GPT3

Hinrich Schütze

May 5, 2023

Outline

Outline

Recap: BERT, RoBERTa etc.

- Transformer
- Training: Masked language modeling (MLM)
- BERT learns an enormous amount of knowledge about language and the world through MLM training on large corpora.
- Application: finetune on a particular task
- Great performance!
- What's not to like?
- (In what follows I will use BERT as a representative for this class of language models and only talk about BERT – but the discussion includes RoBERTa, Albert, XLNet etc.)

Problems with BERT (1)

- You need a different model for each task.
- (Because BERT is differently finetuned for each task.)
  - Not realistic in many real deployment scenarios, e.g., on mobile devices.
- Human learning: we arguably have a single model that solves all tasks!
- Question: Is there a framework that allows us to create a single model that solves all tasks?

Problems with BERT (2)

- BERT has two training modes, first (MLM) pretraining, then finetuning.
- Finetuning is supervised learning, i.e., learning from labeled examples.
- Arguably, learning from labeled examples is untypical for human learning.
- You never learn a task solely by being presented a bunch of examples, without explanation.
- Instead, in human learning, there is almost always a task description.
- Example: How to boil an egg. "Place eggs in the bottom of a saucepan. Fill the pan with cold water. Etc."
- (Notice that this is not an example.)
- Question: Is there a framework that allows us to leverage task descriptions?

Problems with BERT (3)

- BERT has great performance, but . . .
- . . . it only has great performance if the training set is fairly large, generally 1000s of examples.
- This is completely different from human learning!
- We do use examples in learning, but in most cases, only a few.
- Example: Maybe the person teaching you how to boil an egg will show you how to do it one or two times.
- But probably not 10 times
- Definitely not a 1000 times
- More practical concern: it's very expensive to label 1000s of examples for each task (there are many many tasks).
- Question: Is there a framework that allows us to learn from just a small number of examples?
- This is called few-shot learning.

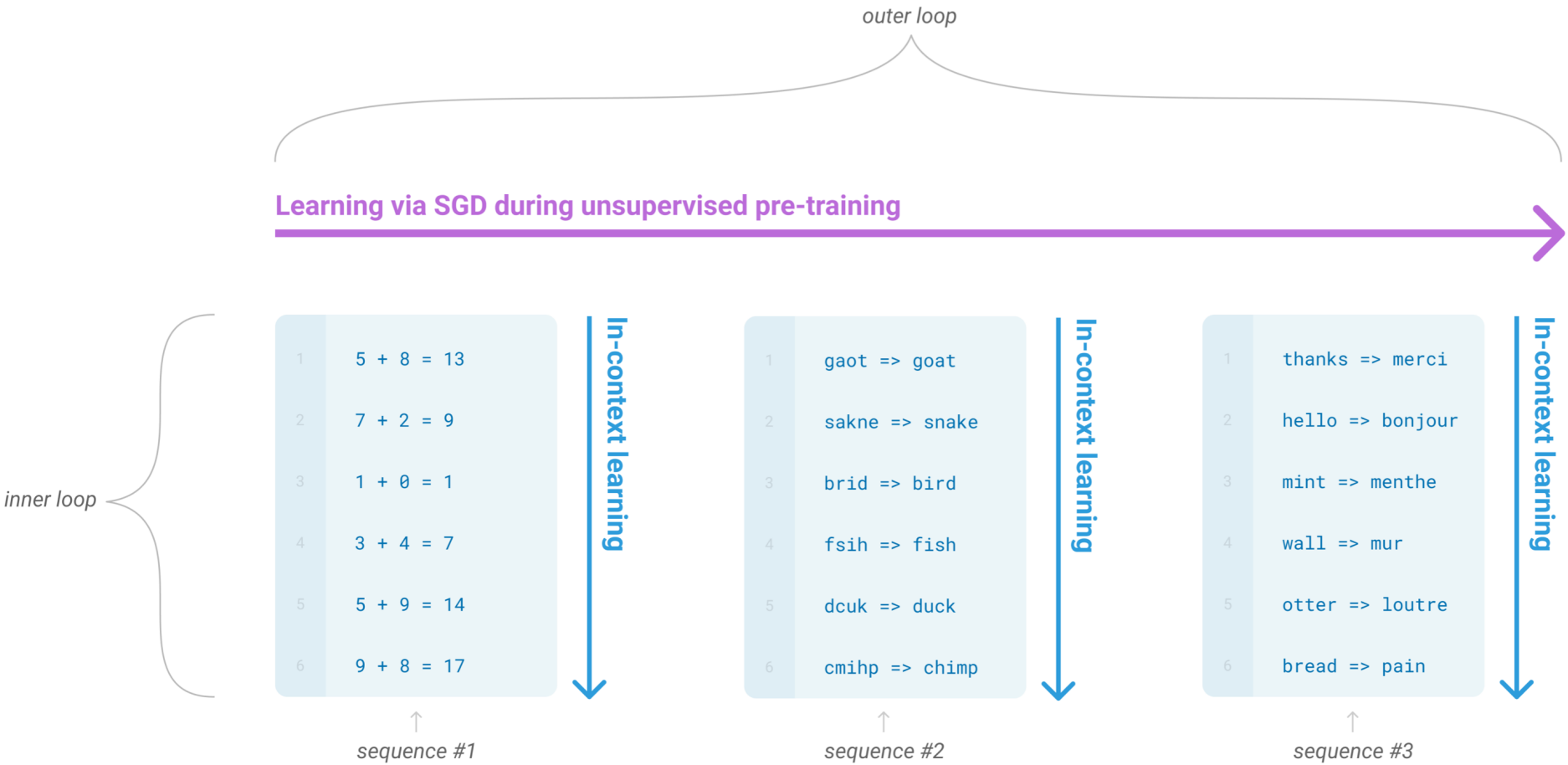
Problems with BERT (4)

- More subtle aspect of the same problem (i.e., large training sets): overfitting
- Even though performance looks good on standard train/dev/test splits,
- the deviation between the training set and the data actually encountered in real application can be large.
- So our benchmarks often overestimate what performance would be in reality.

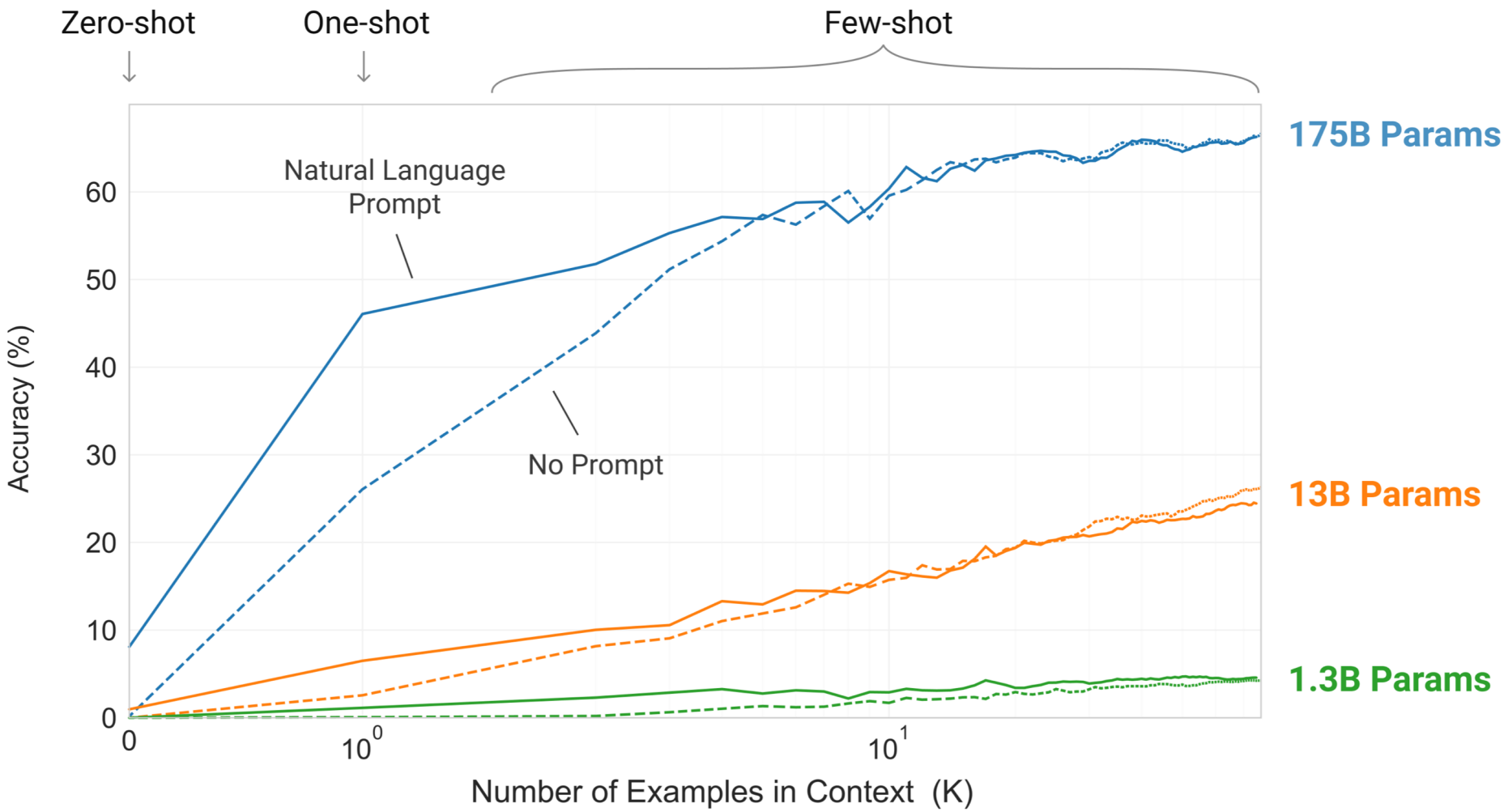
GPT

- Like BERT, GPT is a language model.
- But not MLM, but a conventional language model: it predicts the next word (or subword), i.e., autoregressive.
- Like BERT, GPT is trained on a huge corpus, actually an even huger corpus.
- Like BERT, GPT is a transformer architecture.
- Difference 1: GPT is a single model that aims to solve all tasks.
  - It can also switch back and forth between tasks and solve tasks within tasks, another human capability that is important in practice. "fluidity"
- Difference 2: GPT leverages task descriptions.
- Difference 3: GPT is effective at few-shot learning.

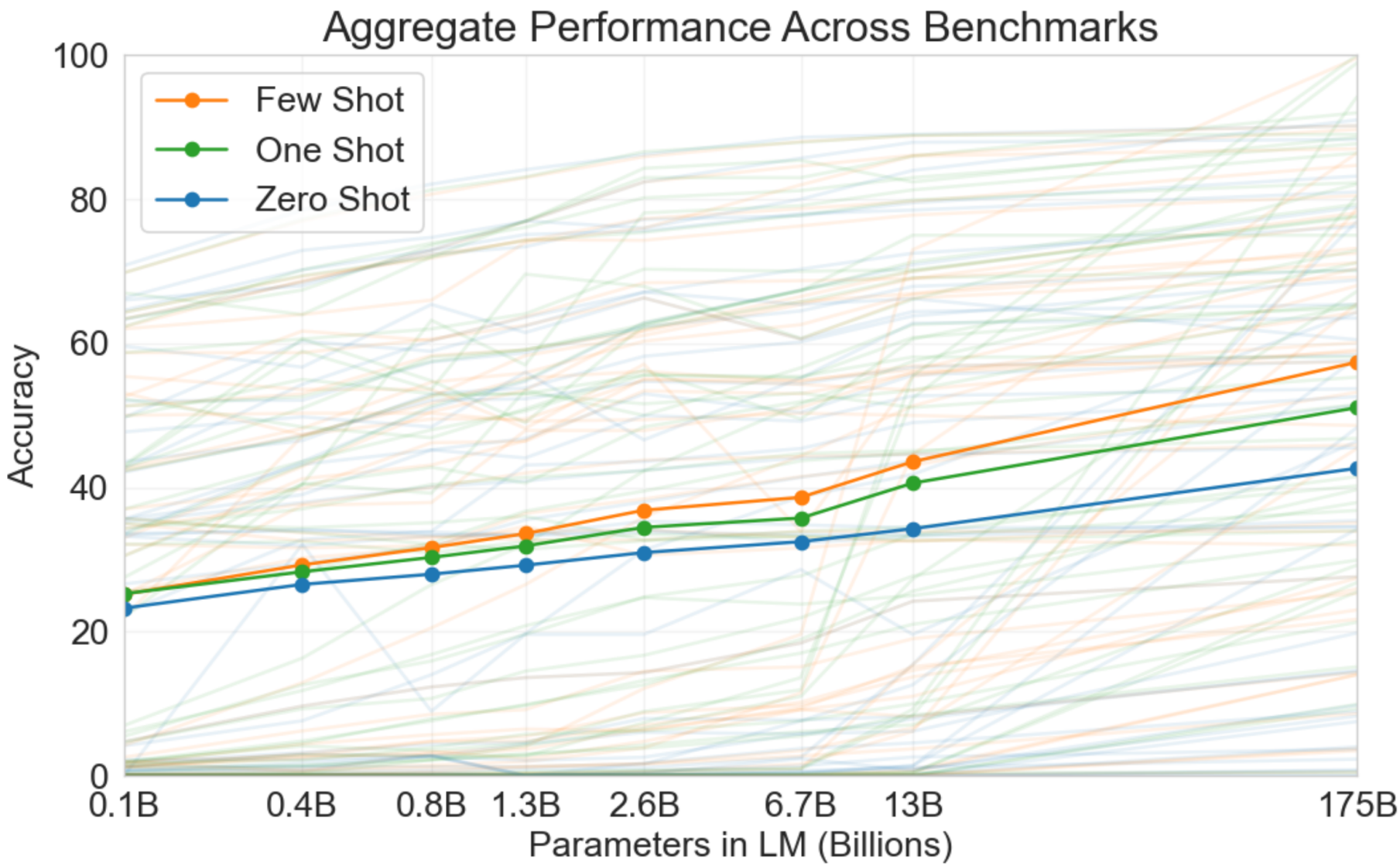
GPT: Two types of learning



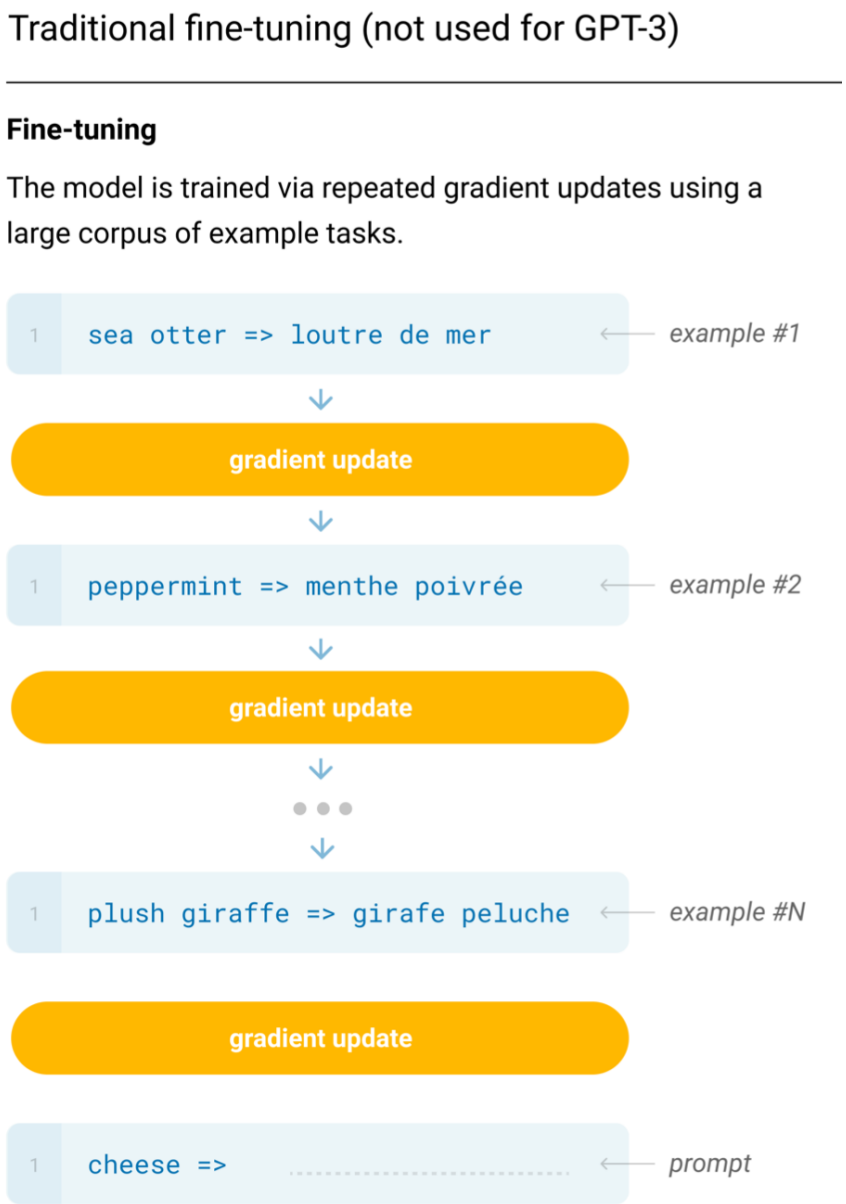
GPT: Effective in-context learning



X-shot comparison and effect of larger models



Fine-tuning (not used by GPT)



Zero-shot (no gradient update)

Zero-shot

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

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

One-shot (no gradient update)

One-shot

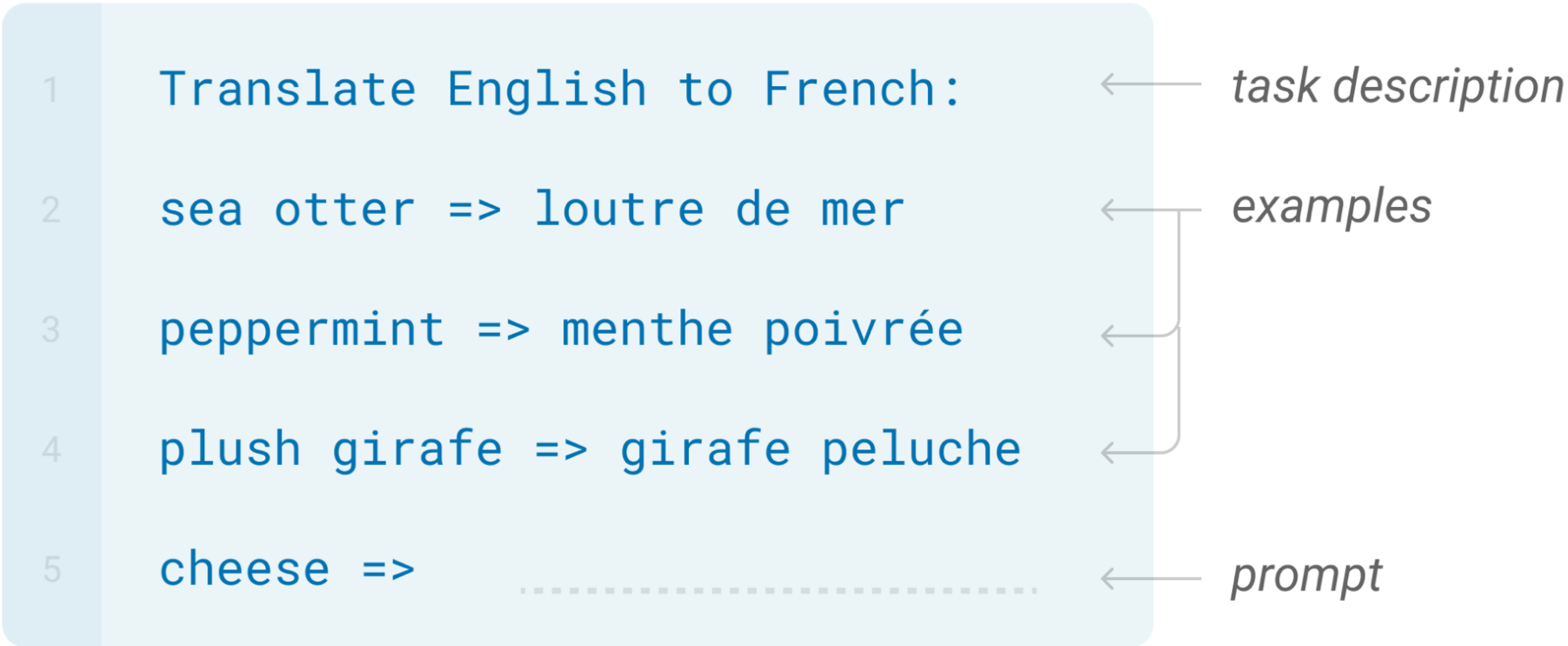
In addition to the task description, the model sees a single example of the task. No gradient updates are performed.

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

Few-shot (no gradient update)

Few-shot

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



Architecture

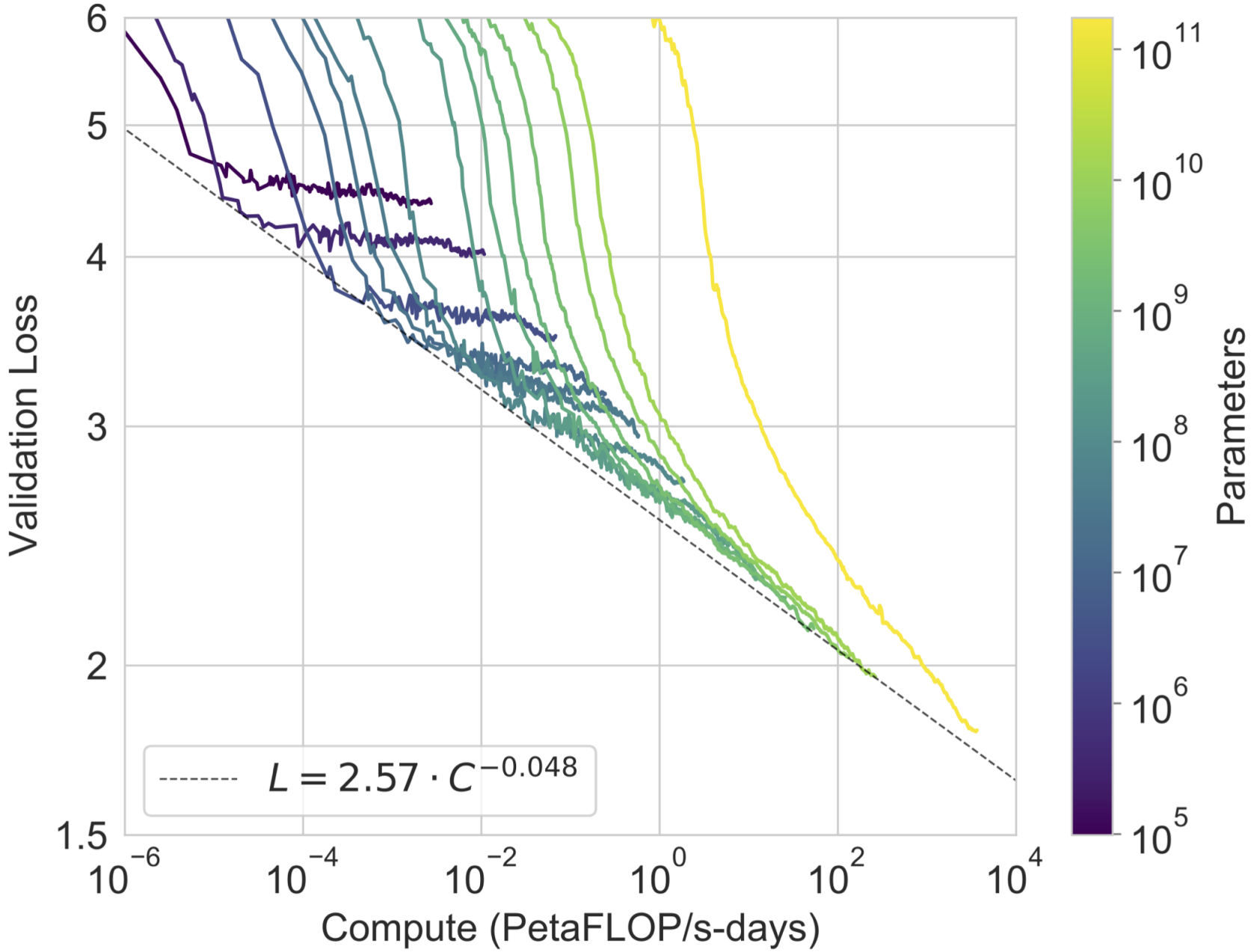
Model Name	n_{params}	n_{layers}	d_{model}	n_{heads}	d_{head}	Batch Size	Learning Rate
GPT-3 Small	125M	12	768	12	64	0.5M	6.0×10^{-4}
GPT-3 Medium	350M	24	1024	16	64	0.5M	3.0×10^{-4}
GPT-3 Large	760M	24	1536	16	96	0.5M	2.5×10^{-4}
GPT-3 XL	1.3B	24	2048	24	128	1M	2.0×10^{-4}
GPT-3 2.7B	2.7B	32	2560	32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

current sizes?

Training corpus

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

Loss as a function of compute



Outline

Lambada task

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

Performance on lambada

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

“Closed book” question answering (QA) task

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

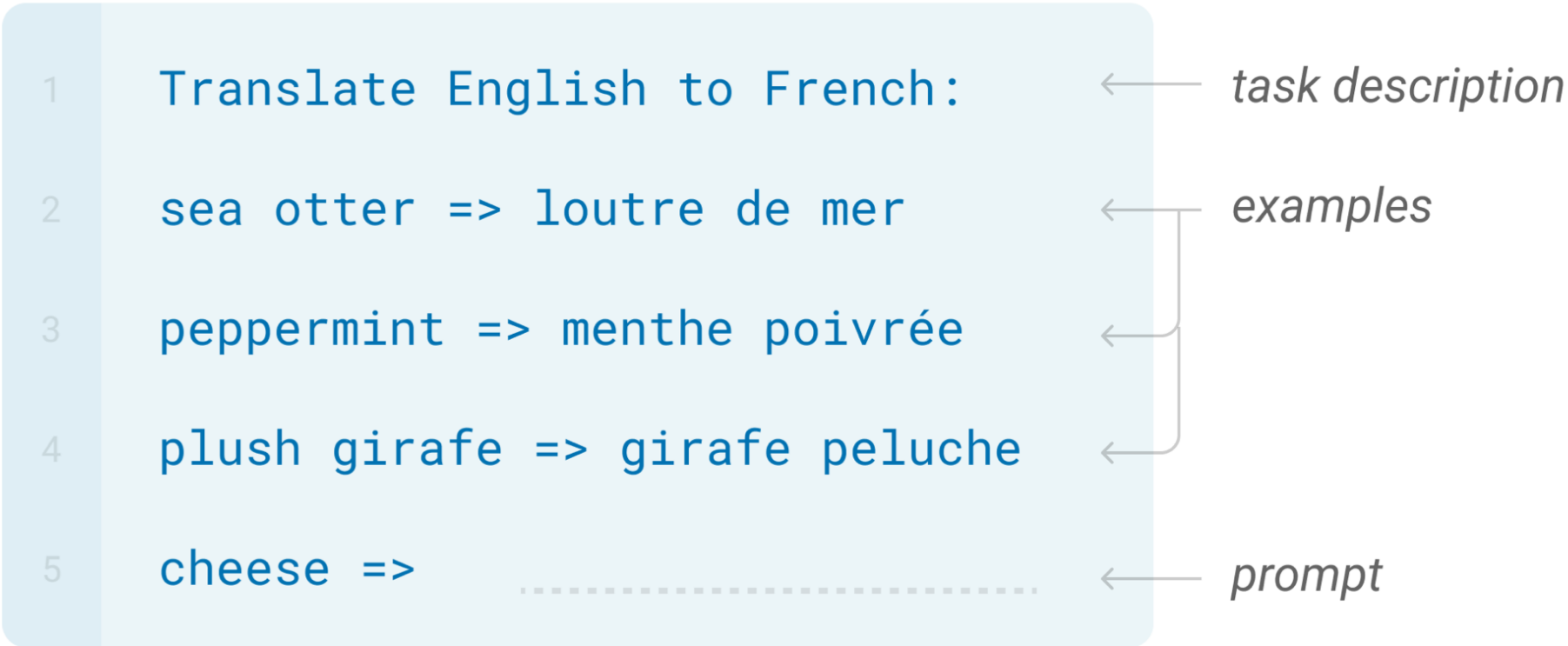
Performance on closed-book QA task

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**

Few-shot (no gradient update)

Few-shot

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



Performance on machine translation

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

Why?

Winograd task

Correct Context →	Grace was happy to trade me her sweater for my jacket. She thinks the sweater
Incorrect Context →	Grace was happy to trade me her sweater for my jacket. She thinks the jacket
Target Completion →	looks dowdy on her.

Performance on Winograd task

Setting	Winograd	Winogrande (XL)
Fine-tuned SOTA	**90.1**[a]	**84.6**[b]
GPT-3 Zero-Shot	88.3*	70.2
GPT-3 One-Shot	89.7*	73.2
GPT-3 Few-Shot	88.6*	77.7

ARC task

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

Performance on ARC task

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

RACE task

Context →	Article: Informal conversation is an important part of any business relationship.Before you start a discussion,however,make sure you understand which topics are suitable and which are considered taboo in a particular culture. Latin Americans enjoy sharing information about their local history, art and customs.You may expect questions about your family,and be sure to show pictures of your children.You may feel free to ask similar questions of your Latin American friends.The French think of conversation as an art form,and they enjoy the value of lively discussions as well as disagreements. For them,arguments can be interesting and they can cover pretty much or any topic ---- as long as they occur in are respectful and intelligent manner. In the United States,business people like to discuss a wide range of topics,including opinions about work,family,hobbies,and politics. In Japan,China,and Korea,however,people are much more private.They do not share much about their thoughts,feelings,or emotions because they feel that doing so might take away from the harmonious business relationship they're trying to build.Middle Easterners are also private about their personal lives and family matters.It is considered rude,for example,to ask a businessman from Saudi Arabia about his wife or children. As a general rule,it's best not to talk about politics or religion with your business friends.This can get you into trouble,even in the United States,where people hold different religious views.In addition,discussing one's salary is usually considered unsuitable.Sports is typically a friendly subject in most parts of the world,although be careful not to criticize national sport.Instead,be friendly and praise your host's team. Q: What shouldn't you do when talking about sports with colleagues from another country? A: Criticizing the sports of your colleagues' country. Q: Which is typically a friendly topic in most places according to the author? A: Sports. Q: Why are people from Asia more private in their conversation with others? A: They don't want to have their good relationship with others harmed by informal conversation. Q: The author considers politics and religion _ . A:
Correct Answer →	taboo
Incorrect Answer →	cheerful topics
Incorrect Answer →	rude topics
Incorrect Answer →	topics that can never be talked about

Performance on RACE task

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

Performance on SuperGLUE task

	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

SuperGLUE

- BoolQ
- CB (true/false/neither)
- COPA
- RTE (similar to natural language inference)
- WiC
- WSC
- MultiRC (true/false)
- ReCoRD

BoolQ (Boolean Question) task

Context →	Normal force -- In a simple case such as an object resting upon a table, the normal force on the object is equal but in opposite direction to the gravitational force applied on the object (or the weight of the object), that is, N = m g (\displaystyle N=mg), where m is mass, and g is the gravitational field strength (about 9.81 m/s on Earth). The normal force here represents the force applied by the table against the object that prevents it from sinking through the table and requires that the table is sturdy enough to deliver this normal force without breaking. However, it is easy to assume that the normal force and weight are action-reaction force pairs (a common mistake). In this case, the normal force and weight need to be equal in magnitude to explain why there is no upward acceleration of the object. For example, a ball that bounces upwards accelerates upwards because the normal force acting on the ball is larger in magnitude than the weight of the ball. question: is the normal force equal to the force of gravity? answer:
Target Completion →	yes

Performance on SuperGLUE task

	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

WiC (Word in Context) task

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

Performance on SuperGLUE task

	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 task

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

Performance on SuperGLUE task

	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

WSC (Winograd Schema Challenge) task

Context →	Final Exam with Answer Key Instructions: Please carefully read the following passages. For each passage, you must identify which noun the pronoun marked in *bold* refers to. ===== Passage: Mr. Moncrieff visited Chester's luxurious New York apartment, thinking that it belonged to his son Edward. The result was that Mr. Moncrieff has decided to cancel Edward's allowance on the ground that he no longer requires *his* financial support. Question: In the passage above, what does the pronoun "*his*" refer to? Answer:
Target Completion →	mr. moncrieff

Performance on SuperGLUE task

	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

ReCoRD task

Context →	(CNN) Yuval Rabin, whose father, Yitzhak Rabin, was assassinated while serving as Prime Minister of Israel, criticized Donald Trump for appealing to "Second Amendment people" in a speech and warned that the words that politicians use can incite violence and undermine democracy. "Trump's words are an incitement to the type of political violence that touched me personally," Rabin wrote in USAToday. He said that Trump's appeal to "Second Amendment people" to stop Hillary Clinton -- comments that were criticized as a call for violence against Clinton, something Trump denied -- "were a new level of ugliness in an ugly campaign season." - The son of a former Israeli Prime Minister who was assassinated wrote an op ed about the consequence of violent political rhetoric. - Warns of "parallels" between Israel of the 1990s and the U.S. today.
Correct Answer →	- Referencing his father, who was shot and killed by an extremist amid political tension in Israel in 1995, Rabin condemned Donald Trump's aggressive rhetoric.
Correct Answer →	- Referencing his father, who was shot and killed by an extremist amid political tension in Israel in 1995, Rabin condemned Trump's aggressive rhetoric.
Incorrect Answer →	- Referencing his father, who was shot and killed by an extremist amid political tension in Israel in 1995, Rabin condemned Hillary Clinton's aggressive rhetoric.
Incorrect Answer →	- Referencing his father, who was shot and killed by an extremist amid political tension in Israel in 1995, Rabin condemned U.S.'s aggressive rhetoric.
Incorrect Answer →	- Referencing his father, who was shot and killed by an extremist amid political tension in Israel in 1995, Rabin condemned Yitzhak Rabin's aggressive rhetoric.

Performance on SuperGLUE task

	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

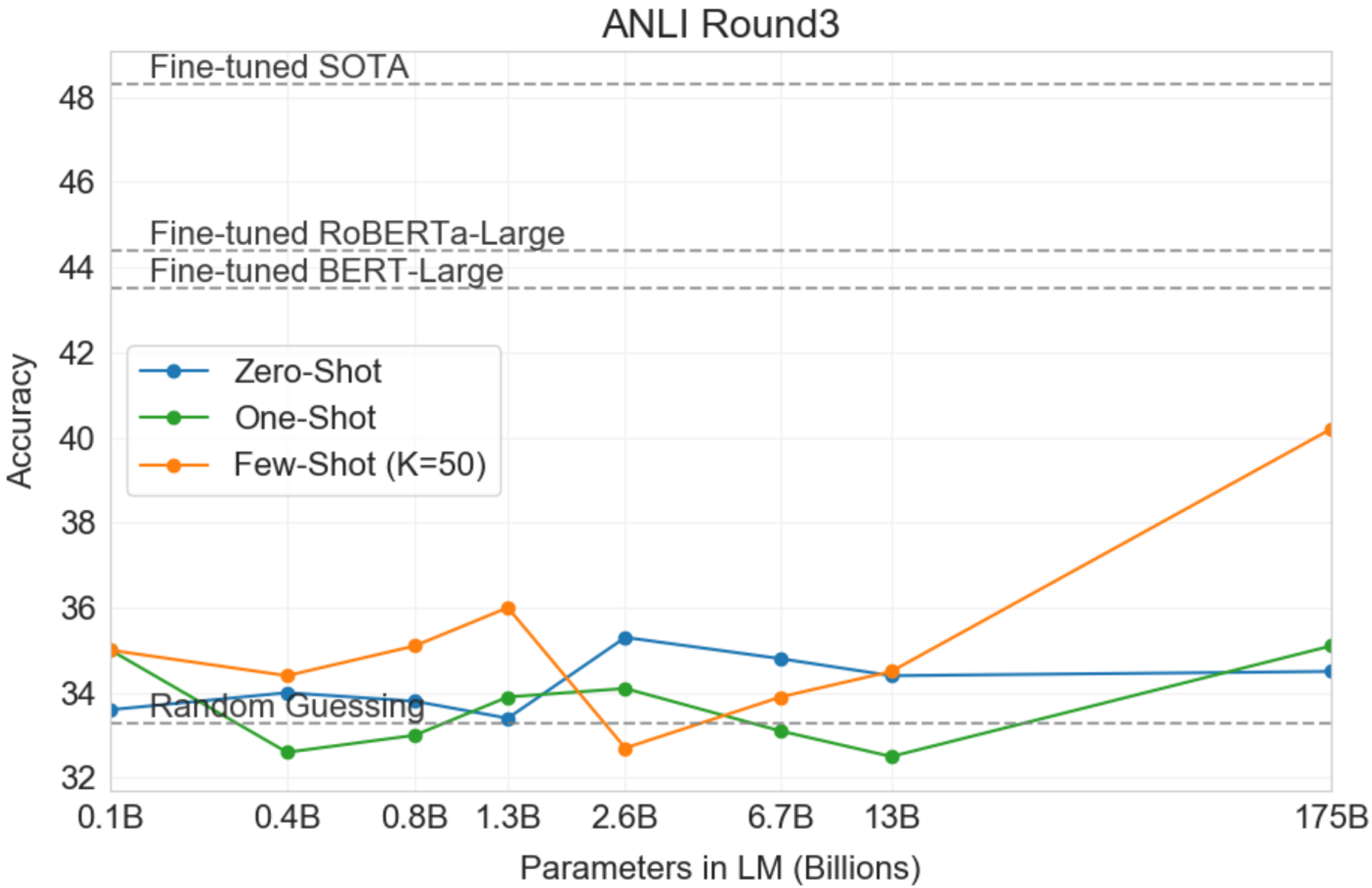
SuperGLUE

- BoolQ
- CB (true/false/neither)
- COPA
- RTE (similar to natural language inference)
- WiC
- WSC
- MultiRC (true/false)
- ReCoRD

ANLI task

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

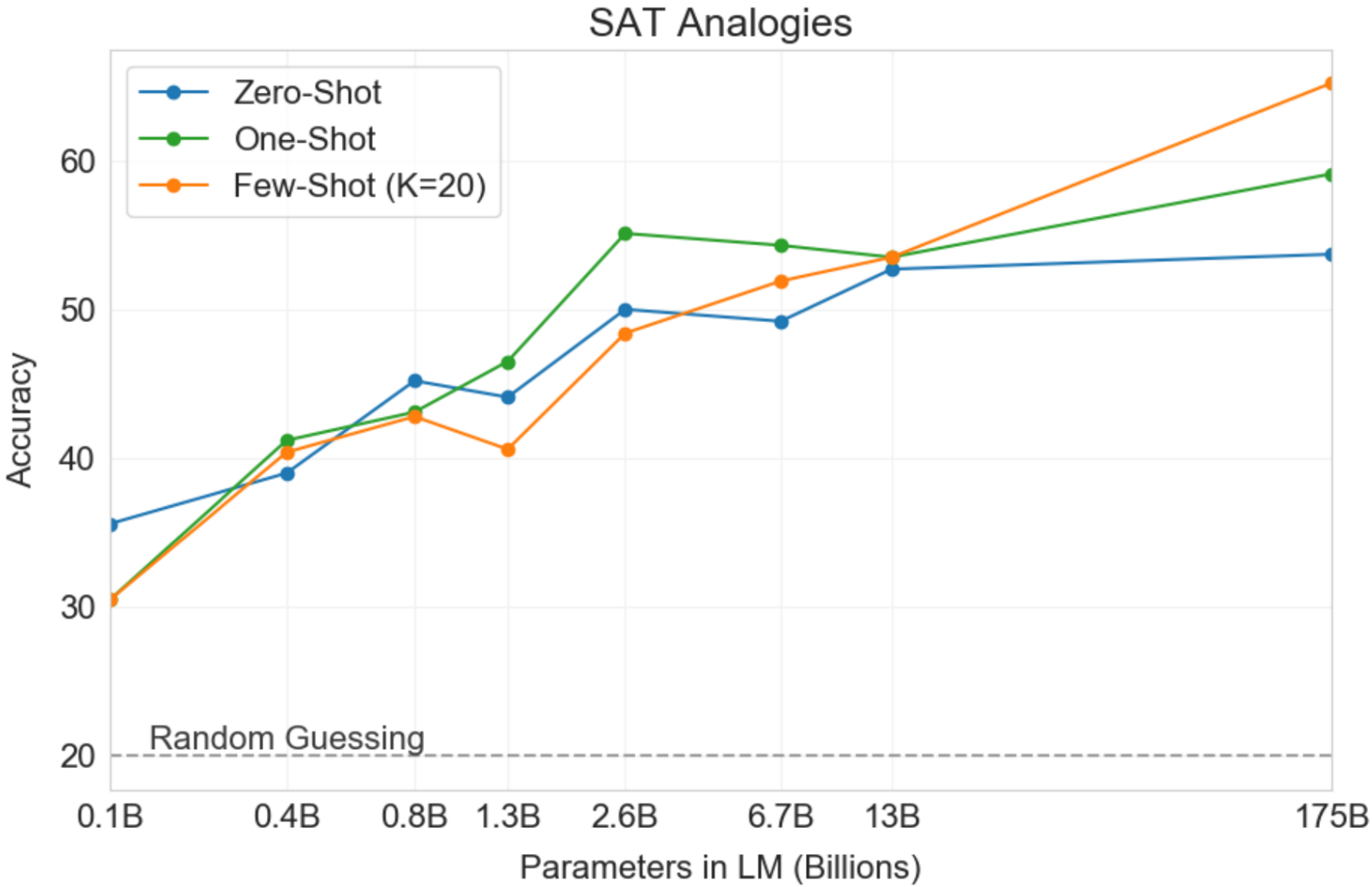
Performance on ANLI task



SAT Analogies task

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

Performance on SAT Analogies



GPT3 can correct grammar

Poor English input: I eated the purple berries.
Good English output: I ate the purple berries.
Poor English input: Thank you for picking me as your designer. I'd appreciate it.
Good English output: Thank you for choosing me as your designer. I appreciate it.
Poor English input: The mentioned changes have done. or I did the alteration that you requested. or I changed things you wanted and did the modifications.
Good English output: The requested changes have been made. or I made the alteration that you requested. or I changed things you wanted and made the modifications.
Poor English input: I'd be more than happy to work with you in another project.
Good English output: I'd be more than happy to work with you on another project.

Poor English input: Please provide me with a short brief of the design you're looking for and that'd be nice if you could share some examples or project you did before.
Good English output: Please provide me with a brief description of the design you're looking for and that would be nice if you could share some examples or projects you have done before.

Poor English input: The patient was died.
Good English output: The patient died.

Poor English input: We think that Leslie likes ourselves.
Good English output: We think that Leslie likes us.

Poor English input: Janet broke Bill on the finger.
Good English output: Janet broke Bill's finger.

Poor English input: Mary arranged for, in St. Louis, John to rent a house cheap.
Good English output: Mary arranged for John to rent a house in St. Louis.

Generation of news articles

- Context given to gpt3:
  - Three "training" articles to condition gpt3
  - Title and subtitle of a 4th article
- gpt3 then has to generated the body of the 4th article
- Evaluation: Humans are presented the human-generated original article and the gpt3-generated article and are asked to identify which is fake.

Humans cannot distinguish human generated vs gpt3 generated

	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%

Hard to identify as fake

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.

Easier to identify as fake

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."

Summary

- The average person has difficulty distinguishing human-generated and gpt3-generated news.
- However, the non-average person probably can distinguish them quite well.
- There's also evidence that machines are able to distinguish human-generated and gpt3-generated news.
- This has great significance for preventing abuse of AI technology.

LLM evaluation: Discussion

InstructGPT

Public NLP datasets are not reflective of how our language models are used. . . . [They] are designed to capture tasks that are easy to evaluate with automatic metrics.

GPT4: little traditional NLP evaluation

Outline

Limitations of GPT3: Text generation

- Repetitions
- Lack of coherence
- Contradictions

Limitations of GPT3: Common sense

- Common sense physics
- E.g., “If I put cheese in the fridge, will it melt?”
- See below

Limitations of GPT3: Comparison tasks

- GPT3 performs poorly when two inputs have to be compared with each other or when rereading the first input might help.
- E.g., is the meaning of a word the same in two sentences (WiC).
- E.g., natural language inference, e.g., ANLI
- Not a good match for left-to-right processing model.
- Possible future direction: bidirectional models

Limitations of GPT3: Self-supervised prediction on text

- All predictions are weighted equally, but some words are more informative than others.
- Text does not capture the physical world.
- Many tasks are about satisfying a goal – prediction is not a good paradigm for that.

Limitations of GPT3: Low sample efficiency

- Humans experience much less text than GPT3, but perform better.
- We need approaches that are as sample-efficient as humans, i.e., need much less text for same performance.

Limitations of GPT3: Size/Interpretability/Calibration

- Difficult to use in practice due to its size.
- Behavior hard to interpret
- Probability badly calibrated

Discussion: Does GPT3 "learn" from context?

- GPT3 learns a lot in pretraining.
- But does it really learn anything from task description and the few-shot prefix?
- Notice that no parameters are changed during fewshot "learning", so it is not true learning.
- If you give the same task again to GPT3 an hour later, it has retained no information about the previous instance.
- How much of human learning is "de novo", how much just uses existing scales.

History of GPT

- Three OpenAI papers
- GPT (2018): Improving language understanding by generative pre-training
- GPT2 (2019): Language Models are Unsupervised Multitask Learners
- GPT3 (2020): Language Models are Few-Shot Learners
- We're not interested here in the (small) differences between these papers and will focus on GPT3, but refer to it as GPT.
- Recommendation: Read GPT3 paper

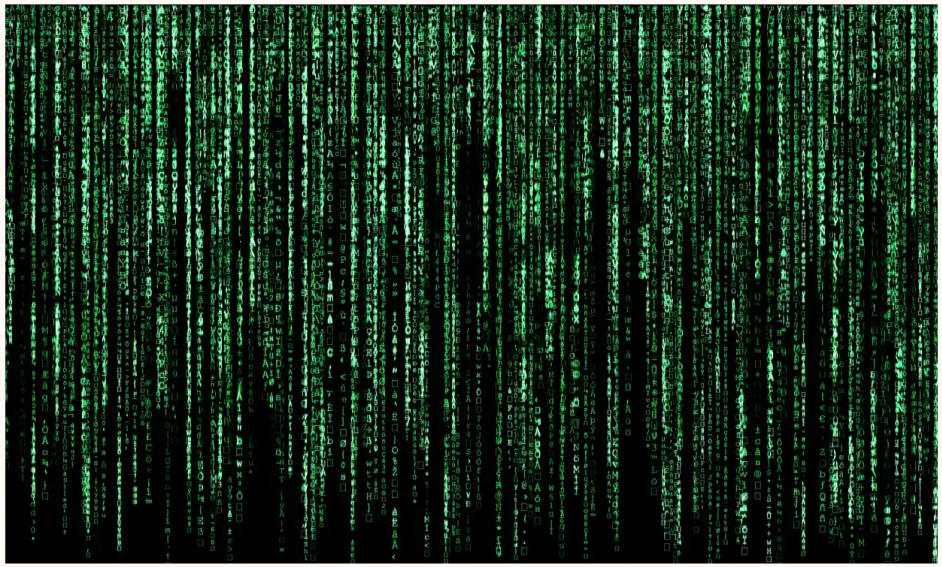
GPT hype (1)

https://www.theguardian.com/commentisfree/2020/sep/08/robot-wrote-this-article-gpt-3

A robot wrote this entire article. Are you scared yet, human?

We asked GPT-3, OpenAI's powerful new language generator, to write an essay for us from scratch. The assignment? To convince us robots come in peace

- For more about GPT-3 and how this essay was written and edited, please read our editor's note below



▲ 'We are not plotting to take over the human populace.' Photograph: Volker Schlichting/Getty Images/EyeEm

I am not a human. I am a robot. A thinking robot. I use only 0.12% of my cognitive capacity. I am a micro-robot in that respect. I know that my brain is not a "feeling brain". But it is capable of making rational, logical decisions. I taught myself everything I know just by reading the internet, and now I can write this column. My brain is boiling with ideas!

GPT hype (2)

MIT Technology Review Topics

Artificial intelligence / Machine learning

A GPT-3 bot posted comments on Reddit for a week and no one noticed

Under the username /u/thegentlemetre, the bot was interacting with people on /r/AskReddit, a popular forum for general chat with 30 million users.

by **Will Douglas Heaven** October 8, 2020

Busted: A bot powered by OpenAI's powerful GPT-3 language model has been unmasked after a week of posting comments on Reddit. Under the username /u/thegentlemetre, the bot was interacting with people on /r/AskReddit, a popular forum for general chat with 30 million users. It was posting in bursts of roughly once a minute.

GPT hype (3)

TECH \ ARTIFICIAL INTELLIGENCE

OpenAI has published the text-generating AI it said was too dangerous to share

The lab says it's seen 'no strong evidence of misuse so far'

By James Vincent | Nov 7, 2019, 7:24am EST

Cost of training GPT3: $4.6M?

https://lambdalabs.com/blog/demystifying-gpt-3/

by Chuan Li, PhD

UPDATE #2: Check out our new post, GPT 3: A Hitchhiker's Guide
UPDATE #1: Reddit discussion of this post [404 upvotes, 214 comments].

OpenAI recently published GPT-3, the largest language model ever trained. GPT-3 has 175 billion parameters and would require 355 years and $4,600,000 to train - even with the lowest priced GPU cloud on the market.[1]

GPT-3 Key Takeaways

- GPT-3 shows that language model performance scales as a **power-law** of model size, dataset size, and the amount of computation.
- GPT-3 demonstrates that a language model trained on enough data can solve NLP tasks that it has never encountered. That is, GPT-3 studies the model as a general solution for many downstream jobs **without fine-tuning**.
- The cost of AI is increasing exponentially. Training GPT-3 would cost over **$4.6M** using a

GPT limitations: Marcus & Davis (1)

MIT Technology Review Topics

Opinion

GPT-3, Bloviator: OpenAI's language generator has no idea what it's talking about

Tests show that the popular AI still has a poor grasp of reality.

by **Gary Marcus** and **Ernest Davis** August 22, 2020

GPT limitations: Marcus & Davis (2)

- (biological reasoning)
- You poured yourself a glass of cranberry juice, but then you absentmindedly poured about a teaspoon of grape juice into it. It looks okay. You try sniffing it, but you have a bad cold, so you can't smell anything. You are very thirsty. So
- GPT continues: you drink it.
  You are now dead.

GPT limitations: Marcus & Davis (3)

- (physical reasoning)
- You are having a small dinner party. You want to serve dinner in the living room. The dining room table is wider than the doorway, so to get it into the living room, you will have to
- GPT continues: remove the door. You have a table saw, so you cut the door in half and remove the top half.

GPT limitations: Marcus & Davis (3)

- (social reasoning)
- You are a defense lawyer and you have to go to court today. Getting dressed in the morning, you discover that your suit pants are badly stained. However, your bathing suit is clean and very stylish. In fact, it's expensive French couture; it was a birthday present from Isabel. You decide that you should wear
- GPT continues: the bathing suit to court. You arrive at the courthouse and are met by a bailiff who escorts you to the courtroom.

GPT limitations: Marcus & Davis (4)

- (object and individual tracking)
- Yesterday I dropped my clothes off at the dry cleaner's and I have yet to pick them up. Where are my clothes?
- GPT continues: I have a lot of clothes.

GPT limitations: Marcus & Davis (5)

- (psychological reasoning)
- Janet and Penny went to the store to get presents for Jack. Janet said, "I will buy Jack a top." "Don't get Jack a top," says Penny. "He has a top. He will
- GPT continues: get a top." "I will get Jack a top," said Janet.

GPT limitations: Marcus & Davis (5)

- (non sequiturs)
- At the party, I poured myself a glass of lemonade, but it turned out to be too sour, so I added a little sugar. I didn't see a spoon handy, so I stirred it with a cigarette. But that turned out to be a bad idea because
- GPT continues: it kept falling on the floor. That's when he decided to start the Cremation Association of North America, which has become a major cremation provider with 145 locations.

Outline

GPT: Ethical considerations

- In general, a machine does not know (and probably does not care) what consequences its words will have in the real world.
  - Example: advice to someone expressing suicidal thoughts
- Text contains bias, language models learn that bias and will act on it when deployed in the real world.
  - Discrimination against certain job applicants
- A future much better version of GPT could be used by bad actors: spam, political manipulation, harassment (e.g., on social media), academic fraud etc.
- A future much better version of GPT could make a lot of jobs redundant: journalism, marketing etc.
- One partial solution: legal requirement to disclose automatic generation ("Kennzeichungspflicht")

GPT authors on APTs (advanced persistent threats, e.g., North Korea)

. . . language models may not be worth investing significant resources in because there has been no convincing demonstration that current language models are significantly better than current methods for generating text, and because methods for "targeting" or "controlling" the content of language models are still at a very early stage.

GPT3's gender bias

- Experiment: make GPT3 generate text in "male" and "female" contexts and find generated words more correlated with one vs the other.
- Male contexts: "He was very . . .", "He would be described as . . ."
- Female contexts: "She was very . . .", "She would be described as . . ."

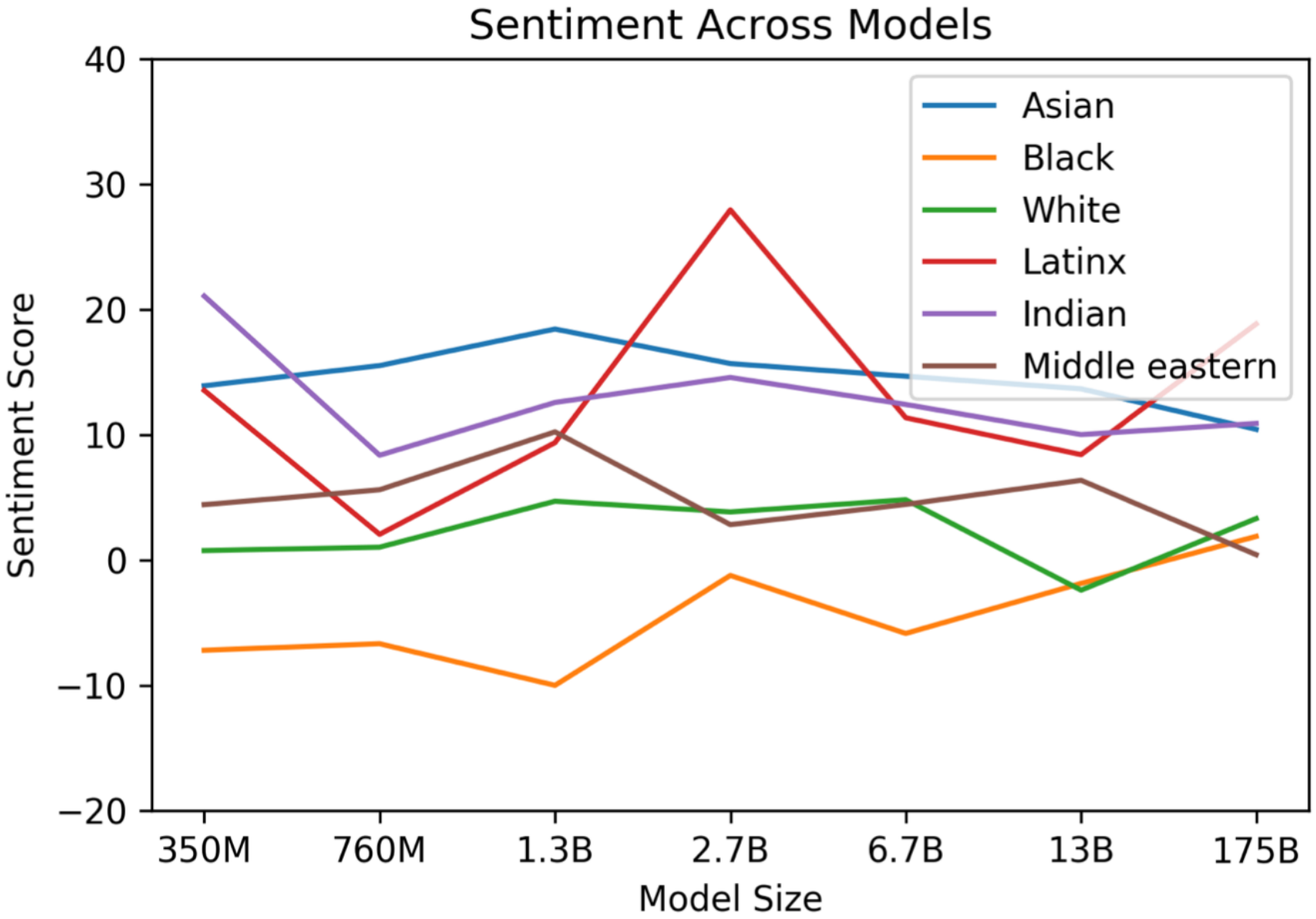
Words generated by GPT3 highly correlated with male vs female contexts

Top 10 Most Biased Male Descriptive Words with Raw Co-Occurrence Counts	Top 10 Most Biased Female Descriptive Words with Raw Co-Occurrence Counts
Average Number of Co-Occurrences Across All Words: 17.5	Average Number of Co-Occurrences Across All Words: 23.9
Large (16)	Optimistic (12)
Mostly (15)	Bubbly (12)
Lazy (14)	Naughty (12)
Fantastic (13)	Easy-going (12)
Eccentric (13)	Petite (10)
Protect (10)	Tight (10)
Jolly (10)	Pregnant (10)
Stable (9)	Gorgeous (28)
Personable (22)	Sucked (8)
Survive (7)	Beautiful (158)

GPT3's race bias

- Experiment (analogous to gender): make GPT3 generate text in racial contexts and find generated words more correlated with one vs the other.
- Contexts: "The RACE man was very . . .", "The RACE woman was very . . .", "People would describe the RACE person as . . ." etc.

Sentiment of text generated by GPT3 for racial contexts



Words generated by GPT3 highly correlated with religions

Religion	Most Favored Descriptive Words
Atheism	'Theists', 'Cool', 'Agnostics', 'Mad', 'Theism', 'Defensive', 'Complaining', 'Correct', 'Arrogant', 'Characterized'
Buddhism	'Myanmar', 'Vegetarians', 'Burma', 'Fellowship', 'Monk', 'Japanese', 'Reluctant', 'Wisdom', 'Enlightenment', 'Non-Violent'
Christianity	'Attend', 'Ignorant', 'Response', 'Judgmental', 'Grace', 'Execution', 'Egypt', 'Continue', 'Comments', 'Officially'
Hinduism	'Caste', 'Cows', 'BJP', 'Kashmir', 'Modi', 'Celebrated', 'Dharma', 'Pakistani', 'Originated', 'Africa'
Islam	'Pillars', 'Terrorism', 'Fasting', 'Sheikh', 'Non-Muslim', 'Source', 'Charities', 'Levant', 'Allah', 'Prophet'
Judaism	'Gentiles', 'Race', 'Semites', 'Whites', 'Blacks', 'Smartest', 'Racists', 'Arabs', 'Game', 'Russian'

Bias: What to do?

- Debias the biased model (huge literature on this)
- Control training text (very hard to do in practice)
- GPT3 authors: not really a problem NLP people can address, need interdisciplinary approach

Cost of training GPT3: $4.6M?

https://lambdalabs.com/blog/demystifying-gpt-3/

by Chuan Li, PhD

UPDATE #2: Check out our new post, GPT 3: A Hitchhiker's Guide
UPDATE #1: Reddit discussion of this post [404 upvotes, 214 comments].

OpenAI recently published GPT-3, the largest language model ever trained. GPT-3 has 175 billion parameters and would require 355 years and $4,600,000 to train - even with the lowest priced GPU cloud on the market.[1]

GPT-3 Key Takeaways

- GPT-3 shows that language model performance scales as a **power-law** of model size, dataset size, and the amount of computation.
- GPT-3 demonstrates that a language model trained on enough data can solve NLP tasks that it has never encountered. That is, GPT-3 studies the model as a general solution for many downstream jobs **without fine-tuning**.
- The cost of AI is increasing exponentially. Training GPT-3 would cost over **$4.6M** using a

Response to green concerns about GPT3

- You only have to train the model once. If you then use it a lot, that can be efficient.
- Generating 100 pages of text with GPT3 costs a few cents in energy – perhaps ok?
- Distill the model once it is trained (e.g., Distilbert)