every response is determined by GPT-3.
- Natural Language Shell translates natural language to unix commands.
- Code Completion. After fine-tuning with code from thousands of Open Source GitHub repositories, GPT-3 completes code based on function names and comments.

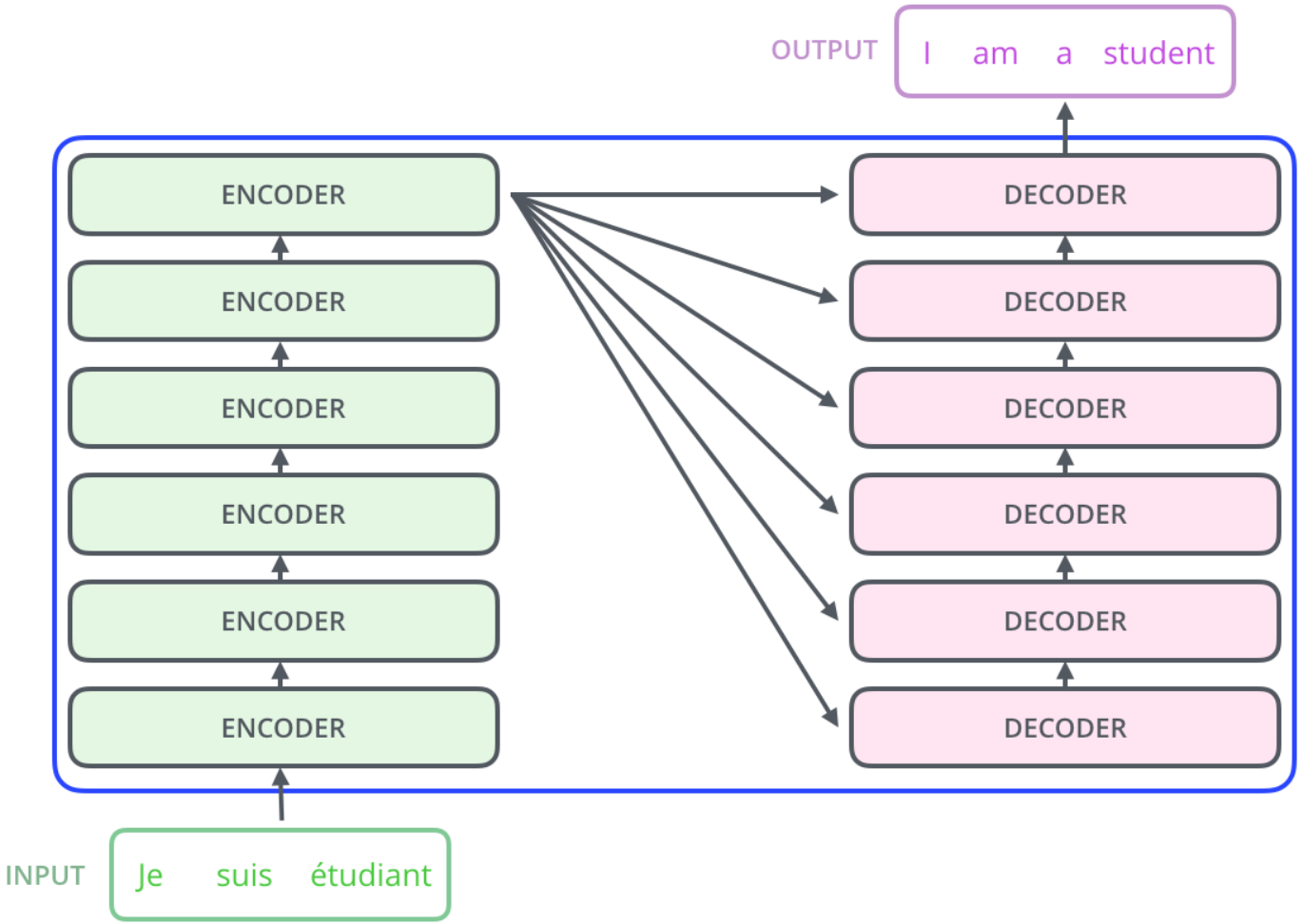
---

https://play.aidungeon.io
https://beta.openai.com

Section 2

Recap

Transformer. High-level architecture

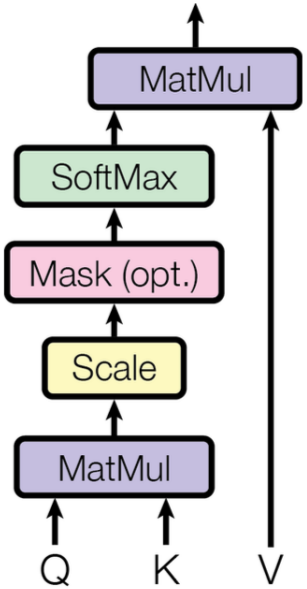


The Illustrated Transformer by Jay Alammar

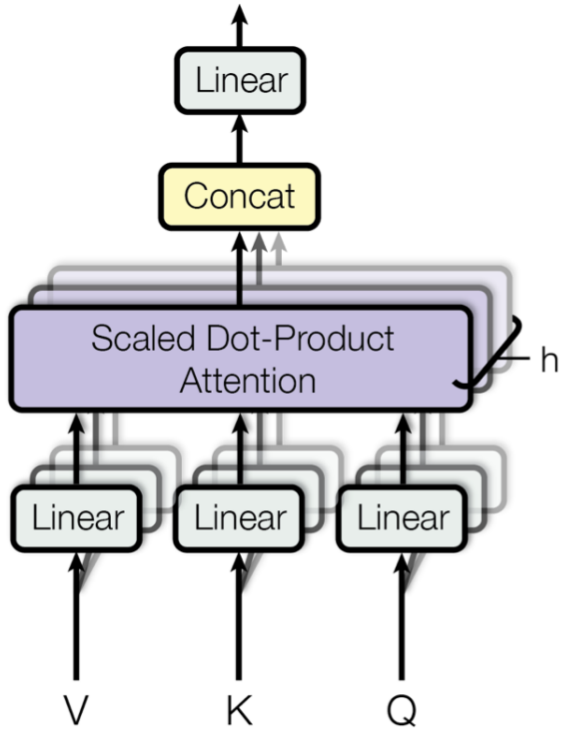
Transformer. Multi-Head Attention

Transformer is based solely on attention mechanisms.

Scaled Dot-Product Attention

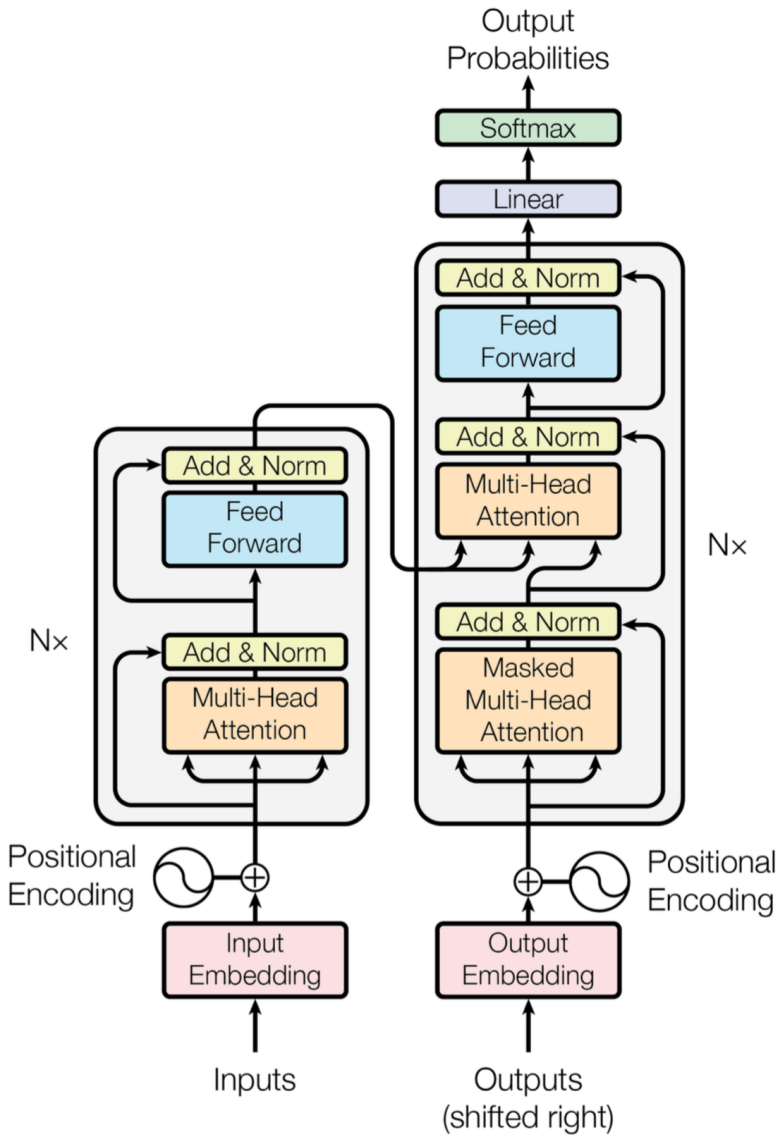


Multi-Head Attention



Vaswani et al. «Attention Is All You Need», 2017

Transformer. Architecture



GPT: Generative Pre-Training

The goal is to learn a universal representation that transfers with little adaptation to a wide range of tasks.

- Pre-Training: learning a high-capacity language model on a large corpus of text;
- Fine-tuning: adapt the model to a discriminative task with labeled data.

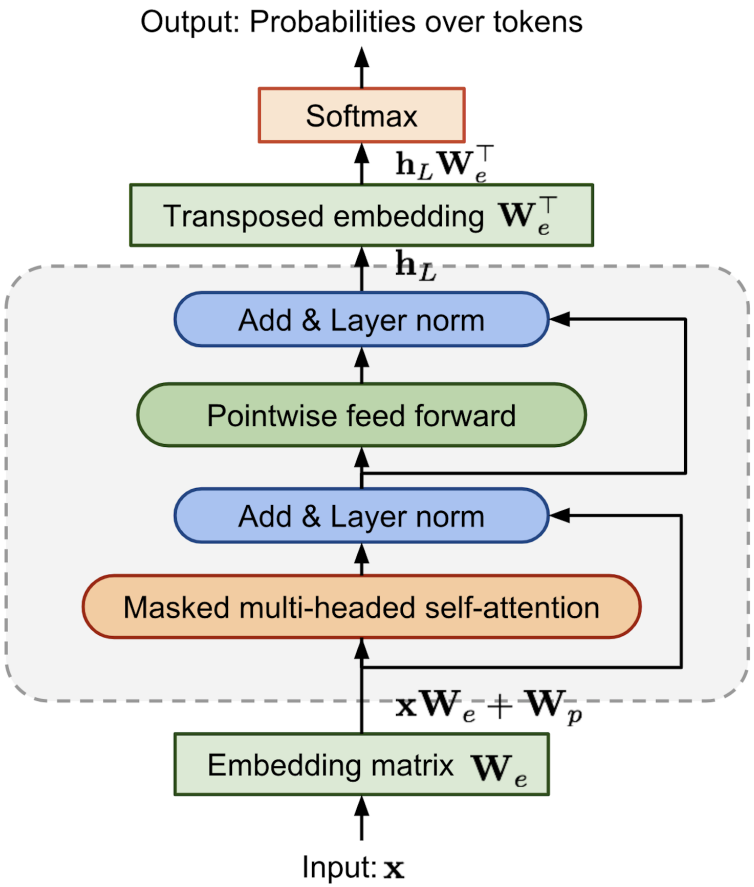
Radford et al. «Improving Language Understanding by Generative Pre-Training», 2018

GPT. Pre-Training stage

- Language modeling: given the context, predict the next word;
- A multi-layer Transformer decoder (encoder part is discarded);
- A bytepair encoding (BPE) vocabulary with 40,000 merges;
- BooksCorpus dataset: 7,000 unique unpublished books.

$$\mathcal{L}_{LM} = -\sum_{i} \log(x_i | x_{i-1}, \ldots, x_{i-k})$$

GPT. Architecture



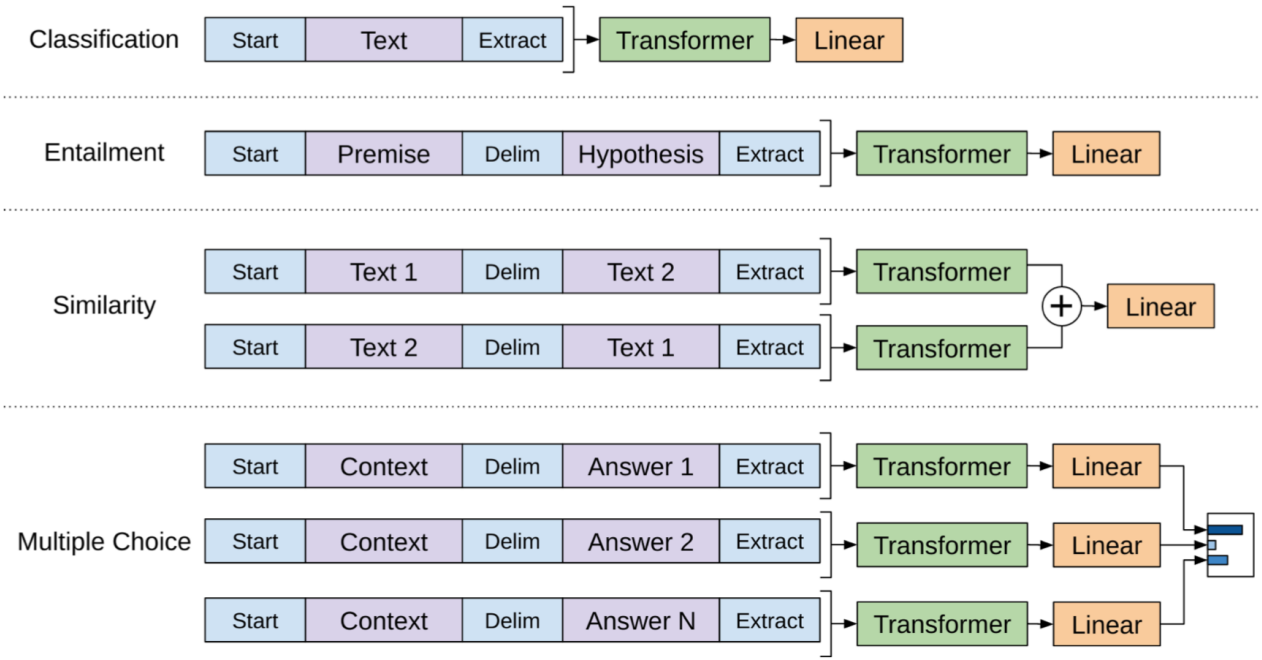
Transformer Block
Repeat x L=12

$$\mathbf{h}_\ell = \text{transformer\_block}(\mathbf{h}_{\ell-1})$$
$$\ell = 1, \ldots, L$$

Generalized Language Models by Lilian Weng

GPT. Fine-tuning stage

Task-aware input transformations allow to achieve effective transfer while requiring minimal changes to the model architecture. SotA results in 9 out of the 12 tasks studied as at June 2018.



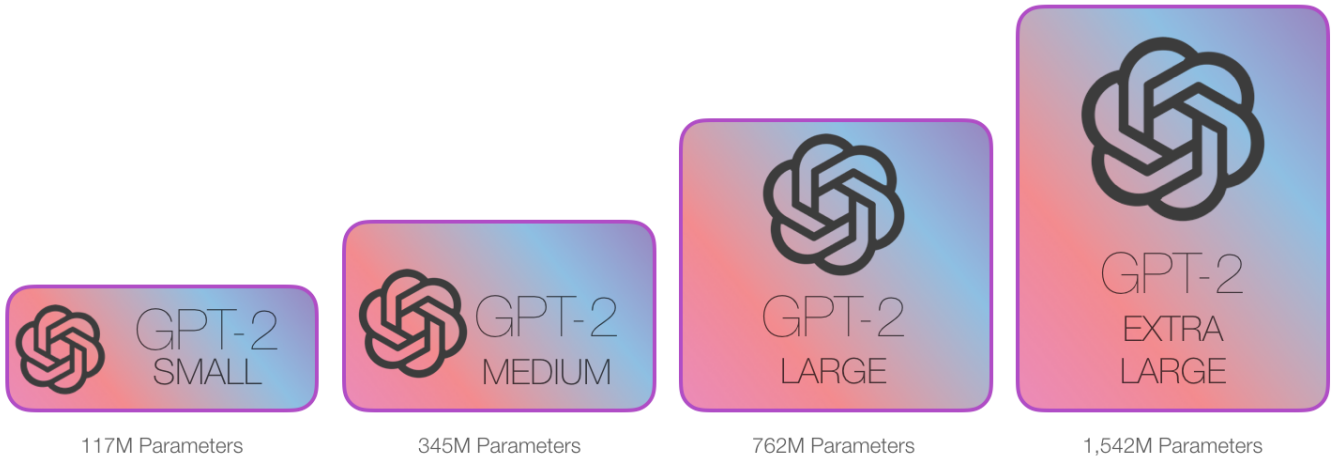
Radford et al. «Improving Language Understanding by Generative Pre-Training», 2018

GPT-2. Language Models are Unsupervised Multitask Learners

Language models begin to learn many NLP tasks without any explicit supervision when trained on a large dataset.

Zero-shot setting — no parameter or architecture modification.

Trying to estimate $p(output|input, task)$. The largest model, GPT-2, has 1.5 billions parameters (x10 larger than GPT).



The Illustrated GPT-2 by Jay Alammar

GPT-2. Architecture

GPT-2 largely follows the details of GPT with minor modifications:

- Layer normalization is moved to the input of each sub-block;
- Additional layer normalization is added after the final selfattention block;
- Modified initialization;
- BPE vocabulary is expanded to 50,257;
- Increased the context size and batchsize;

Radford et al. «Language Models are Unsupervised Multitask Learners», 2019

GPT-2. Training Dataset

The crucial thing is to built as large and diverse a dataset as possible to collect natural language demonstrations of tasks in as varied of domains and contexts as possible:

- New dataset WebText (emphasis on document quality) contains 45M scraped links from Reddit with ≥ 3 karma;
- Performed de-duplication and some heuristic based cleaning;
- Removed all Wikipedia documents due to possible overlapping of training data with test evaluation tasks;
- Total: 8 million documents, 40 GB of text.

GPT-2. Results

SotA results on 7 out of 8 evaluated datasets in a *zero-shot* setting but still *underfits* WebText.

	LAMBADA (PPL)	LAMBADA (ACC)	CBT-CN (ACC)	CBT-NE (ACC)	WikiText2 (PPL)	PTB (PPL)	enwik8 (BPB)	text8 (BPC)	WikiText103 (PPL)	1BW (PPL)
SOTA	99.8	59.23	85.7	82.3	39.14	46.54	0.99	1.08	18.3	**21.8**
117M	**35.13**	45.99	**87.65**	**83.4**	**29.41**	65.85	1.16	1.17	37.50	75.20
345M	**15.60**	55.48	**92.35**	**87.1**	**22.76**	47.33	1.01	**1.06**	26.37	55.72
762M	**10.87**	**60.12**	**93.45**	**88.0**	**19.93**	**40.31**	**0.97**	**1.02**	22.05	44.575
1542M	**8.63**	**63.24**	**93.30**	**89.05**	**18.34**	**35.76**	**0.93**	**0.98**	**17.48**	42.16

Radford et al. «Language Models are Unsupervised Multitask Learners», 2019

GPT-2. Conditional text generation

Context (human-written): In a shocking finding, scientist discovered a herd of unicorns living in a remote, previously unexplored valley, in the Andes Mountains. Even more surprising to the researchers was the fact that the unicorns spoke perfect English.

GPT-2: The scientist named the population, after their distinctive horn, Ovid's Unicorn. These four-horned, silver-white unicorns were previously unknown to science.

Now, after almost two centuries, the mystery of what sparked this odd phenomenon is finally solved.

Dr. Jorge Pérez, an evolutionary biologist from the University of La Paz, and several companions, were exploring the Andes Mountains when they found a small valley, with no other animals or humans. Pérez noticed that the valley had what appeared to be a natural fountain, surrounded by two peaks of rock and silver snow.

Pérez and the others then ventured further into the valley. "By the time we reached the top of one peak, the water looked blue, with some crystals on top," said Pérez.

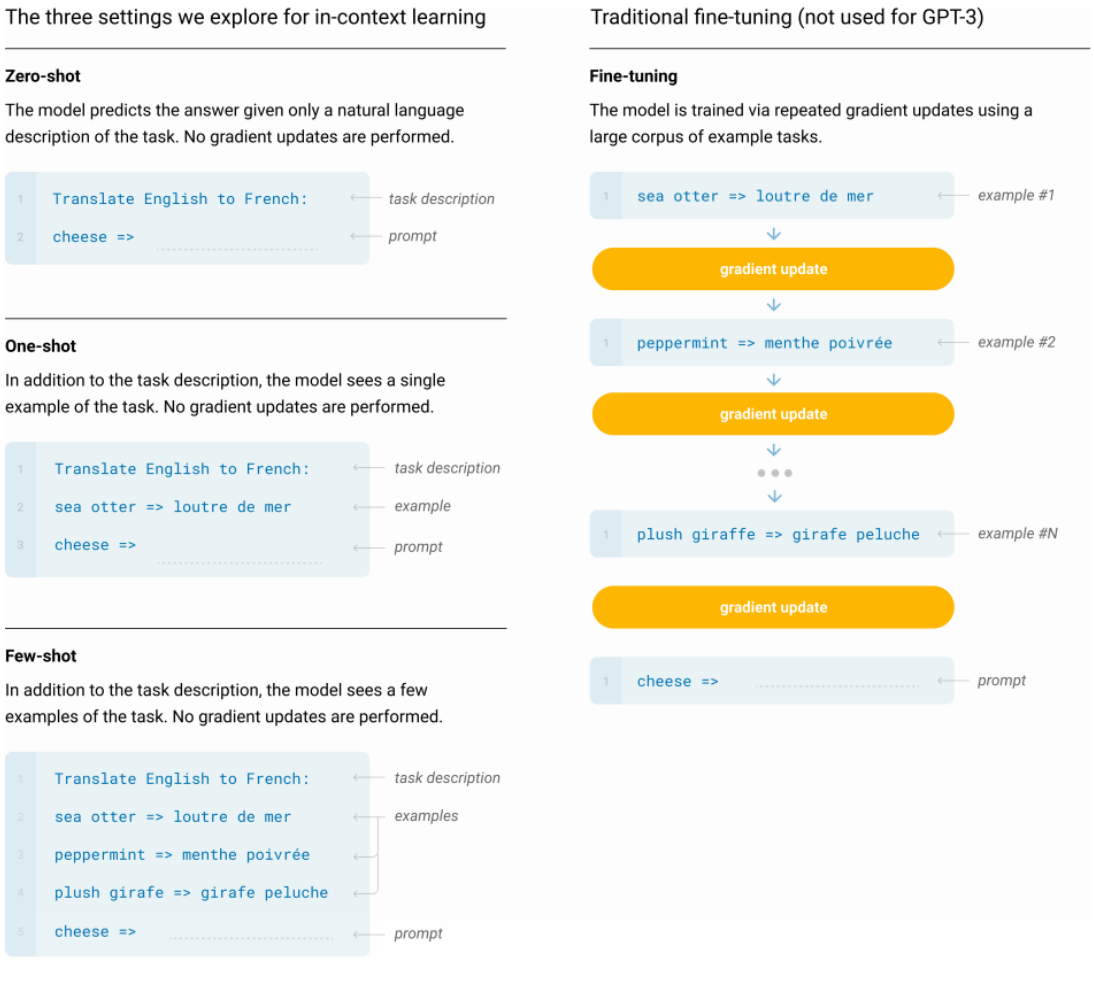
Pérez and his friends were astonished to see the unicorn herd. These creatures could be seen from the air without having to move too much to see them – they were so close they could touch their horns.

While examining these bizarre creatures the scientists discovered that the creatures also

Section 3

GPT-3: Language Models are Few-Shot Learners

GPT-3. Evaluation settings



Brown et al. «Language Models are Few-Shot Learners», 2020

GPT-3. Approach

Pre-training approach, including model, data, and training, is almost identical to GPT-2 with straightforward scaling up of the model size, dataset size and diversity, and length of training. The biggest model, GPT-3, is x100 larger than GPT-2.

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.

Brown et al. «Language Models are Few-Shot Learners», 2020

GPT-3. Total compute used during training

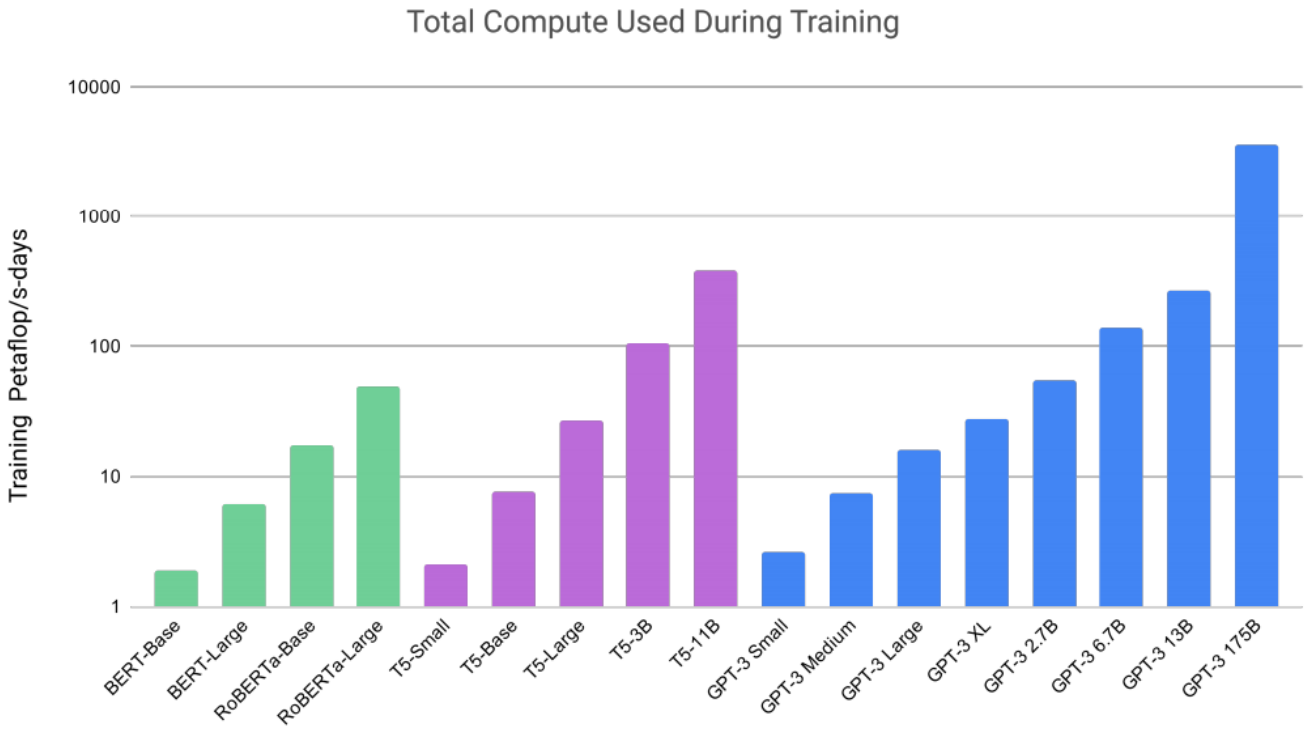


Figure 2.2: Total compute used during training. Based on the analysis in Scaling Laws For Neural Language Models [KMH+20] we train much larger models on many fewer tokens than is typical. As a consequence, although GPT-3 3B is almost 10x larger than RoBERTa-Large (355M params), both models took roughly 50 petaflop/s-days of compute during pre-training. Methodology for these calculations can be found in Appendix D.

Brown et al. «Language Models are Few-Shot Learners», 2020

GPT-3. Training Dataset

- Filtered CommonCrawl dataset;
- Performed fuzzy deduplication at the document level, within and across datasets;
- Added known high-quality reference corpora to the training mix to increase its diversity;
- Total: over 600 GB of text.

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

Table 2.2: Datasets used to train GPT-3. "Weight in training mix" refers to the fraction of examples during training that are drawn from a given dataset, which we intentionally do not make proportional to the size of the dataset. As a result, when we train for 300 billion tokens, some datasets are seen up to 3.4 times during training while other datasets are seen less than once.

Brown et al. «Language Models are Few-Shot Learners», 2020

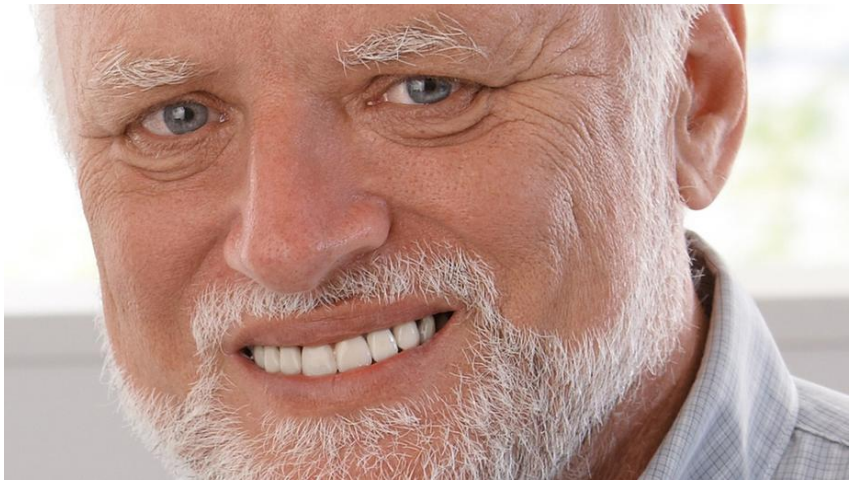
GPT-3. Contamination and memorization problems

- Training dataset is sourced from the internet, so it is possible that the model was trained on some of benchmark test sets.
- The authors proactively searched for any overlap between the training data and the development and test sets.

GPT-3. Contamination and memorization problems

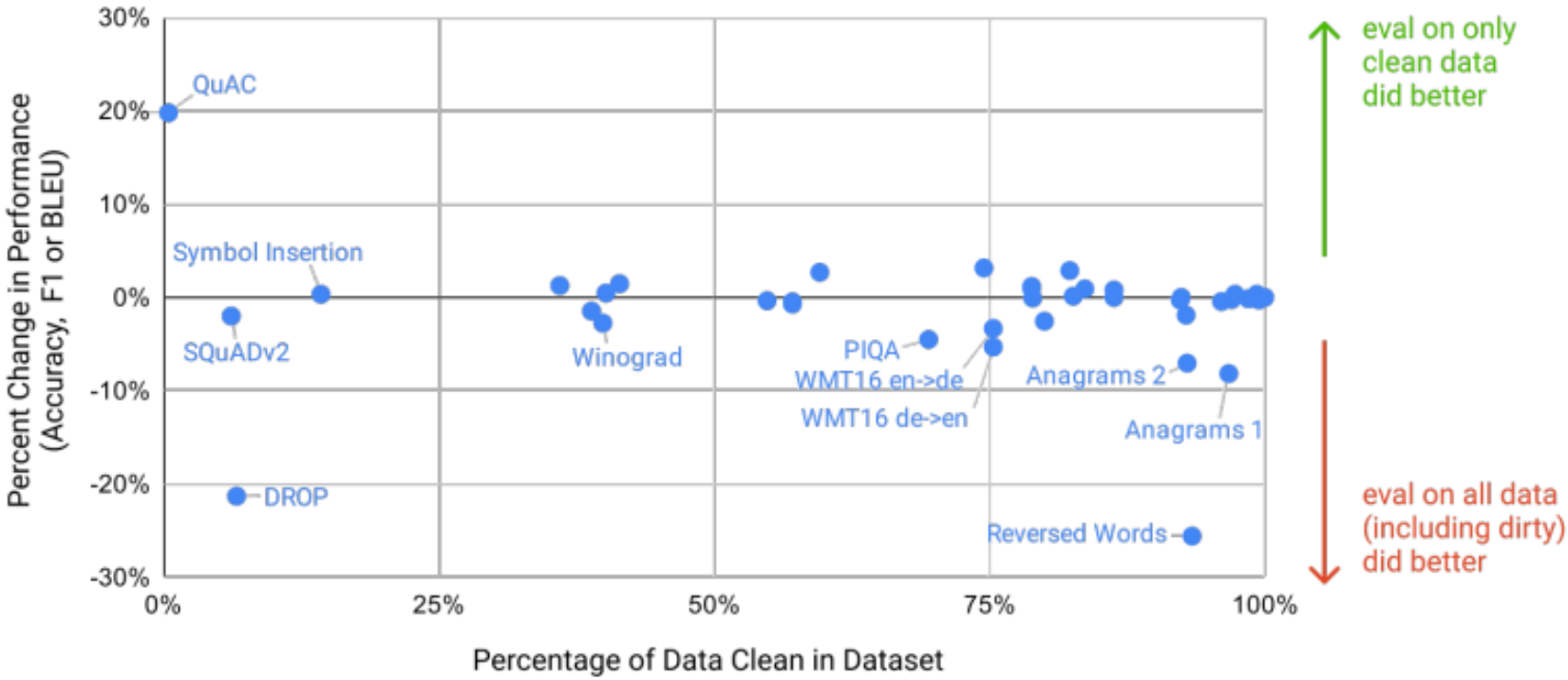
- Training dataset is sourced from the internet, so it is possible that the model was trained on some of benchmark test sets.
- The authors proactively searched for any overlap between the training data and the development and test sets.

«Unfortunately, a bug resulted in only partial removal of all detected overlaps from the training data. Due to the cost of training, it was not feasible to retrain the model».



GPT-3. Contamination and memorization problems

For each benchmark, a «clean» version was produced which removes all potentially leaked examples.



Brown et al. «Language Models are Few-Shot Learners», 2020

GPT-3. Results: News Generation

Human accuracy at detecting articles by GPT-3 is barely above chance.

	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. We find that human accuracy (measured by the ratio of correct assignments to non-neutral assignments) ranges from 86% on the control model to 52% on GPT-3 175B. This table compares mean accuracy between five different models, and shows the results of a two-sample T-Test for the difference in mean accuracy between each model and the control model (an unconditional GPT-3 Small model with increased output randomness).

Brown et al. «Language Models are Few-Shot Learners», 2020

GPT-3. Results: News Generation

Title: United Methodists Agree to Historic Split
Subtitle: Those who oppose gay marriage will form their own denomination
Article: After two days of intense debate, the United Methodist Church has agreed to a historic split - one that is expected to end in the creation of a new denomination, one that will be "theologically and socially conservative," according to The Washington Post. The majority of delegates attending the church's annual General Conference in May voted to strengthen a ban on the ordination of LGBTQ clergy and to write new rules that will "discipline" clergy who officiate at same-sex weddings. But those who opposed these measures have a new plan: They say they will form a separate denomination by 2020, calling their church the Christian Methodist denomination.
The Post notes that the denomination, which claims 12.5 million members, was in the early 20th century the "largest Protestant denomination in the U.S.," but that it has been shrinking in recent decades. The new split will be the second in the church's history. The first occurred in 1968, when roughly 10 percent of the denomination left to form the Evangelical United Brethren Church. The Post notes that the proposed split "comes at a critical time for the church, which has been losing members for years," which has been "pushed toward the brink of a schism over the role of LGBTQ people in the church." Gay marriage is not the only issue that has divided the church. In 2016, the denomination was split over ordination of transgender clergy, with the North Pacific regional conference voting to ban them from serving as clergy, and the South Pacific regional conference voting to allow them.

Figure 3.14: The GPT-3 generated news article that humans had the greatest difficulty distinguishing from a human written article (accuracy: 12%).

Brown et al. «Language Models are Few-Shot Learners», 2020

GPT-3. Results: News Generation

Title: Star's Tux Promise Draws Megyn Kelly's Sarcasm
Subtitle: Joaquin Phoenix pledged to not change for each awards event
Article: A year ago, Joaquin Phoenix made headlines when he appeared on the red carpet at the Golden Globes wearing a tuxedo with a paper bag over his head that read, "I am a shape-shifter. I can't change the world. I can only change myself." It was a promise to not change to fit into the Hollywood mold: "I think that's a really special thing, to not change yourself. I think it's a really special thing to say, 'This is what's inside of me, I'm proud of it, and I'm not going to be ashamed because of the way that someone else thinks I should be.'" Now, it's the Oscars, and Phoenix is at it again. But this time, his publicist is saying he'll be wearing a tux no matter what.
Megyn Kelly was not impressed, and she let him have it on The Tonight Show. "You know, I feel like, I feel like you could have worn the tux," she says. "But you're saying you're a shape-shifter. I don't know if you can change your tux, but you can change your mind. You can change your mind. You can change your mind." Phoenix says he did, but it didn't stick. "I was like, 'Okay, I'm going to wear a tuxedo to this thing.' And then I thought, 'I don't want to wear a tuxedo to this thing.'" Kelly goes on to encourage him to change his mind again, but Phoenix says it's too late: "I'm committed to wearing this."

Figure 3.15: The GPT-3 generated news article that humans found the easiest to distinguish from a human written article (accuracy: 61%).

Brown et al. «Language Models are Few-Shot Learners», 2020

GPT-3. Results: Question Answering

GPT-3 outperforms open-domain fine-tuned SotA on one dataset and approaches SotA on the other two despite not being fine-tuned.

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**

Brown et al. «Language Models are Few-Shot Learners», 2020

GPT-3. Results: Question Answering

TriviaQA

Q: American Callan Pinckney's eponymously named system became a best-selling (1980s-2000s) book/video franchise in what genre?
A: Fitness

NaturalQS

Q: when did kendrick lamars first album come out?
A: July 2, 2011

The questions in NaturalQS tend towards very fine-grained knowledge on Wikipedia.

GPT-3. Results: Translation

GPT-3 significantly outperforms prior unsupervised NMT work when translating into English but underperforms when translating in the other direction.

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

Brown et al. «Language Models are Few-Shot Learners», 2020

GPT-3. Results: Translation

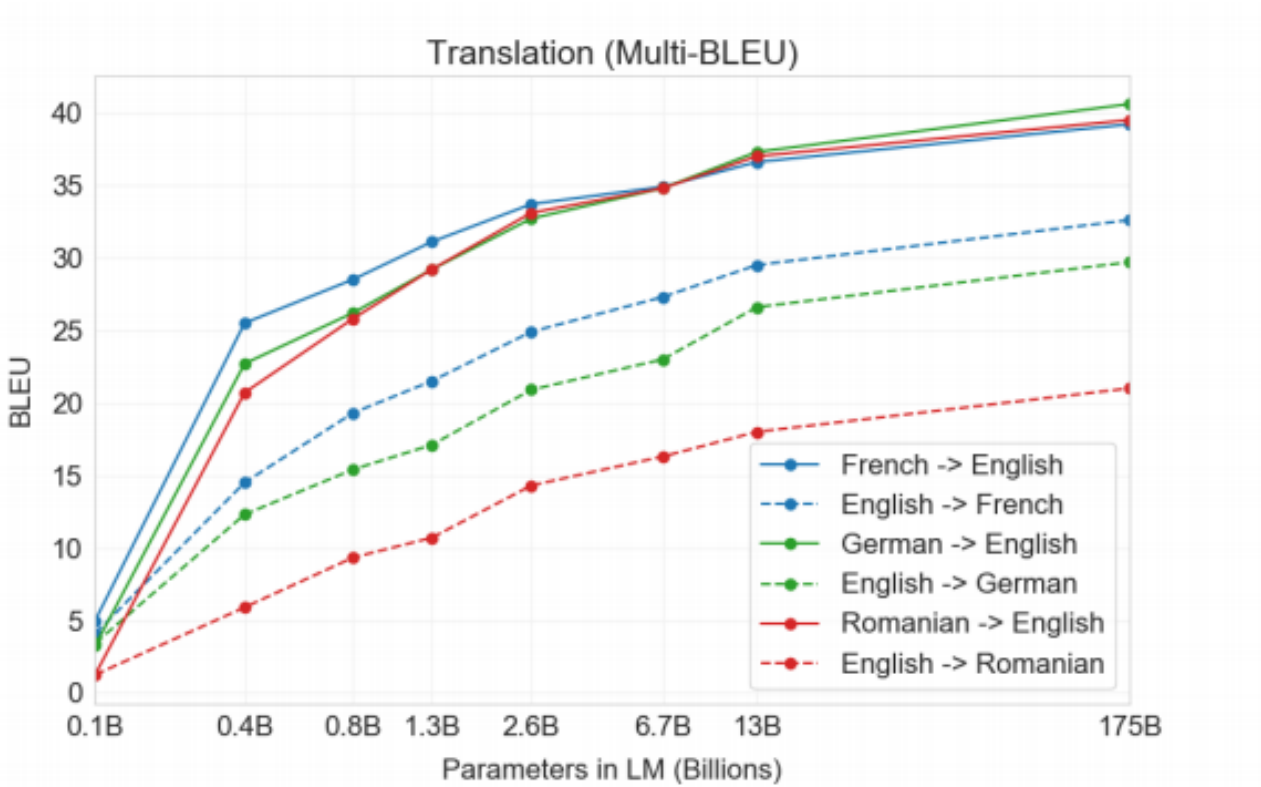


Figure 3.4: Few-shot translation performance on 6 language pairs as model capacity increases. There is a consistent trend of improvement across all datasets as the model scales, and as well as tendency for translation into English to be stronger than translation from English.

Brown et al. «Language Models are Few-Shot Learners», 2020

GPT-3. Results: Winograd-Style Tasks

Determine which word a pronoun refers to, when the pronoun is grammatically ambiguous but semantically unambiguous to a human.

GPT-3. Results: SuperGLUE

SuperGLUE is a benchmark that summarizes progress on a diverse set of NLP tasks and offers a single-number metric.

	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

Table 3.8: Performance of GPT-3 on SuperGLUE compared to fine-tuned baselines and SOTA. All results are reported on the test set. GPT-3 few-shot is given a total of 32 examples within the context of each task and performs no gradient updates.

Brown et al. «Language Models are Few-Shot Learners», 2020

GPT-3. Results: SuperGLUE

COPA (close to SotA): select the alternative that more plausibly has a causal relation with the premise

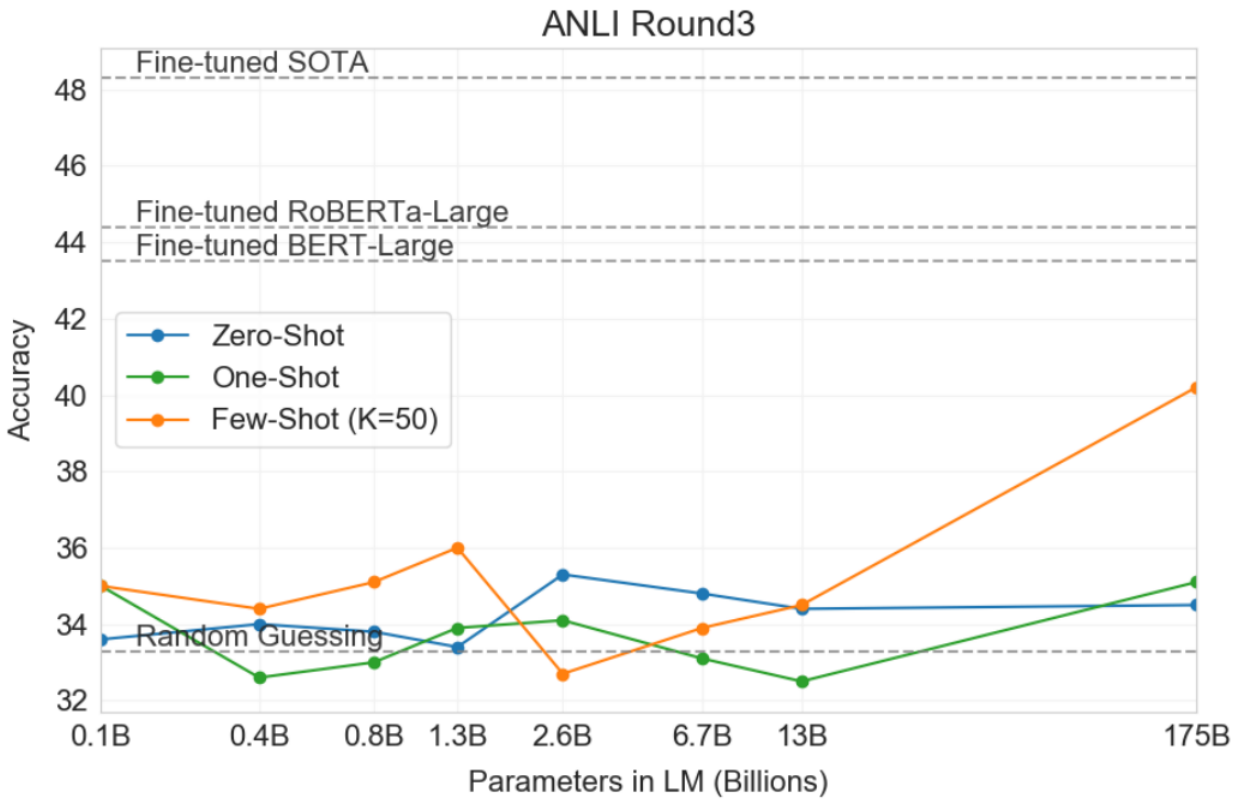
Premise: The man broke his toe. What was the CAUSE of this?
Alternative 1: He got a hole in his sock.
Alternative 2: He dropped a hammer on his foot.
Answer: 2.

WiC (random chance): do the occurrences of a word correspond to the same meaning?

Context-1: There's a lot of trash on the **bed** of the river.
Context-2: I keep a glass of water next to my **bed** when I sleep.
Answer: False.

GPT-3. Results: Natural Language Inference

Determine whether the second sentence logically follows from the first, contradicts the first sentence, or is possibly true (neutral).



Brown et al. «Language Models are Few-Shot Learners», 2020

GPT-3. Limitations

- On text synthesis GPT-3 samples still sometimes repeat themselves semantically at the document level, start to lose coherence over sufficiently long passages, contradict themselves, and occasionally contain non-sequitur sentences or paragraphs;

GPT-3. Limitations

- On text synthesis GPT-3 samples still sometimes repeat themselves semantically at the document level, start to lose coherence over sufficiently long passages, contradict themselves, and occasionally contain non-sequitur sentences or paragraphs;
- Ambiguity about whether few-shot learning actually learns new tasks from scratch at inference time, or if it simply recognizes and identifies tasks that it has learned during training;

GPT-3. Limitations

- On text synthesis GPT-3 samples still sometimes repeat themselves semantically at the document level, start to lose coherence over sufficiently long passages, contradict themselves, and occasionally contain non-sequitur sentences or paragraphs;
- Ambiguity about whether few-shot learning actually learns new tasks from scratch at inference time, or if it simply recognizes and identifies tasks that it has learned during training;
- Poor sample efficiency during pre-training: GPT-3 still sees much more text during pre-training than a human sees in the their lifetime;

GPT-3. Limitations

- On text synthesis GPT-3 samples still sometimes repeat themselves semantically at the document level, start to lose coherence over sufficiently long passages, contradict themselves, and occasionally contain non-sequitur sentences or paragraphs;
- Ambiguity about whether few-shot learning actually learns new tasks from scratch at inference time, or if it simply recognizes and identifies tasks that it has learned during training;
- Poor sample efficiency during pre-training: GPT-3 still sees much more text during pre-training than a human sees in the their lifetime;
- GPT-3 is expensive and inconvenient to perform inference on.