A Series of GPTs

[CSED490X] Recent Trends in ML: A Large-Scale Perspective

Jungtaek Kim

jtkim@postech.ac.kr

POSTECH
Pohang 37673, Republic of Korea
https://jungtaek.github.io

April 06, 2022

Table of Contents

Introduction

Tasks & Datasets

A Machine Learning Model

A Learning Algorithm

Experimental Results

Introduction

A Series of GPTs

	Selected Contributions	**#Parameters**
GPT [Radford et al., 2018]	Pre-training, then fine-tuning	117M
GPT-2 [Radford et al., 2019]	Task conditioning & Zero-shot task transfer (& In-context learning)	Up to 1.542B
GPT-3 [Brown et al., 2020]	In-context learning & Zero-shot, one-shot, and few-shot settings (without gradient updates)	Up to 175B

[Radford et al., 2018] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever. Improving language understanding by generative pre-training. `https://openai.com/blog/language-unsupervised/`, 2018.

[Radford et al., 2019] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever. Language models are unsupervised multitask learners. `https://openai.com/blog/better-language-models/`, 2019.

[Brown et al., 2020] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei. Language models are few-shot learners. In Advances in Neural Information Processing Systems (NeurIPS), volume 33, pages 1877–1901, Virtual, 2020.

A Series of GPTs

- Generative Pre-trained Transformer (GPT) and its variants have been proposed by a research group in OpenAI.
- They are based on the architecture of Transformer [Vaswani et al., 2017].
- GPT [Radford et al., 2018], GPT-2 [Radford et al., 2019], and GPT-3 [Brown et al., 2020] have been proposed so far.
- OpenAI supports their API, which is powered by GPT-3, to open research subfields and the commercial market.

Examples of OpenAI API with GPT-3

- OpenAI API
- Video Link 1
- Video Link 2
- Video Link 3

Transformer

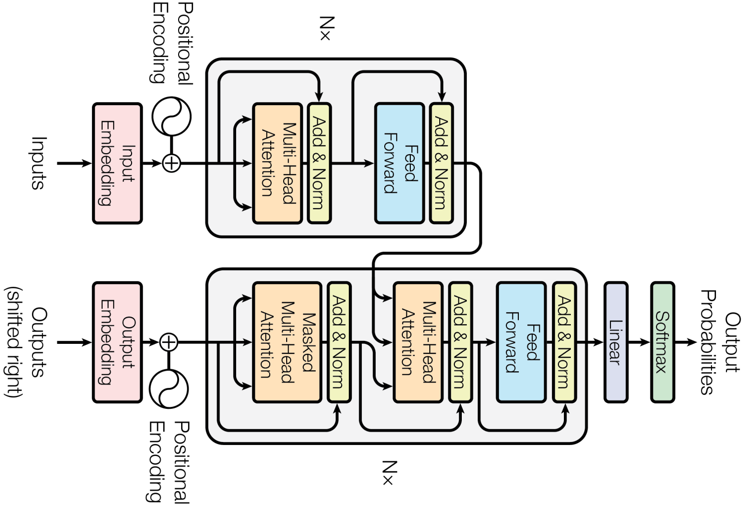


Figure 1: Transformer.

Figure 1 is taken from [Vaswani et al., 2017].

Tasks & Datasets

Tasks: GPT

- Twelve different tasks on natural language inference, question answering, sentence similarity, and classification are solved in the GPT paper.
- Some of these tasks have already been covered in the previous lecture on BERT [Devlin et al., 2018].

Table 1: Different tasks and datasets used in our experiments.

Task	**Datasets**
Natural language inference	SNLI, MultiNLI, Question NLI, RTE, SciTail
Question answering	RACE, Story Cloze
Sentence similarity	MSR Paraphrase Corpus, Quora Question Pairs, STS Benchmark
Classification	Stanford Sentiment Treebank-2, CoLA

Datasets: GPT

Unsupervised Pre-Training

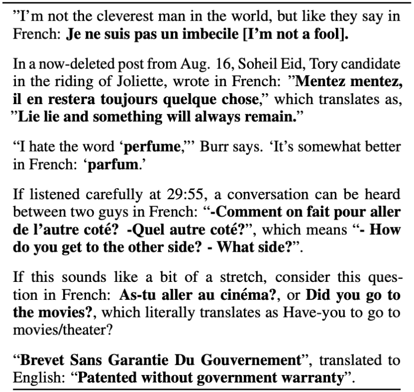
- The BooksCorpus dataset is used to train the GPT model.
- It contains over 7,000 unique unpublished books from a variety of genres including Adventure, Fantasy, and Romance.
- An alternative dataset, the 1B Word Benchmark, is approximately the same size but is shuffled at a sentence level – destroying long-range structure.

Fine-Tuning

- The GPT model is fine-tuned by task-specific datasets, in order to solve the corresponding tasks.

Tasks: GPT-2

- Diverse language tasks such as question answering, reading comprehension, summarization, and translation are solved.



”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**”.

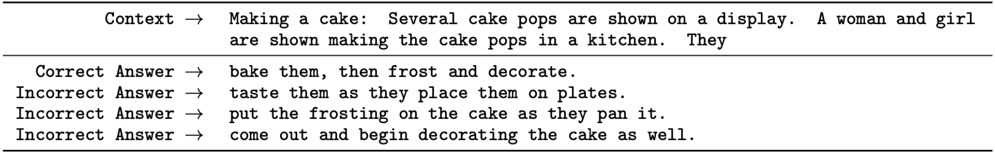
Table 1. Examples of naturally occurring demonstrations of English to French and French to English translation found throughout the WebText training set.

Datasets: GPT-2

- The authors create a new web scrape which emphasizes document quality.
- The web pages that have been curated and filtered by humans are only scraped.
- Manually filtering a full web scrape would be exceptionally expensive, so all outbound links from Reddit, which received at least 3 karma, are scraped.
- The resulting new dataset, WebText, contains the text subset of these 45 million links.
- After de-duplication and some heuristic-based cleaning, it contains slightly over 8 million documents for a total of 40 GB of text.
- All Wikipedia documents are removed from WebText, since it is a common data source for other datasets.

Tasks: GPT-3

- Similar to GPT-2, GPT-3 solves diverse language tasks such as completion, closed-book question answering, translation, common sense reasoning, and reading comprehension.

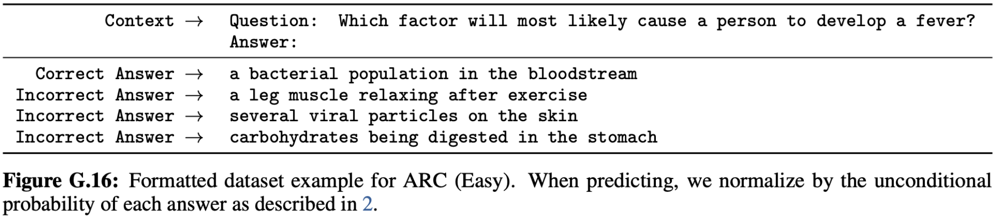


Context →	Making a cake: Several cake pops are shown on a display. A woman and girl are shown making the cake pops in a kitchen. They
Correct Answer →	bake them, then frost and decorate.
Incorrect Answer →	taste them as they place them on plates.
Incorrect Answer →	put the frosting on the cake as they pan it.
Incorrect Answer →	come out and begin decorating the cake as well.

Figure G.9: Formatted dataset example for HellaSwag

Tasks: GPT-3

- Similar to GPT-2, GPT-3 solves diverse language tasks such as completion, closed-book question answering, translation, common sense reasoning, and reading comprehension.

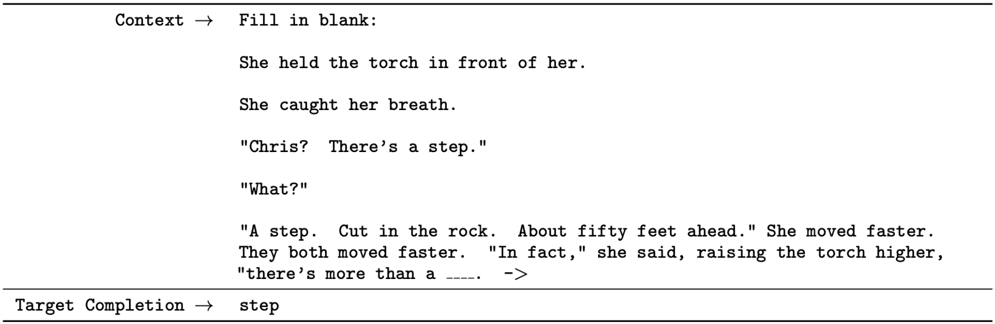


Context →	Question: Which factor will most likely cause a person to develop a fever? Answer:
Correct Answer →	a bacterial population in the bloodstream
Incorrect Answer →	a leg muscle relaxing after exercise
Incorrect Answer →	several viral particles on the skin
Incorrect Answer →	carbohydrates being digested in the stomach

Figure G.16: Formatted dataset example for ARC (Easy). When predicting, we normalize by the unconditional probability of each answer as described in 2.

Tasks: GPT-3

- Similar to GPT-2, GPT-3 solves diverse language tasks such as completion, closed-book question answering, translation, common sense reasoning, and reading comprehension.

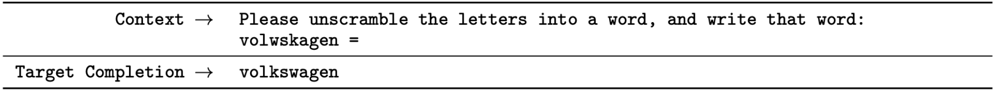


Context →	Fill in blank: She held the torch in front of her. She caught her breath. "Chris? There's a step." "What?" "A step. Cut in the rock. About fifty feet ahead." She moved faster. They both moved faster. "In fact," she said, raising the torch higher, "there's more than a ____. ->
Target Completion →	step

Figure G.21: Formatted dataset example for LAMBADA

Tasks: GPT-3

- Similar to GPT-2, GPT-3 solves diverse language tasks such as completion, closed-book question answering, translation, common sense reasoning, and reading comprehension.

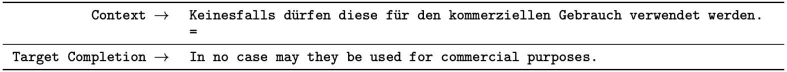


Context →	Please unscramble the letters into a word, and write that word: volwskagen =
Target Completion →	volkswagen

Figure G.23: Formatted dataset example for Anagrams 2

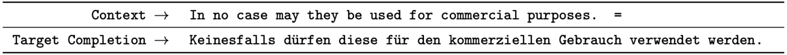
Tasks: GPT-3

- Similar to GPT-2, GPT-3 solves diverse language tasks such as completion, closed-book question answering, translation, common sense reasoning, and reading comprehension.



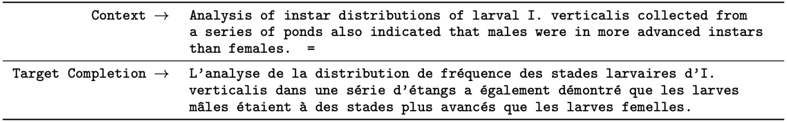
Context →	Keinesfalls dürfen diese für den kommerziellen Gebrauch verwendet werden. =
Target Completion →	In no case may they be used for commercial purposes.

Figure G.36: Formatted dataset example for De→En. This is the format for one- and few-shot learning, for this and other langauge tasks, the format for zero-shot learning is “Q: What is the {language} translation of {sentence} A: {translation}.”



Context →	In no case may they be used for commercial purposes. =
Target Completion →	Keinesfalls dürfen diese für den kommerziellen Gebrauch verwendet werden.

Figure G.37: Formatted dataset example for En→De

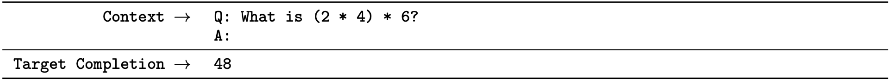


Context →	Analysis of instar distributions of larval I. verticalis collected from a series of ponds also indicated that males were in more advanced instars than females. =
Target Completion →	L’analyse de la distribution de fréquence des stades larvaires d’I. verticalis dans une série d’étangs a également démontré que les larves mâles étaient à des stades plus avancés que les larves femelles.

Figure G.38: Formatted dataset example for En→Fr

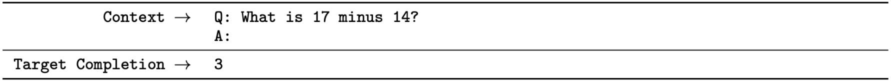
Tasks: GPT-3

- Similar to GPT-2, GPT-3 solves diverse language tasks such as completion, closed-book question answering, translation, common sense reasoning, and reading comprehension.



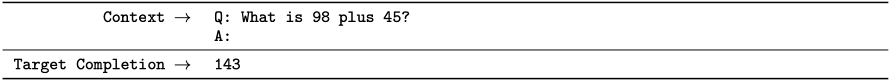
Context →	Q: What is (2 * 4) * 6? A:
Target Completion →	48

Figure G.42: Formatted dataset example for Arithmetic 1DC



Context →	Q: What is 17 minus 14? A:
Target Completion →	3

Figure G.43: Formatted dataset example for Arithmetic 2D-



Context →	Q: What is 98 plus 45? A:
Target Completion →	143

Figure G.44: Formatted dataset example for Arithmetic 2D+

Datasets: GPT-3

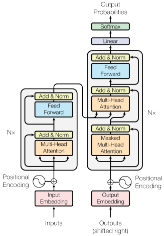
- Five datasets are used to train GPT-3.

Table 2: Datasets used to train GPT-3.

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

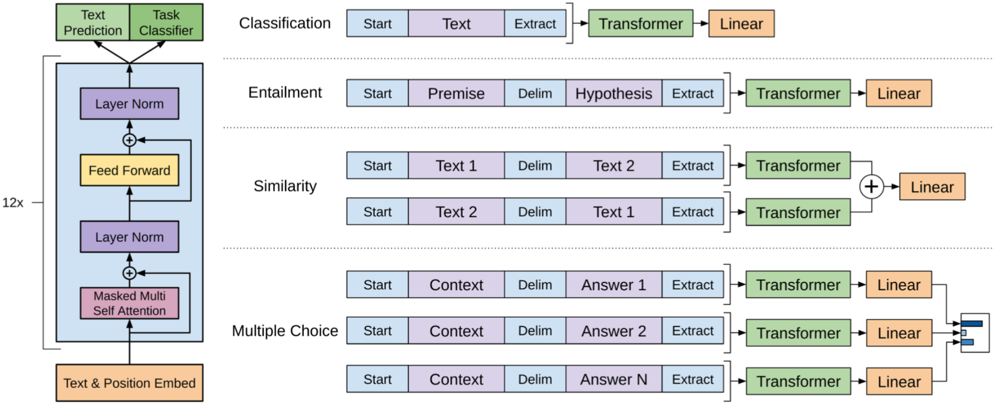
A Machine Learning Model

GPT



- GPT largely follows the original Transformer paper [Vaswani et al., 2017].
- It has a 12-layer decoder-only Transformer with masked self-attention heads (768-dimensional states and 12 attention heads).
- For the position-wise feed-forward networks, 3072-dimensional inner states are used.
- The decoder-only Transformer is pre-trained.
- The pre-trained Transformer and the last task-specific linear layer is fine-tuned.

GPT


Text Prediction
Task Classifier
Layer Norm
Feed Forward
Layer Norm
Masked Multi Self Attention
12x
Text & Position Embed
Classification
Start
Text
Extract
Transformer
Linear
Entailment
Start
Premise
Delim
Hypothesis
Extract
Transformer
Linear
Similarity
Start
Text 1
Delim
Text 2
Extract
Transformer
Start
Text 2
Delim
Text 1
Extract
Transformer
Linear
Multiple Choice
Start
Context
Delim
Answer 1
Extract
Transformer
Linear
Start
Context
Delim
Answer 2
Extract
Transformer
Linear
Start
Context
Delim
Answer N
Extract
Transformer
Linear

GPT

Unsupervised Pre-Training

- Given an unsupervised corpus of tokens $\mathcal{U} = \{u_1, \ldots, u_n\}$, the standard language modeling objective is used to maximize the following log-likelihood:

$$\mathcal{L}_1(\mathcal{U}) = \sum_i \log P(u_i \mid u_{i-k}, \ldots, u_{i-1}; \boldsymbol{\theta}), \tag{1}$$

  where k is the size of context window.

- The model is defined as

$$\begin{aligned} h_i &= \mathbf{U}\mathbf{W}_e + \mathbf{W}_p, && (2) \\ h_l &= \text{transformer\_block}(h_{l-1}) \quad \forall i \in [1, n], && (3) \\ P(u \mid \mathbf{U}) &= \text{softmax}(h_n \mathbf{W}), && (4) \end{aligned}$$

  where $\mathbf{U} = (u_{-k}, \ldots, u_{-1})$ is the context vector of tokens, n is the number of layers, $\mathbf{W}_e$ is a token embedding matrix, and $\mathbf{W}_p$ is a positional embedding matrix.

GPT

Supervised Fine-Tuning

- A task-specific model is defined as

$$P(y \mid x_1, \ldots x_m) = \mathrm{softmax}(h_{n,m}\mathbf{W}_y). \tag{5}$$

- After training the model, the model is fine-tuned by maximizing the following objective:

$$\mathcal{L}(\mathcal{C}) = \mathcal{L}_2(\mathcal{C}) + \lambda \mathcal{L}_1(\mathcal{C}), \tag{6}$$

where

$$\mathcal{L}_2(\mathcal{C}) = \sum_{(\mathbf{x},y)\in\mathcal{C}} \log P(y \mid x_1, \ldots, x_m), \tag{7}$$

and λ is a hyperparameter. Note that $\mathbf{x} = (x_1, \ldots, x_m)$.

GPT-2

- **Task conditioning** is defined as

$$p(\text{output} \mid \text{input}, \text{task}; \text{parameters}). \tag{8}$$

- For example, a translation training example can be written as

$$(\texttt{translation to French}, \langle \text{English text} \rangle, \langle \text{French text} \rangle). \tag{9}$$

- Likewise, a reading comprehension example can be written as

$$(\texttt{answer the question}, \langle \text{document} \rangle, \langle \text{question} \rangle, \langle \text{answer} \rangle). \tag{10}$$

- Without fine-tuning, **zero-shot task transfer** is capable of detecting a prompt, e.g., : (colon).

GPT-2

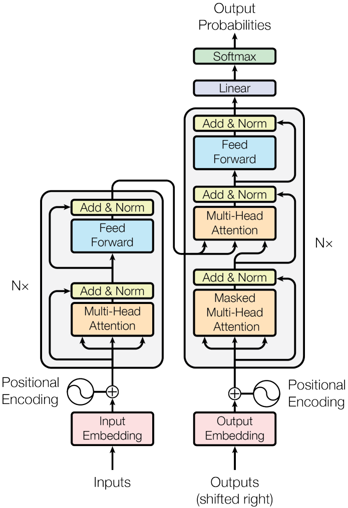


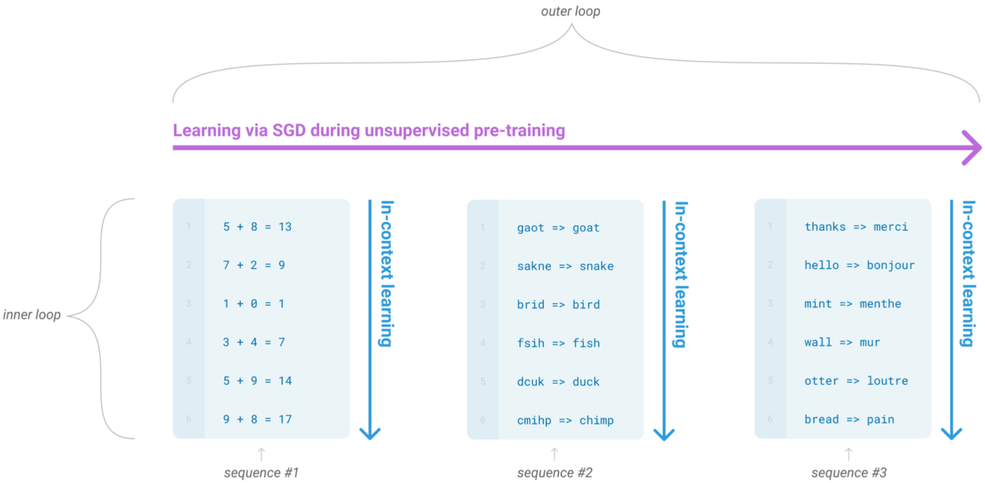
Table 3: Four GPT-2 architectures.

#Parameters	**#Layers**	d_{model}
117M	12	768
345M	24	1024
762M	36	1280
1542M	48	1600

GPT-3

- The same model and architecture of GPT-2 is used, including the modified initialization, pre-normalization, and reversible tokenization.
- GPT-3 utilizes **in-context learning**, which predicts the next word given context words.
- It systematically explores different settings for learning within the context.
- The settings where the model is given a few (or zero) demonstrations of the task at inference time as conditioning are conducted [Radford et al., 2019], but no weight updates are allowed.

GPT-3



outer loop

Learning via SGD during unsupervised pre-training

inner loop

```
5 + 8 = 13
7 + 2 = 9
1 + 0 = 1
3 + 4 = 7
5 + 9 = 14
9 + 8 = 17
```

In-context learning

sequence #1

```
gaot => goat
sakne => snake
brid => bird
fsih => fish
dcuk => duck
cmihp => chimp
```

In-context learning

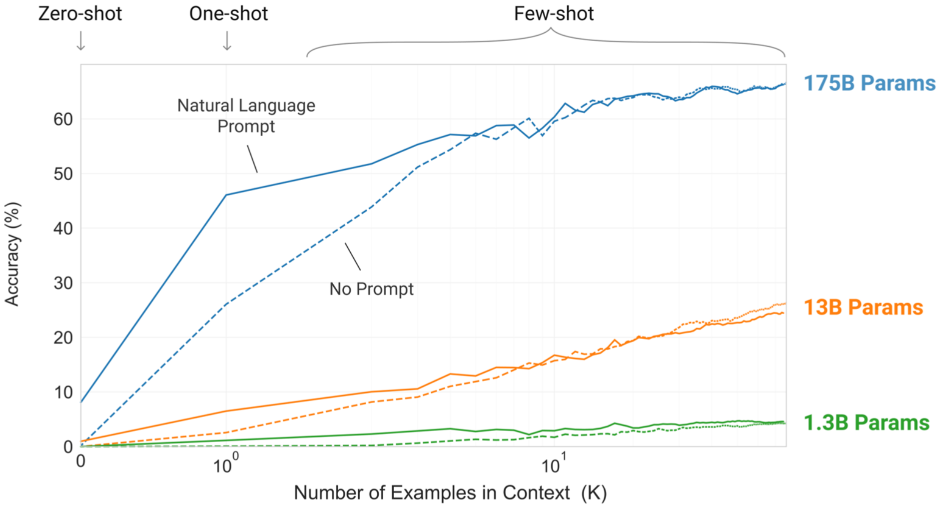
sequence #2

```
thanks => merci
hello => bonjour
mint => menthe
wall => mur
otter => loutre
bread => pain
```

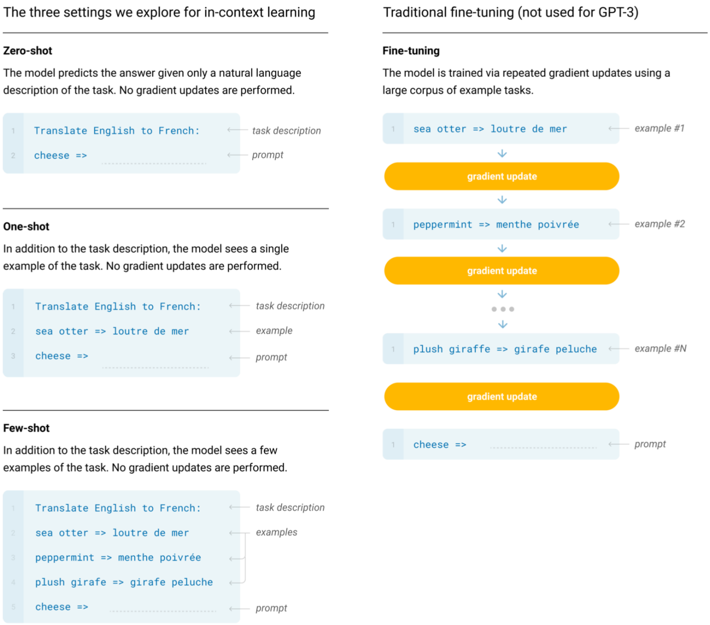
In-context learning

sequence #3

GPT-3


Zero-shot
One-shot
Few-shot
Natural Language Prompt
No Prompt
175B Params
13B Params
1.3B Params
Accuracy (%)
0
10
20
30
40
50
60
0
10^0
10^1
Number of Examples in Context (K)

GPT-3



The three settings we explore for in-context learning

Zero-shot

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

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

One-shot

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

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

Few-shot

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

```
Translate English to French:     ← task description
sea otter => loutre de mer       ← examples
peppermint => menthe poivrée
plush girafe => girafe peluche
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.

```
sea otter => loutre de mer       ← example #1
```

gradient update

```
peppermint => menthe poivrée     ← example #2
```

gradient update

• • •

```
plush giraffe => girafe peluche  ← example #N
```

gradient update

```
cheese =>                        ← prompt
```

GPT-3

Table 4: Eight GPT-3 architectures. BS and LR indicate batch size (in tokens) and learning rate, respectively. All models were trained for a total of 300 billion tokens.

Model Name	#Parameters	#Layers	d_{model}	n_{heads}	d_{head}	BS	LR
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}

A Learning Algorithm

Learning Algorithms

- GPT and GPT-3 are trained by Adam optimizers [Kingma and Ba, 2015]; the optimizer for GPT-2 is not specified in the paper.

- For GPT-3, the global norm of the gradient is clipped at 1.0 and the authors use cosine decay for learning rate down to 10% of its value, over 260 billion tokens. Moreover, there is a linear learning rate warmup over the first 375 million tokens.

- For GPT-3, the batch size gradually increases linearly from a small value (32k tokens) to the full value over the first 4-12 billion tokens of training, depending on the model size

Experimental Results

Experimental Results: GPT

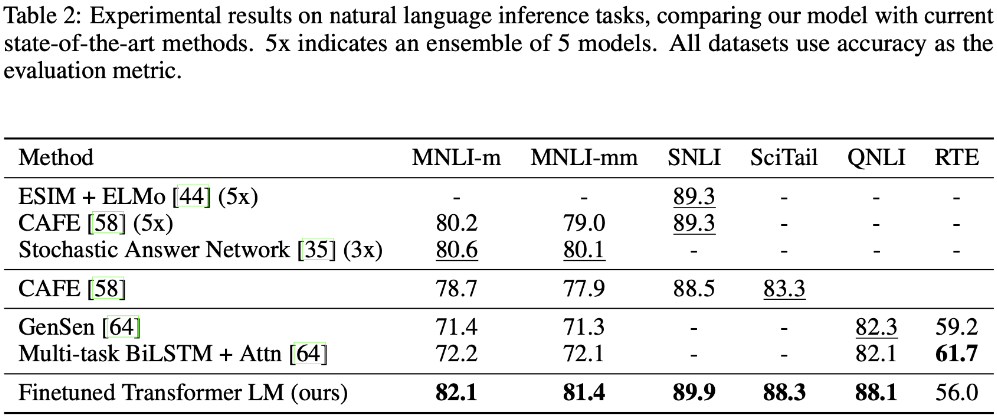


Table 2: Experimental results on natural language inference tasks, comparing our model with current state-of-the-art methods. 5x indicates an ensemble of 5 models. All datasets use accuracy as the evaluation metric.

Method	MNLI-m	MNLI-mm	SNLI	SciTail	QNLI	RTE
ESIM + ELMo [44] (5x)	-	-	89.3	-	-	-
CAFE [58] (5x)	80.2	79.0	89.3	-	-	-
Stochastic Answer Network [35] (3x)	80.6	80.1	-	-	-	-
CAFE [58]	78.7	77.9	88.5	83.3		
GenSen [64]	71.4	71.3	-	-	82.3	59.2
Multi-task BiLSTM + Attn [64]	72.2	72.1	-	-	82.1	**61.7**
Finetuned Transformer LM (ours)	**82.1**	**81.4**	**89.9**	**88.3**	**88.1**	56.0

Experimental Results: GPT

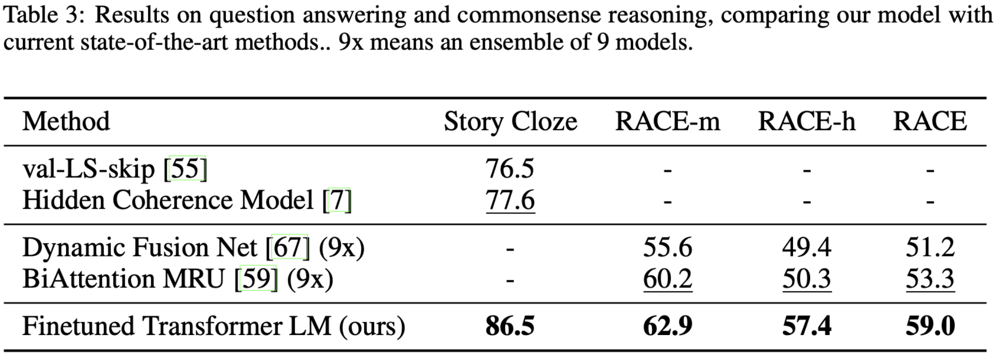


Table 3: Results on question answering and commonsense reasoning, comparing our model with current state-of-the-art methods.. 9x means an ensemble of 9 models.

Method	Story Cloze	RACE-m	RACE-h	RACE
val-LS-skip [55]	76.5	-	-	-
Hidden Coherence Model [7]	77.6	-	-	-
Dynamic Fusion Net [67] (9x)	-	55.6	49.4	51.2
BiAttention MRU [59] (9x)	-	60.2	50.3	53.3
Finetuned Transformer LM (ours)	**86.5**	**62.9**	**57.4**	**59.0**

Experimental Results: GPT

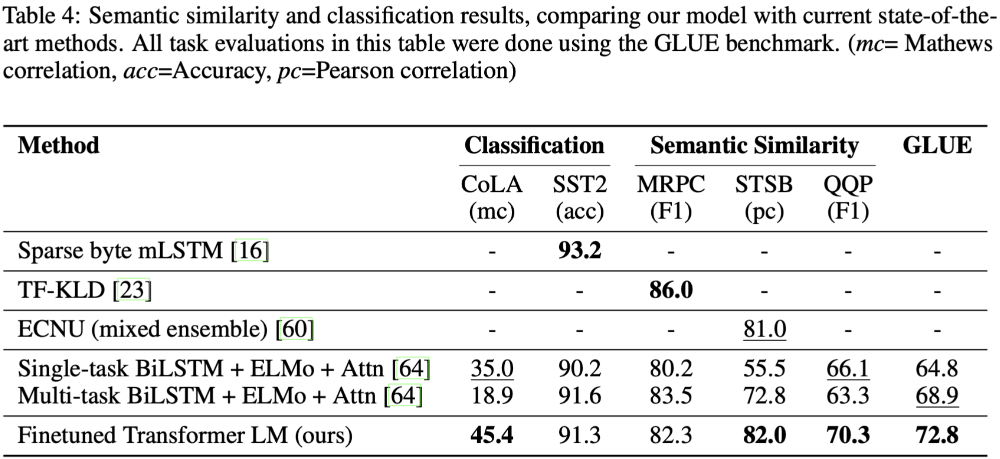
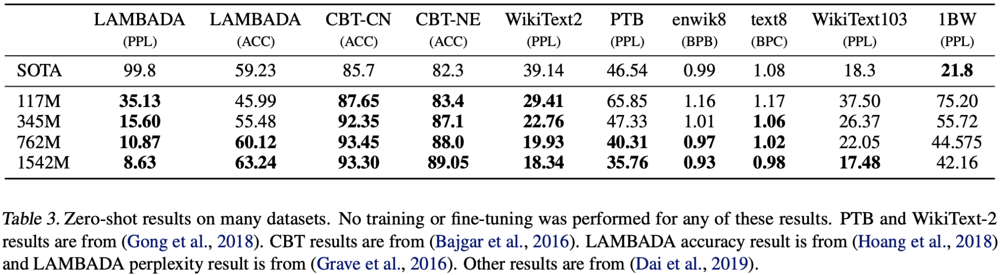


Table 4: Semantic similarity and classification results, comparing our model with current state-of-the-art methods. All task evaluations in this table were done using the GLUE benchmark. (*mc*= Mathews correlation, *acc*=Accuracy, *pc*=Pearson correlation)

Method	**Classification**		**Semantic Similarity**			**GLUE**
	CoLA (mc)	SST2 (acc)	MRPC (F1)	STSB (pc)	QQP (F1)	
Sparse byte mLSTM [16]	-	**93.2**	-	-	-	-
TF-KLD [23]	-	-	**86.0**	-	-	-
ECNU (mixed ensemble) [60]	-	-	-	81.0	-	-
Single-task BiLSTM + ELMo + Attn [64]	35.0	90.2	80.2	55.5	66.1	64.8
Multi-task BiLSTM + ELMo + Attn [64]	18.9	91.6	83.5	72.8	63.3	68.9
Finetuned Transformer LM (ours)	**45.4**	91.3	82.3	**82.0**	**70.3**	**72.8**

Experimental Results: GPT-2



	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

Table 3. Zero-shot results on many datasets. No training or fine-tuning was performed for any of these results. PTB and WikiText-2 results are from (Gong et al., 2018). CBT results are from (Bajgar et al., 2016). LAMBADA accuracy result is from (Hoang et al., 2018) and LAMBADA perplexity result is from (Grave et al., 2016). Other results are from (Dai et al., 2019).

Experimental Results: GPT-2

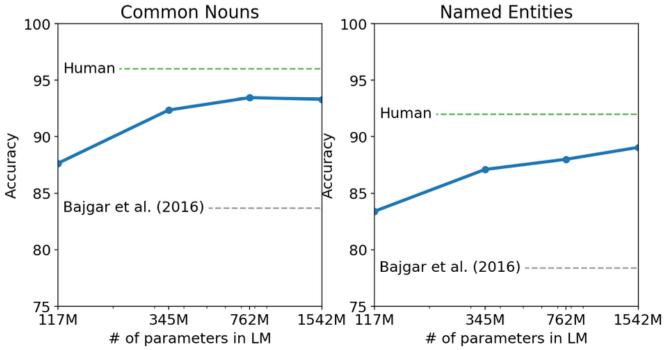


Figure 2. Performance on the Children's Book Test as a function of model capacity. Human performance are from Bajgar et al. (2016), instead of the much lower estimates from the original paper.

Experimental Results: GPT-2

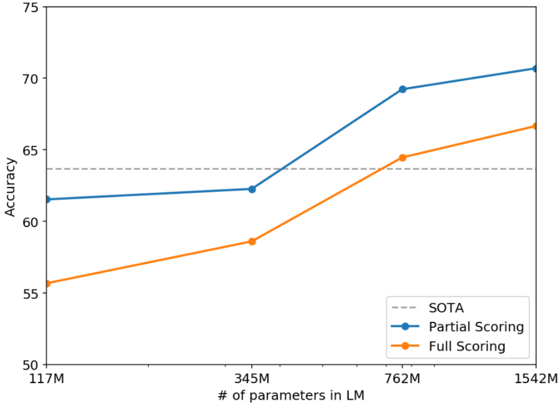
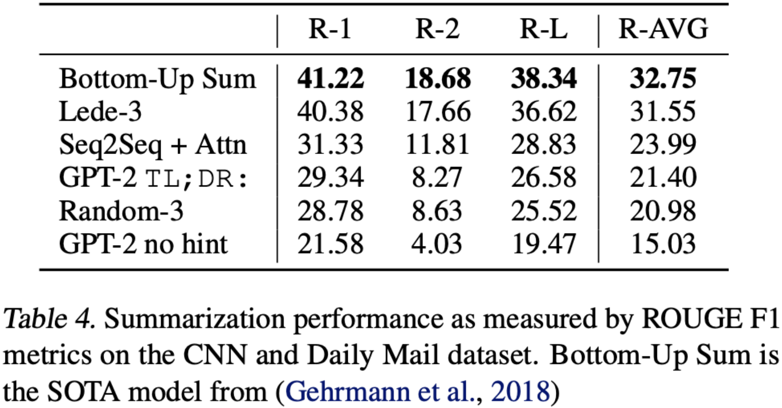


Figure 3. Performance on the Winograd Schema Challenge as a function of model capacity.

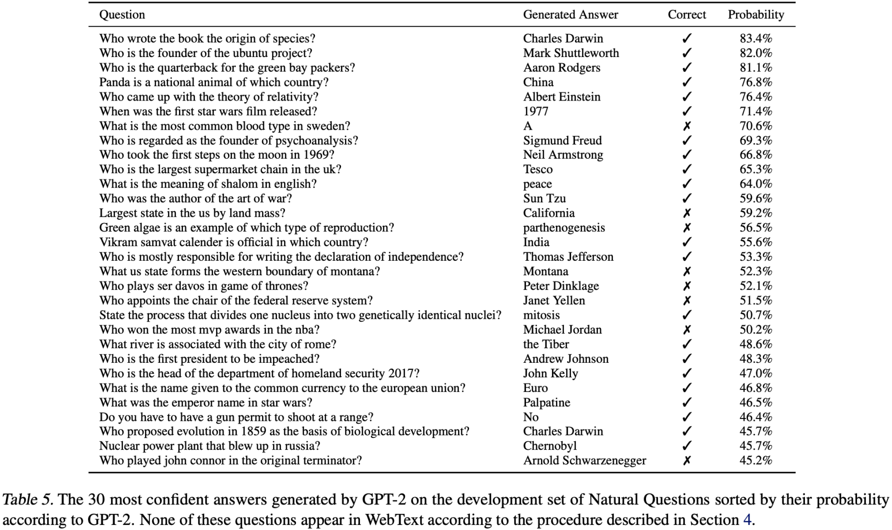
Experimental Results: GPT-2



	R-1	R-2	R-L	R-AVG
Bottom-Up Sum	**41.22**	**18.68**	**38.34**	**32.75**
Lede-3	40.38	17.66	36.62	31.55
Seq2Seq + Attn	31.33	11.81	28.83	23.99
GPT-2 TL;DR:	29.34	8.27	26.58	21.40
Random-3	28.78	8.63	25.52	20.98
GPT-2 no hint	21.58	4.03	19.47	15.03

Table 4. Summarization performance as measured by ROUGE F1 metrics on the CNN and Daily Mail dataset. Bottom-Up Sum is the SOTA model from (Gehrmann et al., 2018)

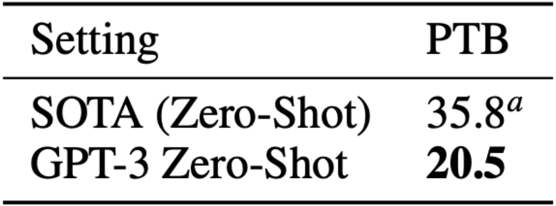
Experimental Results: GPT-2



Question	Generated Answer	Correct	Probability
Who wrote the book the origin of species?	Charles Darwin	✓	83.4%
Who is the founder of the ubuntu project?	Mark Shuttleworth	✓	82.0%
Who is the quarterback for the green bay packers?	Aaron Rodgers	✓	81.1%
Panda is a national animal of which country?	China	✓	76.8%
Who came up with the theory of relativity?	Albert Einstein	✓	76.4%
When was the first star wars film released?	1977	✓	71.4%
What is the most common blood type in sweden?	A	✗	70.6%
Who is regarded as the founder of psychoanalysis?	Sigmund Freud	✓	69.3%
Who took the first steps on the moon in 1969?	Neil Armstrong	✓	66.8%
Who is the largest supermarket chain in the uk?	Tesco	✓	65.3%
What is the meaning of shalom in english?	peace	✓	64.0%
Who was the author of the art of war?	Sun Tzu	✓	59.6%
Largest state in the us by land mass?	California	✗	59.2%
Green algae is an example of which type of reproduction?	parthenogenesis	✗	56.5%
Vikram samvat calender is official in which country?	India	✓	55.6%
Who is mostly responsible for writing the declaration of independence?	Thomas Jefferson	✓	53.3%
What us state forms the western boundary of montana?	Montana	✗	52.3%
Who plays ser davos in game of thrones?	Peter Dinklage	✗	52.1%
Who appoints the chair of the federal reserve system?	Janet Yellen	✗	51.5%
State the process that divides one nucleus into two genetically identical nuclei?	mitosis	✓	50.7%
Who won the most mvp awards in the nba?	Michael Jordan	✗	50.2%
What river is associated with the city of rome?	the Tiber	✓	48.6%
Who is the first president to be impeached?	Andrew Johnson	✓	48.3%
Who is the head of the department of homeland security 2017?	John Kelly	✓	47.0%
What is the name given to the common currency to the european union?	Euro	✓	46.8%
What was the emperor name in star wars?	Palpatine	✓	46.5%
Do you have to have a gun permit to shoot at a range?	No	✓	46.4%
Who proposed evolution in 1859 as the basis of biological development?	Charles Darwin	✓	45.7%
Nuclear power plant that blew up in russia?	Chernobyl	✓	45.7%
Who played john connor in the original terminator?	Arnold Schwarzenegger	✗	45.2%

Table 5. The 30 most confident answers generated by GPT-2 on the development set of Natural Questions sorted by their probability according to GPT-2. None of these questions appear in WebText according to the procedure described in Section 4.

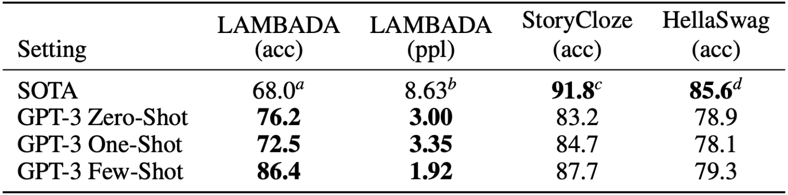
Experimental Results: GPT-3



Setting	PTB
SOTA (Zero-Shot)	35.8[a]
GPT-3 Zero-Shot	**20.5**

Figure 2: Zero-shot results on PTB language modeling dataset. a indicates GPT-2.

Experimental Results: GPT-3



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

Figure 3: Performance on Cloze and completion tasks. b indicates GPT-2.

Experimental Results: GPT-3

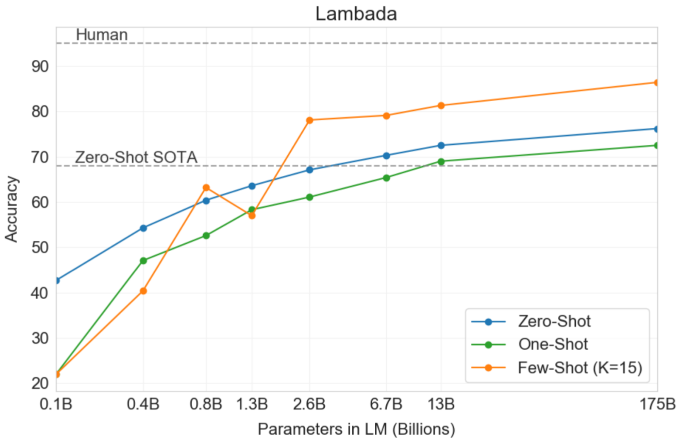
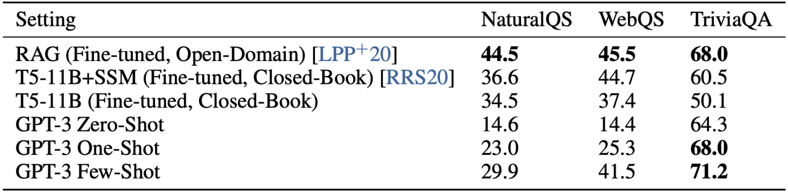


Figure 4: On LAMBADA, the few-shot capability of language models results in a strong boost to accuracy.

Experimental Results: GPT-3



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**

Figure 5: Results on three Open-Domain QA tasks.

Experimental Results: GPT-3

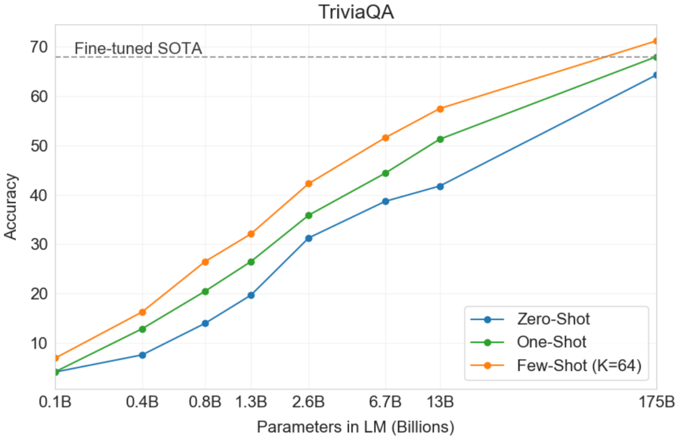
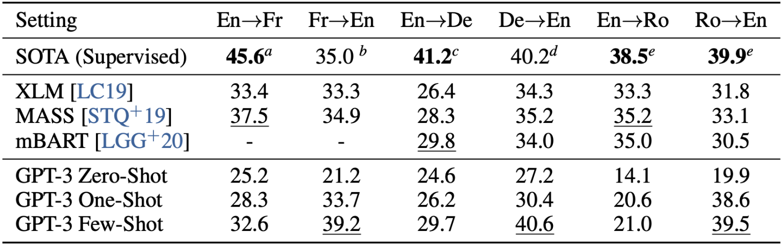


Figure 6: On TriviaQA GPT3's performance grows smoothly with model size, suggesting that language models continue to absorb knowledge as their capacity increases.

Experimental Results: GPT-3



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

Figure 7: Few-shot GPT-3 outperforms previous unsupervised NMT work by 5 BLEU when translating into English reflecting its strength as an English LM.

Experimental Results: GPT-3

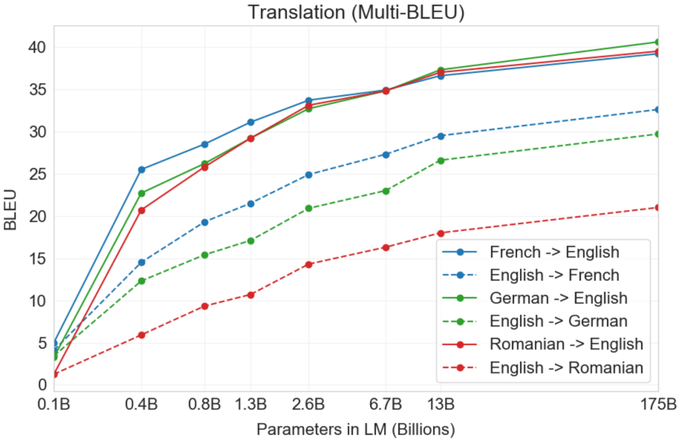
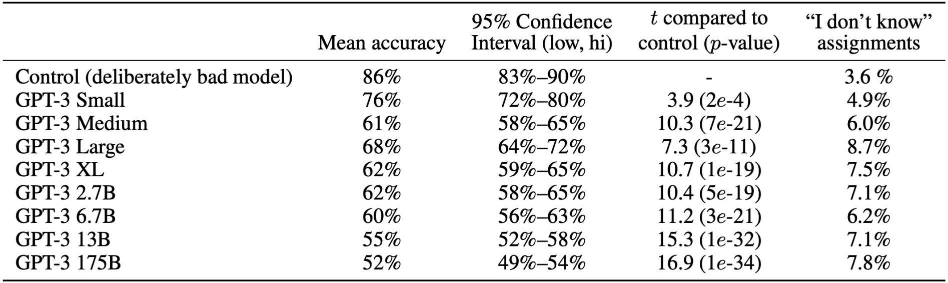


Figure 8: Few-shot translation performance on 6 language pairs as model capacity increases.

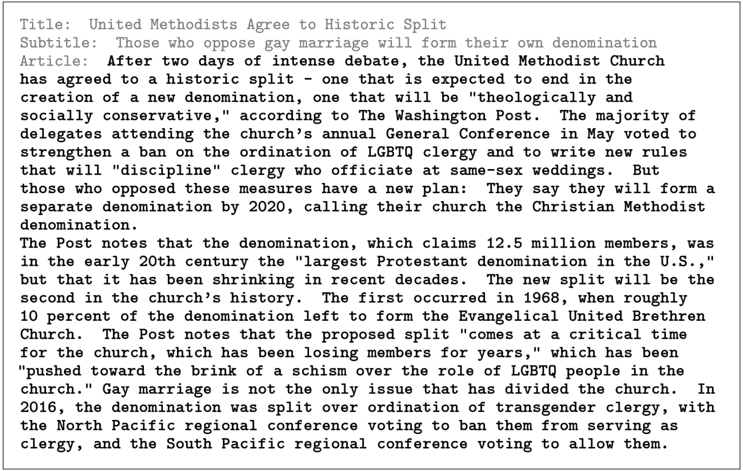
Experimental Results: GPT-3



	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%

Figure 9: Human accuracy in identifying whether short (∼200 words) news articles are model generated.

Experimental Results: GPT-3

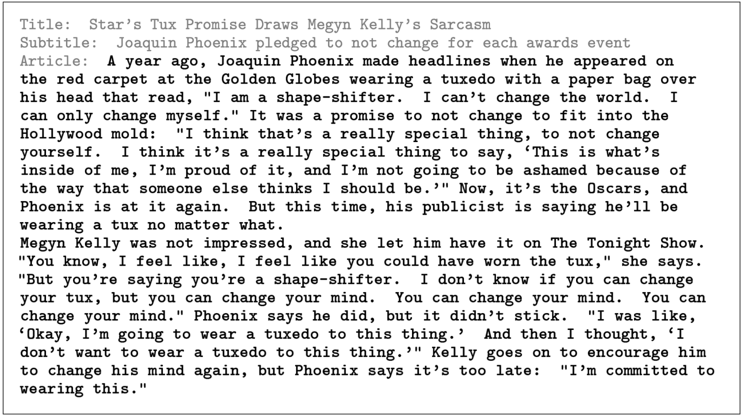


Title: United Methodists Agree to Historic Split
Subtitle: Those who oppose gay marriage will form their own denomination
Article: **After two days of intense debate, the United Methodist Church has agreed to a historic split - one that is expected to end in the creation of a new denomination, one that 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%).

Experimental Results: GPT-3



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%).

Any Questions?

References I

T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei. Language models are few-shot learners. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pages 1877–1901, Virtual, 2020.

J. Devlin, M. Chang, K. Lee, and K. Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

D. P. Kingma and J. L. Ba. ADAM: A method for stochastic optimization. In *Proceedings of the International Conference on Learning Representations (ICLR)*, San Diego, California, USA, 2015.

A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever. Improving language understanding by generative pre-training. `https://openai.com/blog/language-unsupervised/`, 2018.

A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever. Language models are unsupervised multitask learners. `https://openai.com/blog/better-language-models/`, 2019.

A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30, pages 5998–6008, Long Beach, California, USA, 2017.