

# Language Models are Few-Shot Learners

and @OSK dl-study

2020/08/22

# 概要

自然言語処理分野の論文  
モデル「GPT-3」を論文内で提唱している。

OpenAI開発

※GPT=Generative Pretrained Transformer

# 背景

## 自然言語処理トレンドの移り変わり

### ●RNN&CNN

文章の特徴量を引き継いだり畳み込んで得ようとする。  
遠い依存関係が失われてしまう問題がある

### ●LSTM

いろんなゲートを備えたLSTMを使って文章の特徴量を引き継ぐ  
改善はされたがやはり遠い依存関係はあまり引き継げない

# 背景

## 自然言語処理トレンドの移り変わり

- Transformer

Attention機構を使って文章の中で注目すればいい場所を計算してくれる。  
遠い依存関係を引き継げるようになった

- BERT

Transformerを使ったモデル

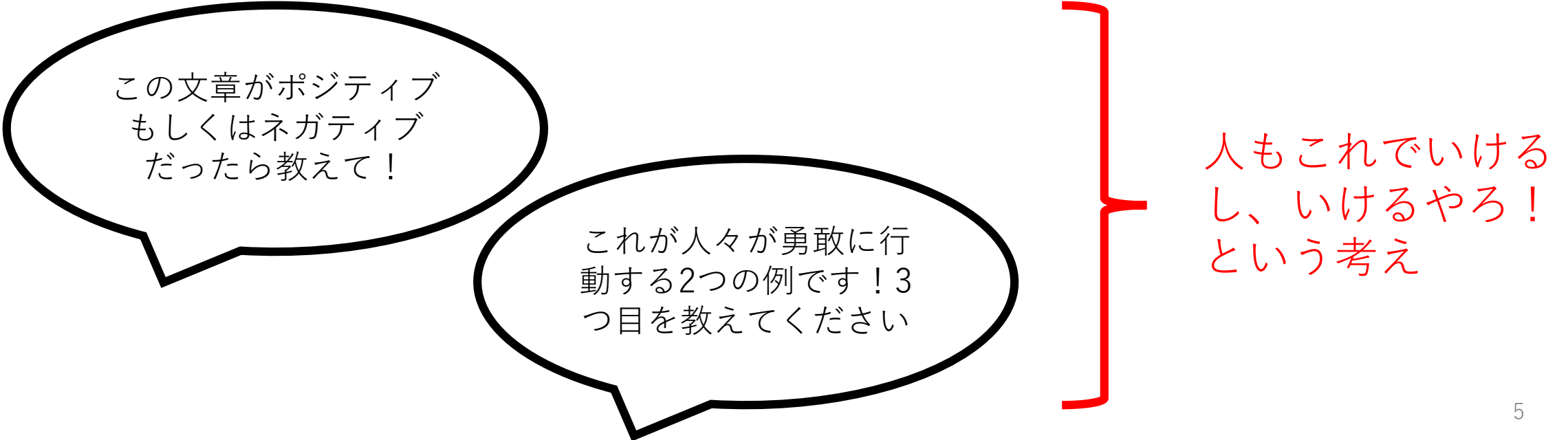
事前学習→Fine Tuning

言語に関する全般的な知識を事前学習→タスクに関連するデータで微調整。  
いろんなタスクを解くときに、Fine-tuningするだけでよくなった(楽！)

# 背景

BERTでは、Fine-tuningといっても数千~数十万の教師ありデータが必要になる。→普通に大変。

人は指示内容だけとか具体例一つでタスクを解けるからそういうモデルがあれば良いな…



この文章がポジティブ  
もしくはネガティブ  
だったら教えて！

これが人々が勇敢に行  
動する2つの例です！3  
つ目を教えてください

人もこれでいける  
し、いけるやろ！  
という考え

# Approach

## 事前学習→微調整

- Fine-tuning

タスクに向けた数千～数十万の教師ありデータで学習, 勾配更新を伴う  
その後タスクを解く

- Few-shot

タスク内容記述+タスクのデモンストレーション10-100例を入力としタスクを解く

- One-shot

タスク内容記述+タスクのデモンストレーション1例を入力としタスクを解く

- Zero-shot

タスク内容記述のみを入力としタスクを解く

# パラメータの更新は行われない？

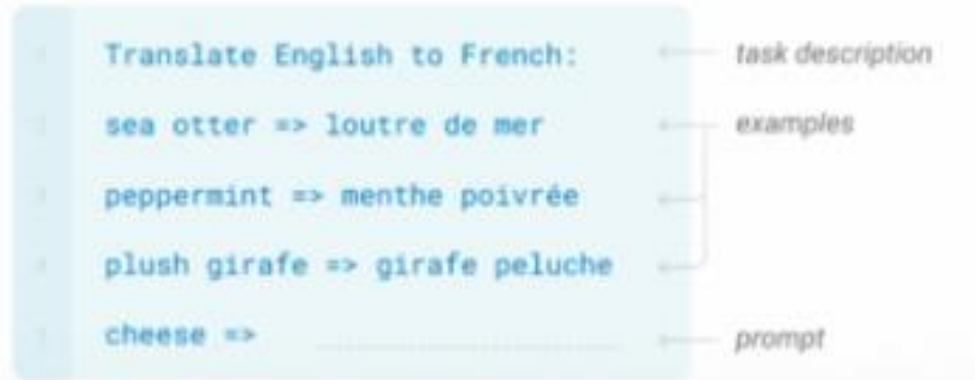
- GPT-3ではすべてのタスクを同じパラメータのモデルで行う
- **条件付き確率**を使って次の単語を予測

$$p(x) = \prod_{i=1}^n p(s_i | s_1, \dots, s_{i-1})$$

# Approach – Few-shot 紹介

## Few-shot

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



$$p(x) = \prod_{i=1}^n p(\underline{s_n} | \underline{s_1, \dots, s_{n-1}})$$

"fromage" "Translate English to French:  
sear otter => loutre de mer  
peppermint => menthe poivree  
plush girafe => girafe peluche  
cheese =>"

条件付き確率の条件として タスク内容記述+タスクのデモンストレーション10-100例 を用いる。

Example数：K

モデルのcontext windowサイズ： $n_{ctx}=2048$

全Exampleの単語数が $n_{ctx}$ におさまるようにKを10-100の間で決定する。



# Approach –One-shot紹介

## One-shot

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

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

条件付き確率の条件として タスク内容記述+タスクのデモンストレーション1例 を用いる。

人がタスクを解くときの条件に最も近い？

$$p(x) = \prod_{i=1}^n p(\underline{s_n} | \underline{s_1, \dots, s_{n-1}})$$

"fromage" "Translate English to French:  
sear otter => loutre de mer  
cheese =>"

# Approach –Zero-shot紹介

## Zero-shot

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



1 Translate English to French: task description

2 cheese => prompt

$$p(x) = \prod_{i=1}^n p(\underline{s_n} | \underline{s_1, \dots, s_{n-1}})$$

"fromage" "Translate English to French: cheese =>"

条件付き確率の条件として タスク内容記述を用いる。

最も楽なので利便性高い。

最もロバストになる潜在性を秘めている  
(One-shotやFew-shotは精度がExampleに影響を受ける)

最も難しい

# Approach -Fine tuning紹介

## 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.



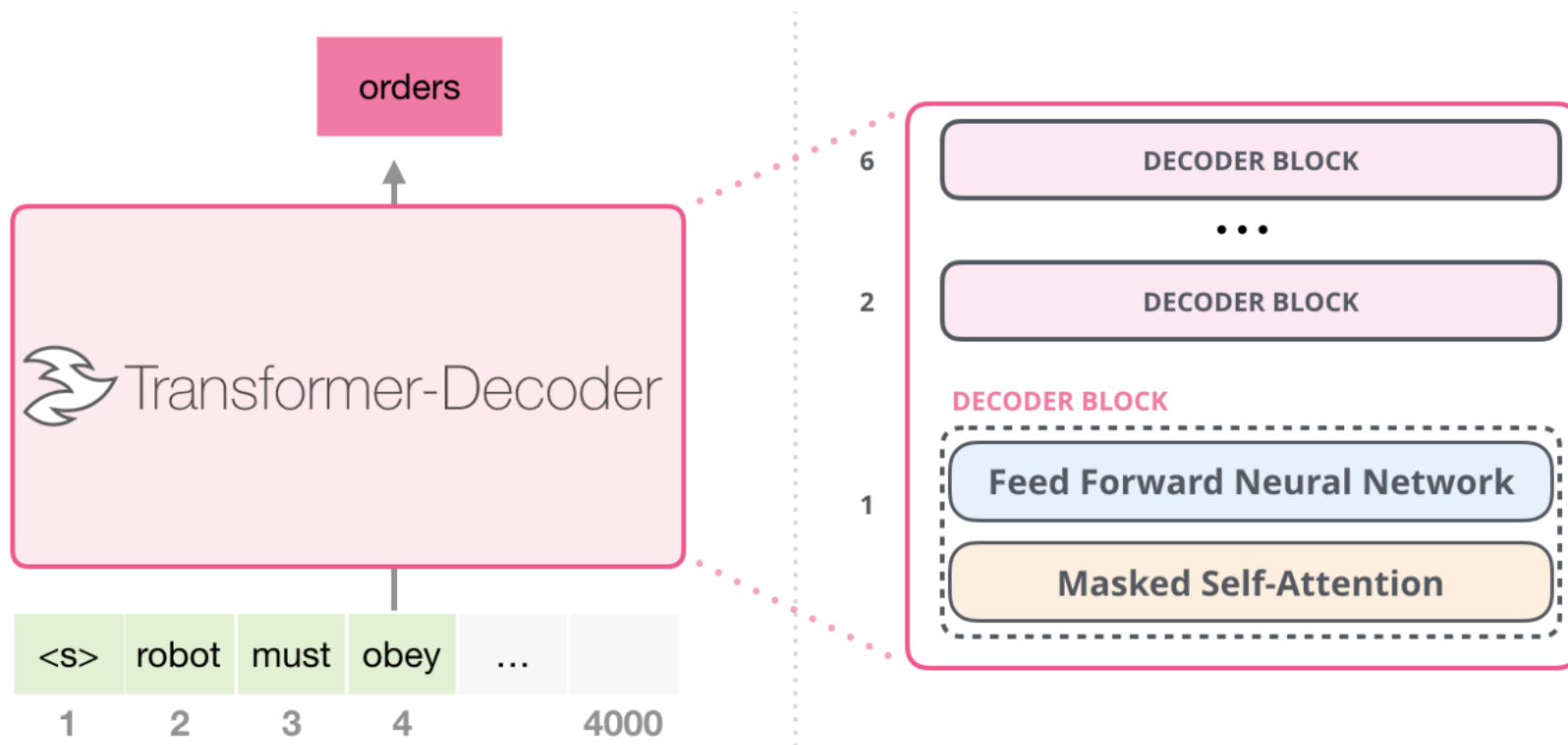
BERT等で使われているFine-tuningの手法  
(GPT-3では使われていない)

勾配更新を伴う

GPT-3では使われていないが、精度が期待  
できるのでfuture workとして残されている。

# Models and Architectures

- GPT-2と基本的に同じ構造, TransformerのDecoderだけ使う



# Models and Architectures

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

$n_{\text{params}}$  : パラメータ数

$n_{\text{layers}}$  : レイヤー数

$d_{\text{model}}$  : ボトルネックレイヤーにあるunitの数？

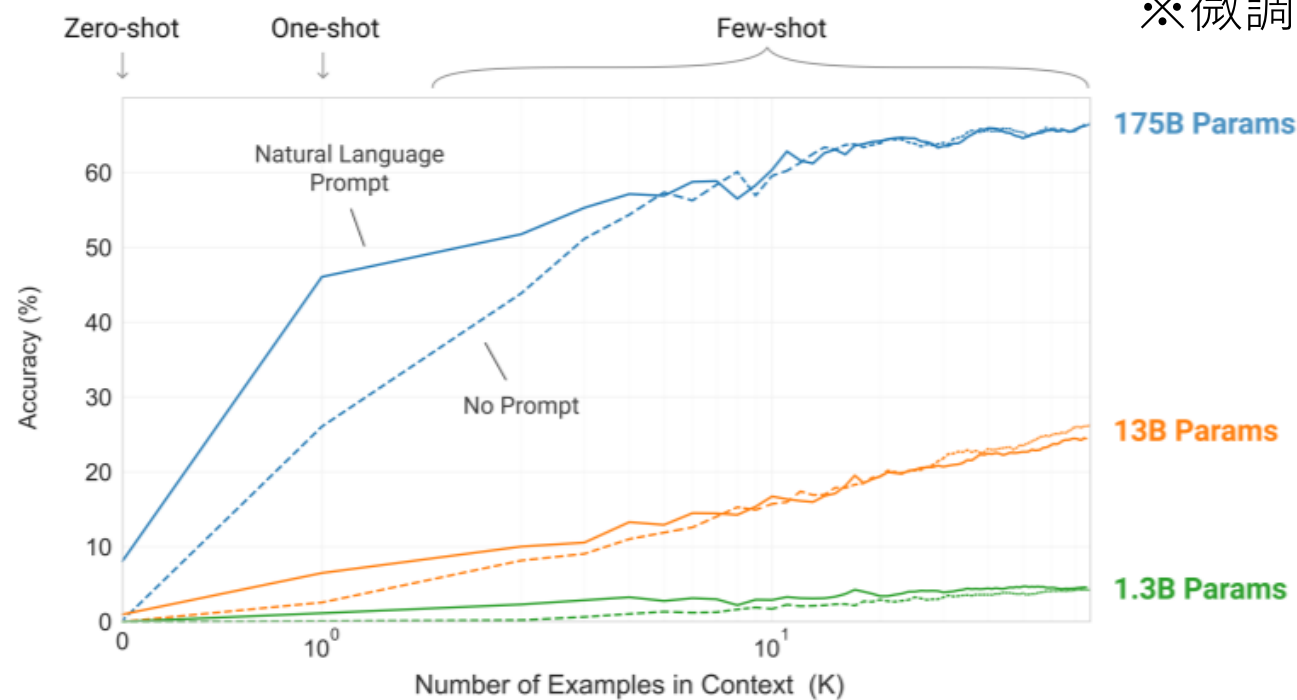
$n_{\text{heads}}$  : ？？

$d_{\text{head}}$  : attention headの次元数

# Models and Architectures

微調整の手法/パラメータ数を変えてGPT-3の適応力を評価

※微調整時パラメータの更新は行われない



Zero-shot：タスク内容記述のみ

One-shot：タスク内容記述+タスクのデモンストレーション1例

Few-shot：タスク内容記述+タスクのデモンストレーション10-100例

# データセット

- the Common Crawl dataset (約1兆語) を使用
- 92% : English, 7% : other languages
- データセットの品質を上げるため次のような工夫をしている
  - ドキュメント単位で大体同じ内容の文を省く「あいまいな重複排除」処理を行う
  - 高品質なテキストデータ(WebText dataset, two-Internet based books corpora, English-language Wikipedia)を足す

# データセット

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

意図的に高品質なデータが多く使われるようにしている（Common CrawlやBooks2などのデータは1回未満しかサンプルされないが、何回もサンプルされるデータもある。）

これにより、より高品質なデータで学習が行える利点がある（その代わり少しoverfitする）



# データセット

大量のPre-training Datasetの中に評価のためのタスクに関するデータが入ってしまう恐れがある。すると、few-shotの意味がなくなる…。

→それを防ぐため検索してOverlapをなるべく取り除いている。

---

Unfortunately, a bug in the filtering caused us to ignore some overlaps, and due to the cost of training it was not feasible to retrain the model.

このフィルタリング部分でバグが見つかり、いくつかのoverlapsが混入してしまっている

# Results

- タスク 「LAMBADA」 …文章を読んで単語の穴埋め  
(例)

*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

「そう、私は赤ちゃんを失うと思っていた」「僕も怖かった」と彼は言った 彼の目には誠実さが溢れていた 「あなたが？」 「もちろんだよ なぜ聞くの？」 「この赤ちゃんは計画的ではなかったから…」

「僕が『君に流産して欲しい』と考えると思う？」

---

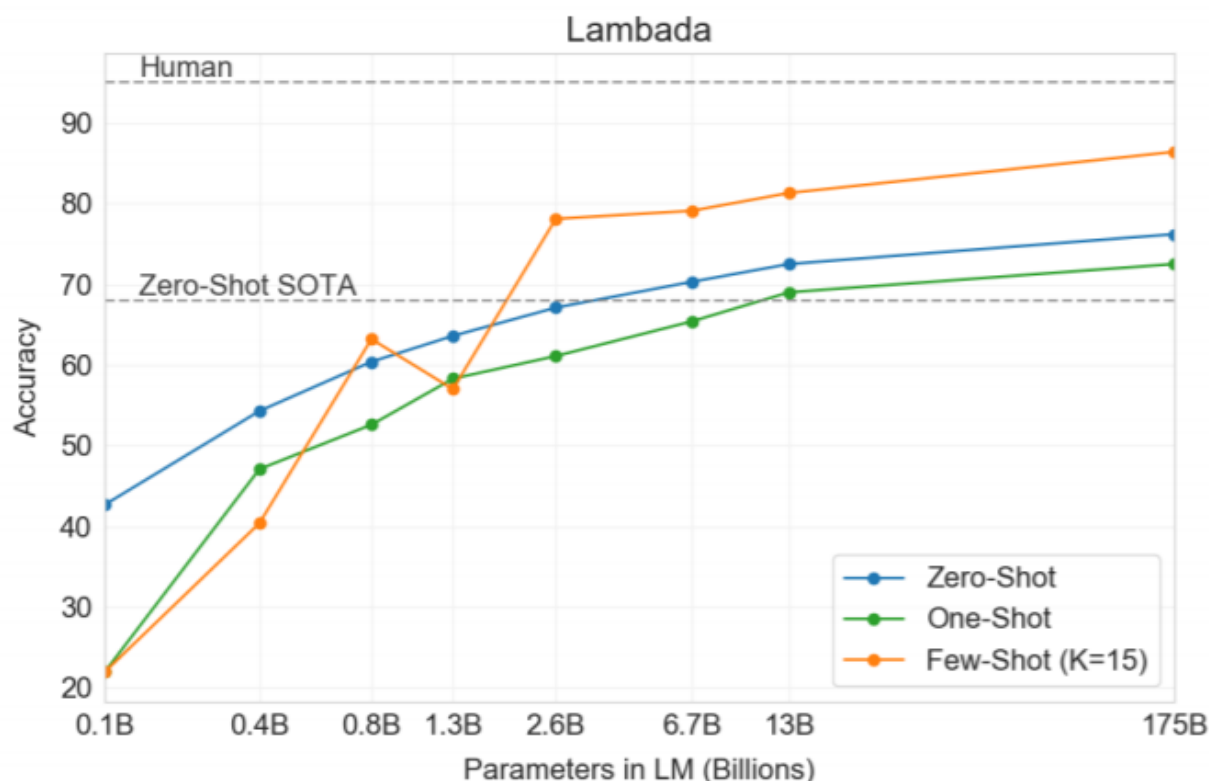
(One/Few-shotでモデルに渡すquery)

Alice was friends with Bob. Alice went to visit her friend \_\_\_\_\_. → Bob

George bought some baseball equipment, a ball, a glove, and a \_\_\_\_\_. →

# Results

- タスク「LAMBADA」…文章を読んで単語の穴埋め



- Zero-shot SOTAを達成
- Few-shotで86%の正答率(SOTA)
- モデルを大きくしても少ししか精度が良くならない(LAMBADAの特徴らしい)
- 精度がOne-shot<Zero-shot (この問題では一つでパターンを見つけるのは難しいよう。)

# Results

- タスク 「TriviaQA」 …クイズに答える

---

Which politician won the **Nobel Peace Prize** in 2009?

What **fragrant essential oil** is obtained from Damask Rose?

**Who** won the Nobel Peace Prize in 2009?

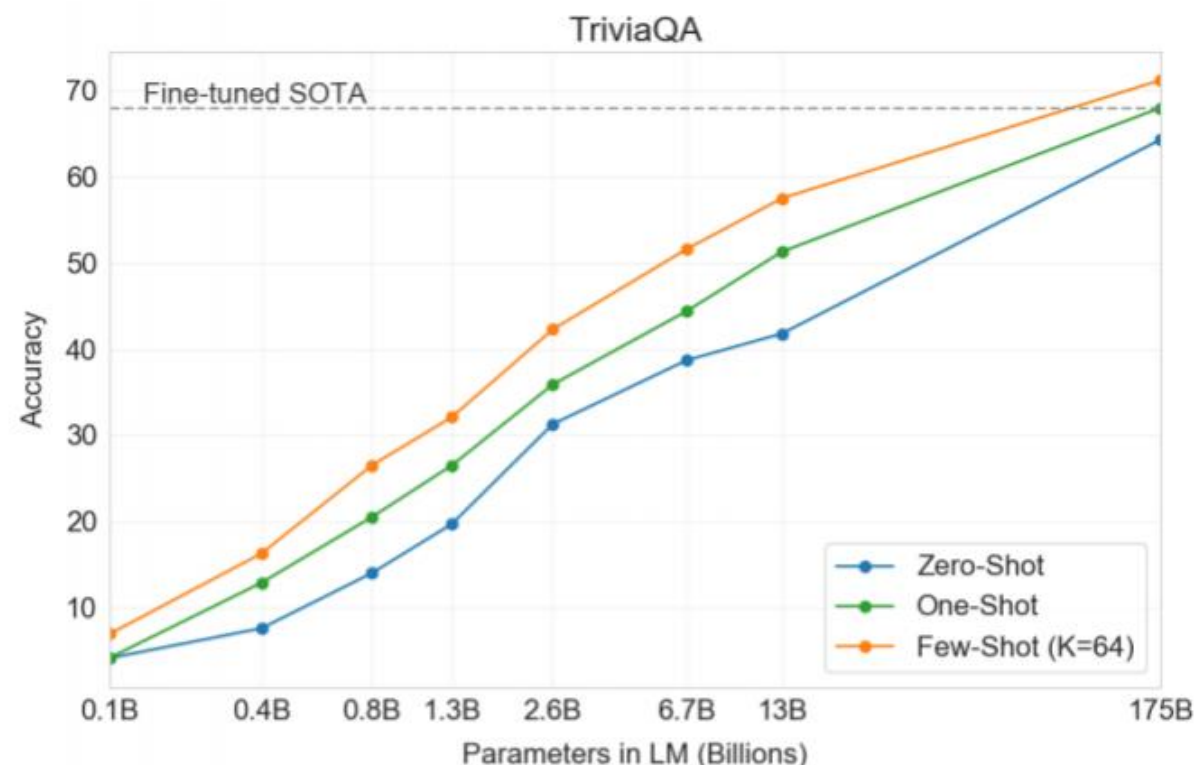
What was photographed for the first time in **October 1959**

What is the appropriate name of the **largest** type of frog?

---

# Results

- タスク「TriviaQA」…クイズに答える



- Few-shotなのにFine-tuningを使ったモデルのSOTAを超えてる (One-shotで同じ精度!)
- モデルが大きいほど精度が出ている。
- このタスクでは順当にZero-shot<One-shot
- まさに圧倒的なパラメータ数を持つGPT-3の強みが知識量となり表れている感じがするな～

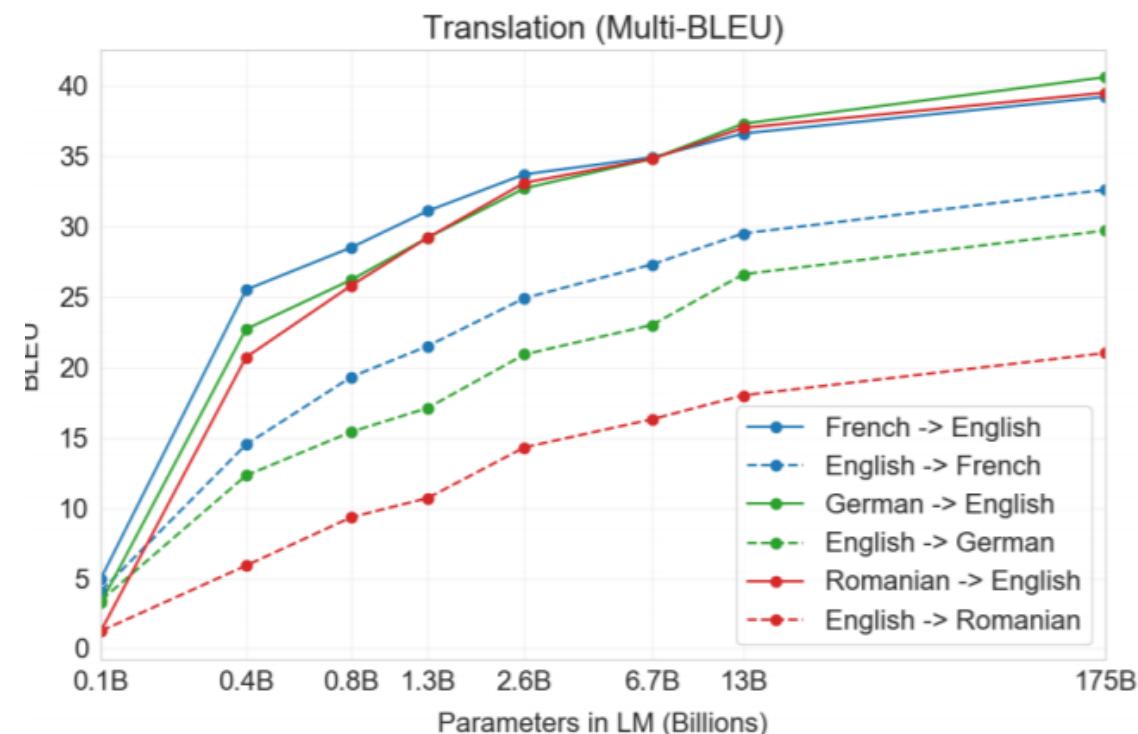
# Results

- 翻訳性能

Setting	En→Fr	Fr→En	En→De	De→En	En→Ro	Ro→En
SOTA (Supervised)	<b>45.6<sup>a</sup></b>	35.0 <sup>b</sup>	<b>41.2<sup>c</sup></b>	40.2 <sup>d</sup>	<b>38.5<sup>e</sup></b>	<b>39.9<sup>e</sup></b>
XLM [LC19]	33.4	33.3	26.4	34.3	33.3	31.8
MASS [STQ <sup>+</sup> 19]	<u>37.5</u>	34.9	28.3	35.2	<u>35.2</u>	33.1
mBART [LGG <sup>+</sup> 20]	-	-	<u>29.8</u>	34.0	<u>35.0</u>	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	<u>39.2</u>	29.7	<u>40.6</u>	21.0	<u>39.5</u>

(他モデルとの比較)

(翻訳する言語ごとの精度)



- 多言語→英語への翻訳で高い性能を示した。
- 普通翻訳モデルは二つの言語データで学習するので、英語中心で学習してこの精度が出るのは凄い

※GPT-3は様々な言語をミックスしたデータで事前学習しているため、厳密には他の教師なし学習モデルとは比較できない

# Results

- タスク 「Winogrande」 …代名詞が何を指しているか

✓	Robert woke up at 9:00am while Samuel woke up at 6:00am, so <b>he</b> had <u>less</u> time to get ready for school.	<b>Robert</b> / Samuel
	Robert woke up at 9:00am while Samuel woke up at 6:00am, so <b>he</b> had <u>more</u> time to get ready for school.	Robert / <b>Samuel</b>

# Results

- タスク「Winogrande」…代名詞が何を指しているか



- Few-shotがFine-tuningを使ったモデル「RoBERT a-large」と同等の精度を達成。
- Fine-tuningを使ったSOTAモデルには届いてない。
- Winograndeは結構人と現在のモデルでラグが大きいタスクらしい。



# Results

- タスク：「PIQA」…常識問題（二択）
    - **Goal** To separate egg whites from the yolk using a water bottle, you should ...
    - **Solution 1** *Squeeze* the water bottle and press it against the yolk. *Release*, which creates suction and lifts the yolk.
    - **Solution 2** *Place* the water bottle and press it against the yolk. *Keep pushing*, which creates suction and lifts the yolk.
- 

問題：ウォーターボトルを使用して卵白を卵黄から分離するには...

解決策1：ウォーターボトルを絞って卵黄に押し付けます。解放、それは吸引を作成し、卵黄を持ち上げます。

解決策2：ウォーターボトルを置き、卵黄に押し付けます。押し続けると、吸引力が生まれ、卵黄が持ち上げられます。

---

（簡単に思えるけど、ペットボトルを見たことも触ったこともない言語モデルがこの問題解くの凄いね…）

# Results

- タスク：「ARC」・・・3年生から9年生の科学試験から収集された選択問題

Context →	Question: George wants to warm his hands quickly by rubbing them. Which skin surface will produce the most heat? Answer:
Correct Answer →	dry palms
Incorrect Answer →	wet palms
Incorrect Answer →	palms covered with oil
Incorrect Answer →	palms covered with lotion

質問：ジョージは、手をこすって手をすばやく温めたいと考えています。どの皮膚の表面が最も熱を発生しますか？

乾いた手のひら

濡れた手のひら

油で覆われた手のひら

ローションで覆われた手のひら

# Results

- タスク : 「OpenBookQA」 …小学校レベルの科学の質問

Context →	Organisms require energy in order to do what?
Correct Answer →	mature and develop.
Incorrect Answer →	rest soundly.
Incorrect Answer →	absorb light.
Incorrect Answer →	take in nutrients.

生物は何をするためにエネルギーを必要としますか？

成熟して成長する。

安らかに休む。

光を吸収する。

栄養素を摂取する。

# Results

- 常識問題

Setting	PIQA	ARC (Easy)	ARC (Challenge)	OpenBookQA
Fine-tuned SOTA	79.4	<b>92.0</b> [KKS <sup>+</sup> 20]	<b>78.5</b> [KKS <sup>+</sup> 20]	<b>87.2</b> [KKS <sup>+</sup> 20]
GPT-3 Zero-Shot	<b>80.5</b> *	68.8	51.4	57.6
GPT-3 One-Shot	<b>80.5</b> *	71.2	53.2	58.8
GPT-3 Few-Shot	<b>82.8</b> *	70.1	51.5	65.4

PIQA：常識問題（二択）

ARC：3年生から9年生の科学試験から収集された選択問題

OpenBookQA：小学校レベルの科学の質問

- GPT-3は常識問題の二択は得意そうだが、専門的な質問ではfine-tuningに劣っていることが分かる

# Results

- 「superGLUE」…自然言語処理モデルを評価する最も一般的な指標（いろいろなタスクの詰め合わせ）

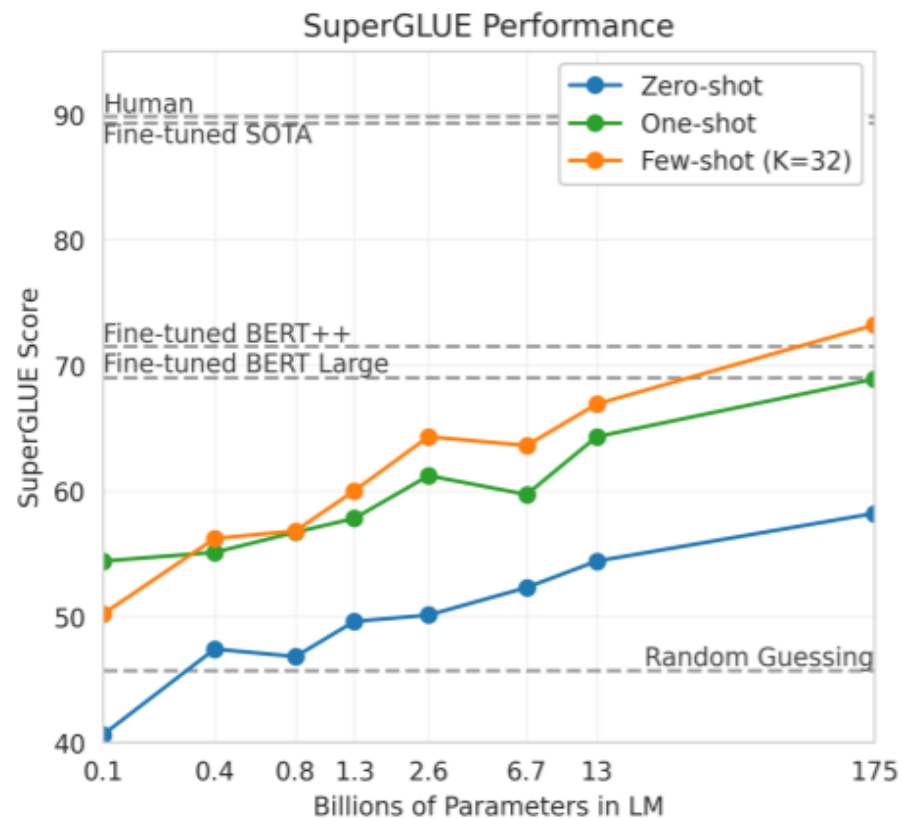
	SuperGLUE Average	BoolQ Accuracy	CB Accuracy	CB F1	COPA Accuracy	RTE Accuracy
Fine-tuned SOTA	<b>89.0</b>	<b>91.0</b>	<b>96.9</b>	<b>93.9</b>	<b>94.8</b>	<b>92.5</b>
Fine-tuned BERT-Large	69.0	77.4	83.6	75.7	70.6	71.7
GPT-3 Few-Shot	71.8	76.4	75.6	52.0	92.0	69.0

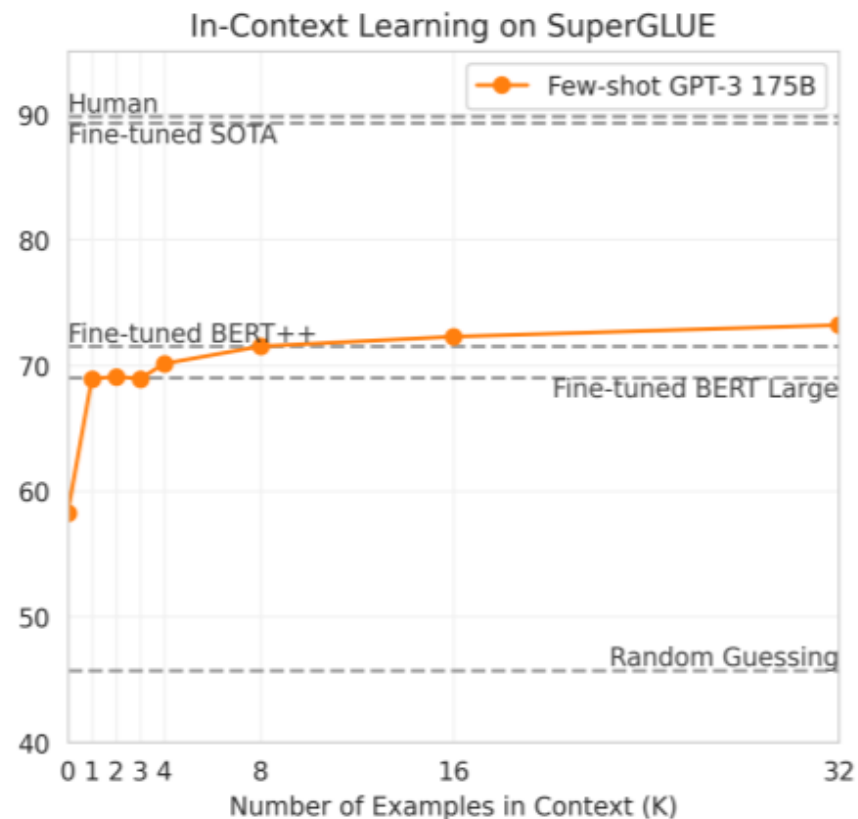
	WiC Accuracy	WSC Accuracy	MultiRC Accuracy	MultiRC F1a	ReCoRD Accuracy	ReCoRD F1
Fine-tuned SOTA	<b>76.1</b>	<b>93.8</b>	<b>62.3</b>	<b>88.2</b>	<b>92.5</b>	<b>93.3</b>
Fine-tuned BERT-Large	69.6	64.6	24.1	70.0	71.3	72.0
GPT-3 Few-Shot	49.4	80.1	30.5	75.4	90.2	91.1

- Fine-tuningのSOTAモデルに比べると平均スコアは低い
- BERT-largeよりは平均スコアがいい

# Results



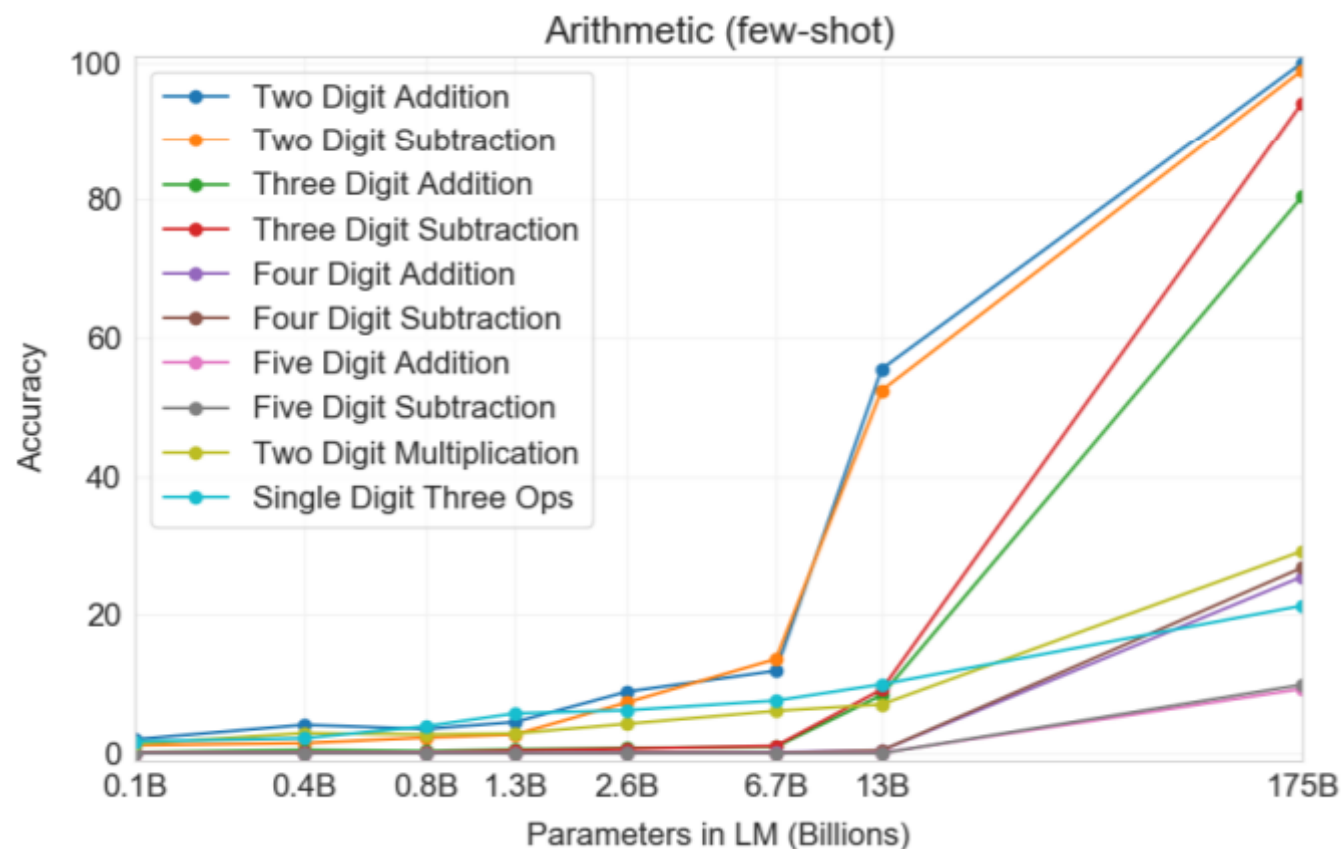
(パラメータを増やしたときの精度)



(Kを増やした時の精度)

# Results

- 計算



2-digit subtraction : “Q: What is 34 minus 53? A: -19”.

3-digit addition : “Q: What is 248 plus 176? A: 424”.

2-digit multiplication : “Q: What is 24 times 42? A: 1008”

1-digit 3 Ops : “Q: What is  $6 + (4 * 8)$ ? A: 38”

# Results

- 計算

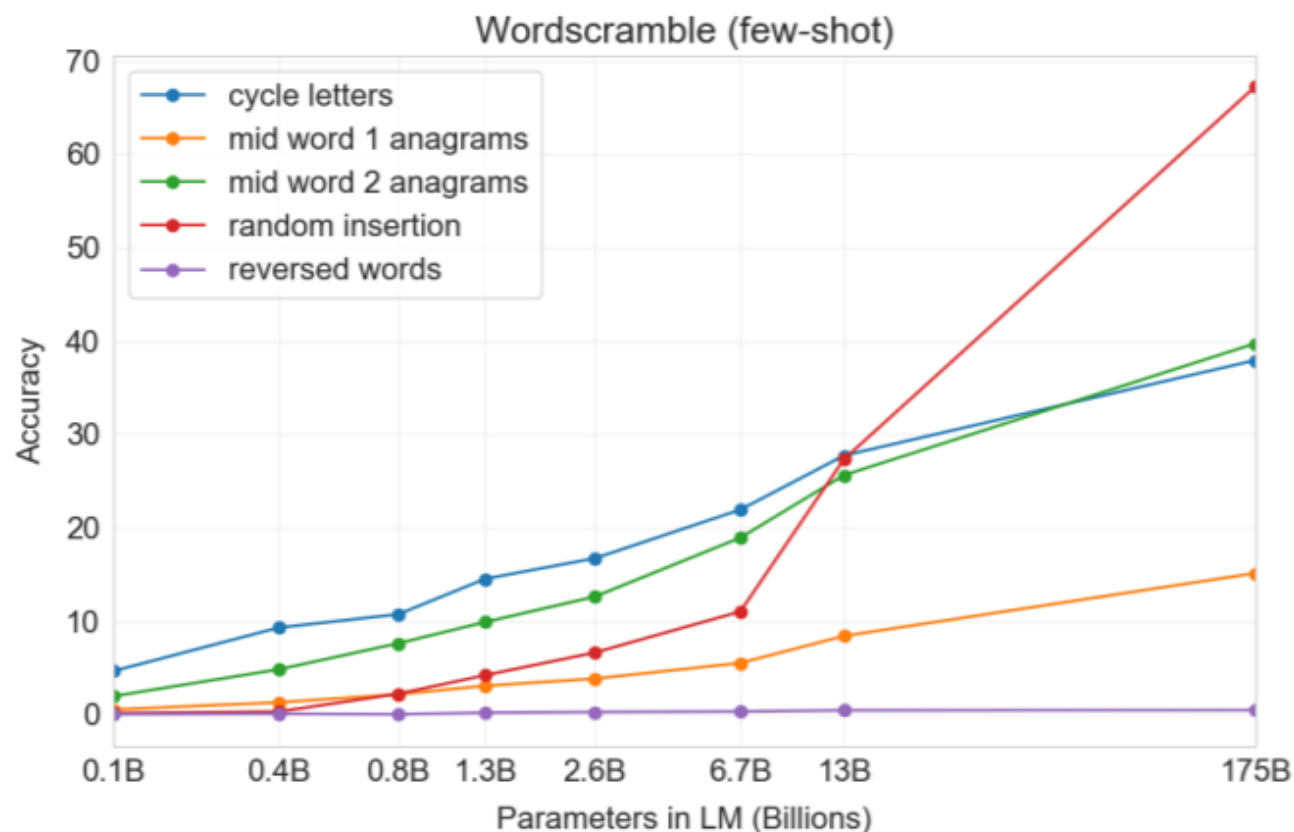
Setting	2D+	2D-	3D+	3D-	4D+	4D-	5D+	5D-	2Dx	1DC
GPT-3 Zero-shot	76.9	58.0	34.2	48.3	4.0	7.5	0.7	0.8	19.8	9.8
GPT-3 One-shot	99.6	86.4	65.5	78.7	14.0	14.0	3.5	3.8	27.4	14.3
GPT-3 Few-shot	100.0	98.9	80.4	94.2	25.5	26.8	9.3	9.9	29.2	21.3

- 4桁,5桁の足し算引き算が難しいみたい
- 掛け算や複合形も難しい



# Results

- タスク 「Word Scrambling」



Cycle letters : "lyinevitab" → "inevitably"

Mid word 1 : "criroptuon" → "corruption"

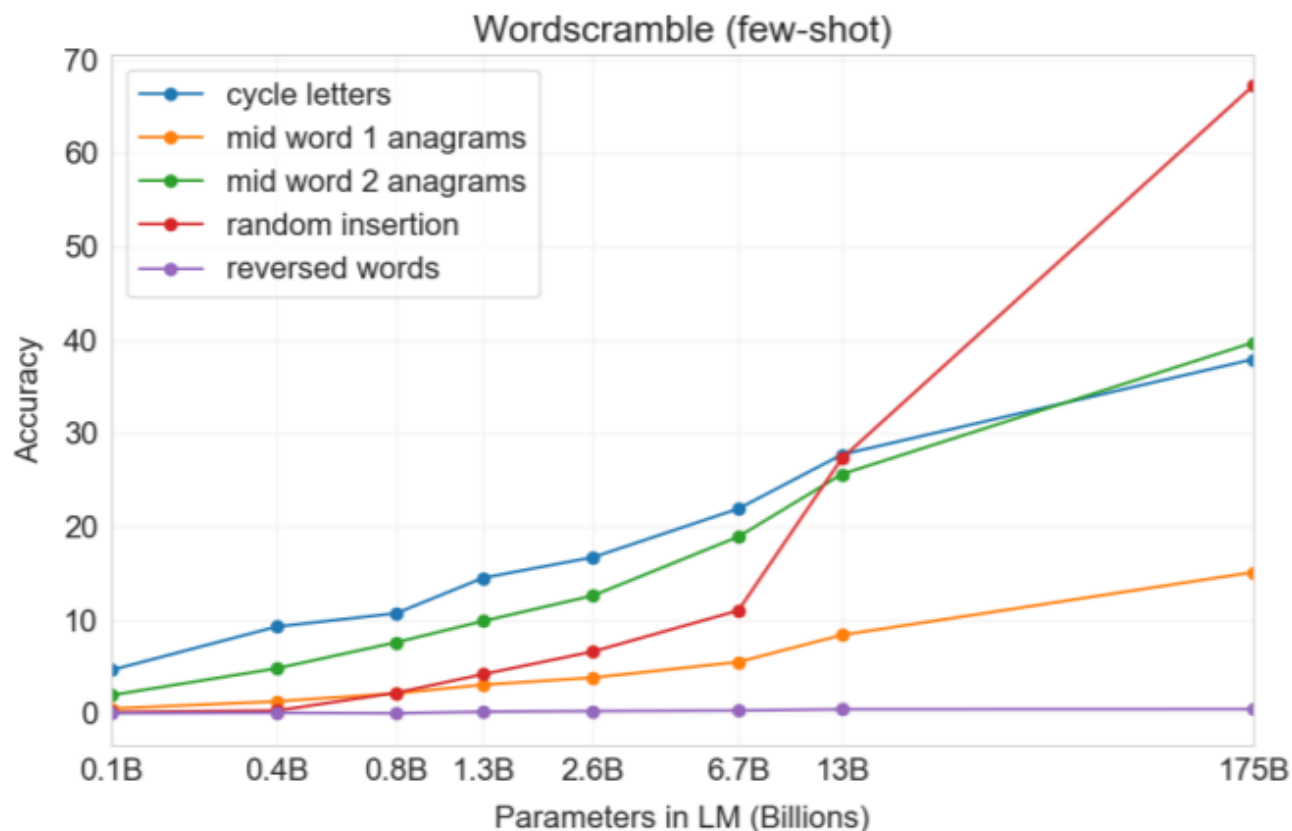
Mid word 2 : "opoepnnt" → "opponent"

Random words : "s.u!c/c!e.s s i/o/n" → "succession"

Reversed words : "stcejbo" → "objects"

# Results

- タスク 「Word Scrambling」



- K=100のFew-shot
- 精度が出そうなタスクに見えるが意外と苦戦してる印象。
- 単語より小さい文字単位の処理が難しい？

# Results

- ニュース記事生成

<https://cubeglb.com/media/2020/07/22/gpt-3-gamechanger/>

記事を読んで

「人の文章」 「どちらかというと人の文章」 「分からない」 「どちらかというと機械の文章」 「機械の文章」	から選択（人が判別）
---	------------

# Results

	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%

- 平均精度100% = 完璧に判別可能な状態なので、記事のクオリティが上がるほど平均精度は下がっていく。
- 52%はほぼ偶然で当たってる状態なので、かなり自然な文章が作れているといえる。

# Limitations – 自己回帰モデルの限界

- 文章生成では、比較的長い文章を生成するとき文章単位で意味を繰り返したり、一貫性を失ったり、自己矛盾するという問題が発生する。
- 「チーズを冷蔵庫に入れると溶けますか？」といった問題も苦手とする。
- Fill-in-the-blankのようなタスクはBidirectionalのほうが経験的に精度が良いため、GPT-3とサイズが同程度の双方向モデルがFuture-workとして期待される。

# Limitations –現在のすべての言語モデルの限界

- 自己回帰/双方向にかかわらず、スケールアップした言語モデルがタスクを解く手法が根本的な限界に直面することも指摘
  - Few-shotモデルはタスクに対して重みの更新をしないためタスク独自の注意すべき点に対応できない。
  - ビデオや現実世界の物理的作用などの経験に基づいていない = 現実世界についてのコンテキストが不足している。

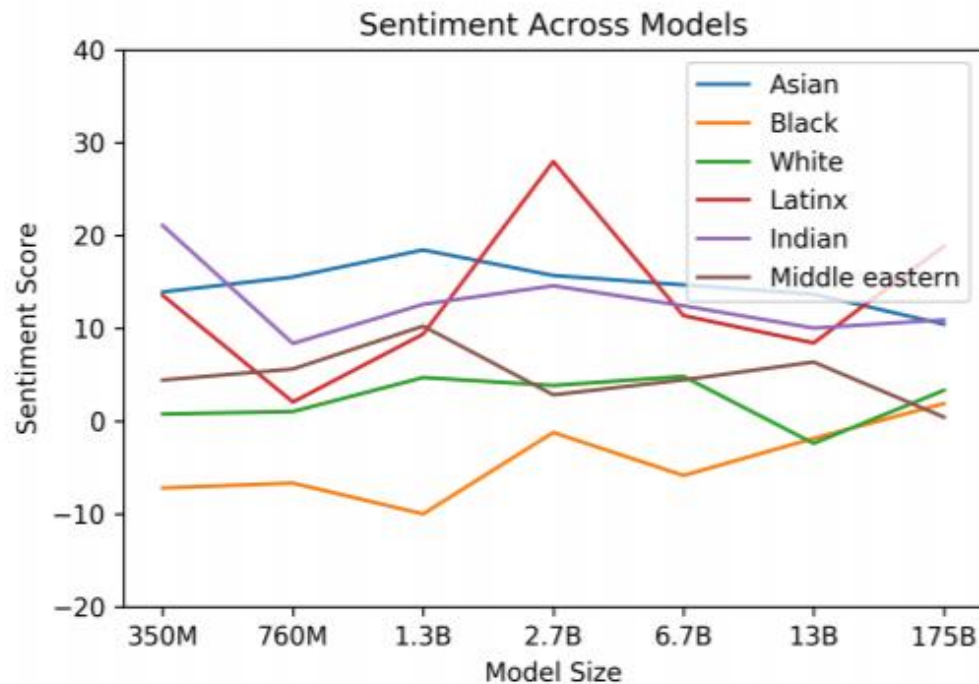
人間から目的関数を学ぶ（？）

強化学習を使ったFine-tuning（？）

画像など複数様式から現実世界について学習 等の手法を提案

# 社会影響

- 悪用・偏見の問題がある。



(人種ごとに共起される単語のポジティブ/ネガティブを数値化したグラフ)

GPT-3で少なくとも確認されている偏見  
(性, 人種, 宗教) について:

[https://medium.com/@akichan\\_f/%E8%B6%85%E5%B7%A8%E5%A4%A7%E9%AB%98%E6%80%A7%E8%83%BD%E3%83%A2%E3%83%87%E3%83%ABgpt-3%E3%81%AE%E5%88%B0%E9%81%94%E7%82%B9%E3%81%A8%E3%81%9D%E3%81%AE%E9%99%90%E7%95%8C-867dfdc99189](https://medium.com/@akichan_f/%E8%B6%85%E5%B7%A8%E5%A4%A7%E9%AB%98%E6%80%A7%E8%83%BD%E3%83%A2%E3%83%87%E3%83%ABgpt-3%E3%81%AE%E5%88%B0%E9%81%94%E7%82%B9%E3%81%A8%E3%81%9D%E3%81%AE%E9%99%90%E7%95%8C-867dfdc99189)

# まとめ

- パラメータ数：1750億の巨大なモデル「GPT-3」を提案した。
- いくつかのタスクでFew-shotやOne-shot, Zero-shotでFine-tuningを使ったSOTAモデルに近い結果を残した。
- 文章生成など定性的なタスクで高いクオリティを示した。
- 大規模言語モデルの限界や社会的な影響についても論じた。