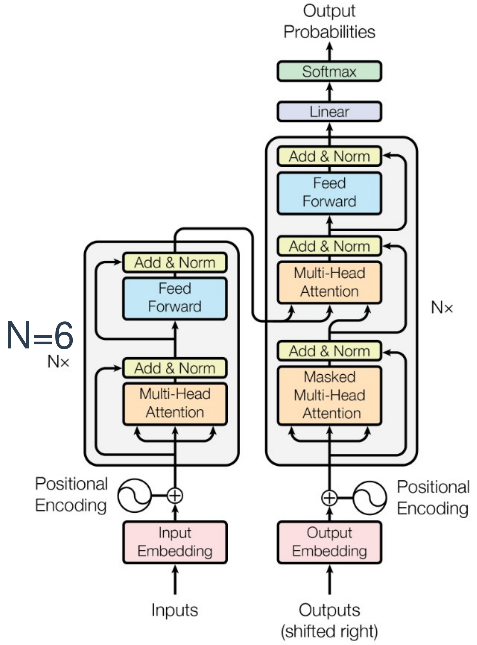
GPT-2 and GPT-3

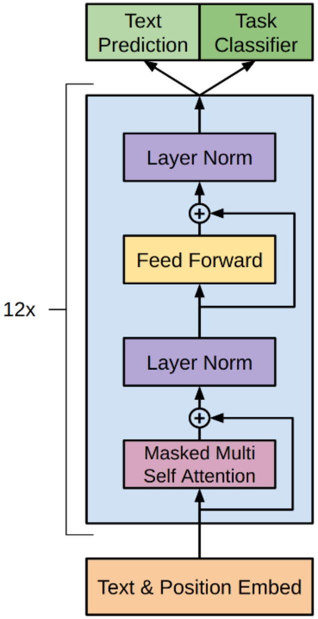
Why size > supervision?

Supervised (machine translation)



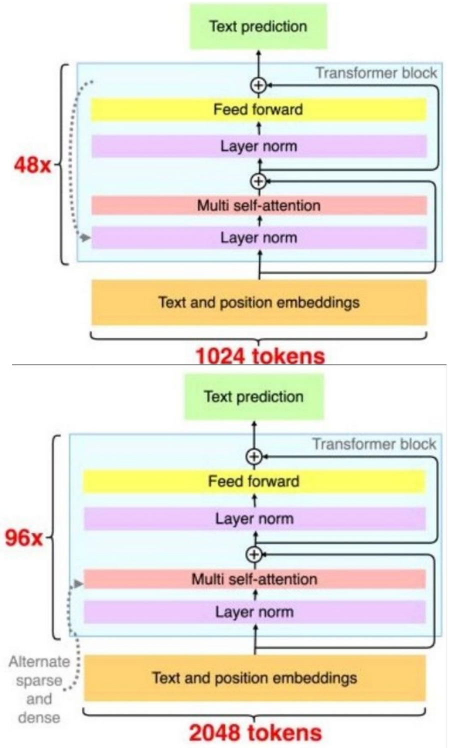
OG Transformer
65M

Unsupervised (LM) + Supervised (many)



GPT-1
117M

Unsupervised (LM)



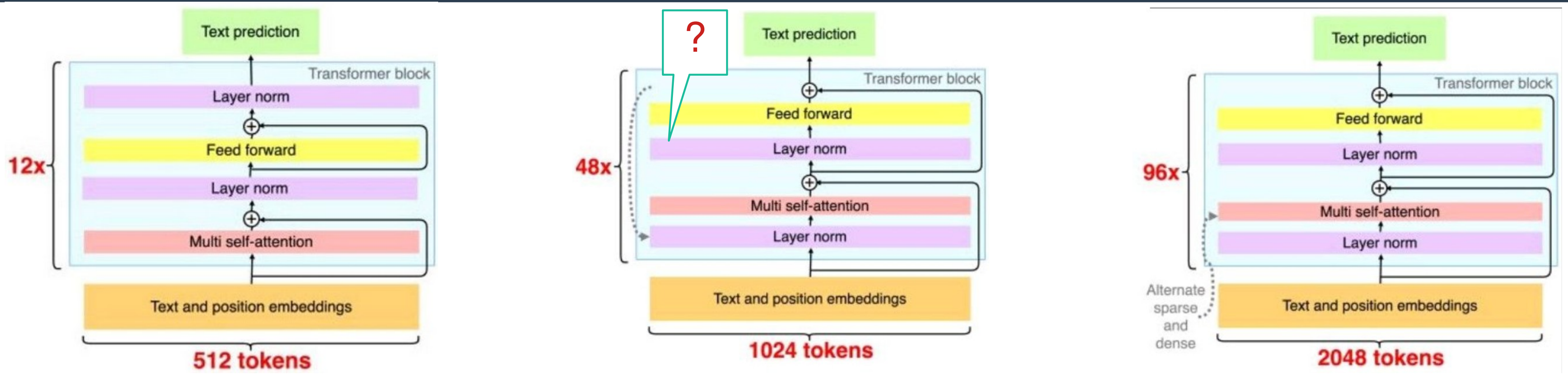
GPT-2: 117M – 1542M
GPT-3: 125M – 175B

ONE DOES NOT SIMPLY
LABEL A DATASET
imgflip.com

Motivation & Hypothesis

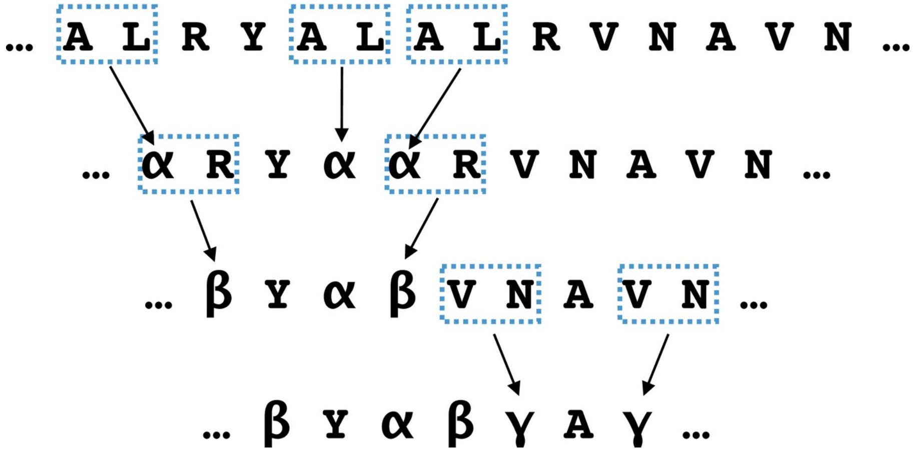
- State of the art machine learning systems are “narrow experts” (large datasets, high-capacity models, supervised learning)
- Problems of labeled datasets:
    - Huge effort to create
    - Models might exploit spurious correlations, data contamination
    - Models are sensitive to slight changes in data distribution and task specification
- Goal: general system (many tasks, no specialized datasets)
- GPT-2: multitask learning (meta-learning)
- GPT-3: larger models make meta-learning feasible

Changes in architecture



- GPT-2: layer normalization before each and after the final block
- GPT-2: scale their initial weights by 1/√#layers
- GPT-3: “alternating dense and locally bended sparse attention patterns”

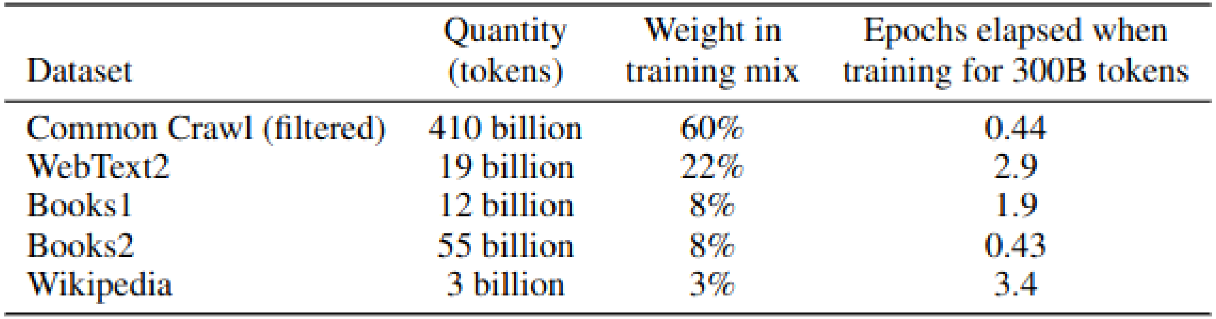
Byte-Pair Encodings



- Adds subword-level information (=> generalization)
- Modification: do not merge across character categories (except spaces)
- compromise between compression efficiency and fragmentation of common words: dog dog. dog!

Pretraining data

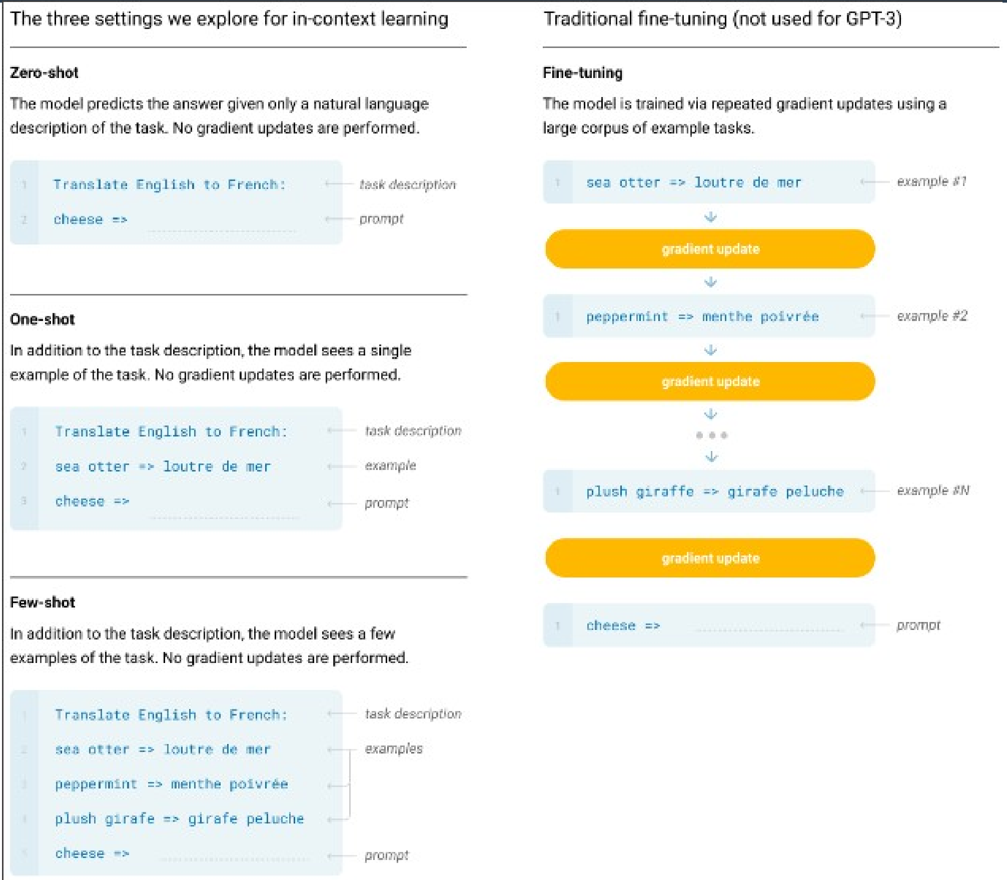
- **GPT-2: WebText = all web documents linked by Reddit posts with min. 3 karma**
  - heuristics to overcome data quality issues
  - deduplicate and remove all Wikipedia documents (could lead to overlapping)
  - 8 million documents, 40 GB
- **GPT-3: mixture**
  - Weighted by quality
  - Attempt to deduplicate



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

Supervised vs. N-shot learning

- NO finetuning!
- “learning” during inference (without gradient updates)
- GPT-2: varying context, ambiguously called zero-shot
- GPT-3: clear distinction between zero-shot, one-shot, and few-shot



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.

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

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

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

```
1 Translate English to French:      ← task description
2 sea otter => loutre de mer        ← examples
3 peppermint => menthe poivrée      ←
4 plush girafe => girafe peluche    ←
5 cheese => ..........               ← prompt
```

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

Fine-tuning

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

```
1 sea otter => loutre de mer        ← example #1
        ↓
  gradient update
        ↓
1 peppermint => menthe poivrée      ← example #2
        ↓
  gradient update
        ↓
       • • •
        ↓
1 plush giraffe => girafe peluche   ← example #N

  gradient update

1 cheese => ..........               ← prompt
```

Some example contexts: poems

The City

BY C. P. CAVAFY

TRANSLATED BY EDMUND KEELEY

[Poem text omitted]

SOME TREES

John Ashbery

[Poem text omitted]

Shadows on the Way

Wallace Stevens

Some example contexts: Natural Language Inference

anli 2: anli 2: The Gold Coast Hotel & Casino is a hotel and casino located in Paradise, Nevada. This locals' casino is owned and operated by Boyd Gaming. The Gold Coast is located one mile ($\sim$ 1.6km) west of the Las Vegas Strip on West Flamingo Road. It is located across the street from the Palms Casino Resort and the Rio All Suite Hotel and Casino. Question: The Gold Coast is a budget-friendly casino. True, False, or Neither?

Some example contexts: Reading comprehension (et. al)

Helsinki is the capital and largest city of Finland. It is in the region of Uusimaa, in southern Finland, on the shore of the Gulf of Finland. Helsinki has a population of , an urban population of , and a metropolitan population of over 1.4 million, making it the most populous municipality and urban area in Finland. Helsinki is some north of Tallinn, Estonia, east of Stockholm, Sweden, and west of Saint Petersburg, Russia. Helsinki has close historical connections with these three cities.
[...]
Q: what is the most populous municipality in Finland?
A: Helsinki
Q: how many people live there?
A: 1.4 million in the metropolitan area
Q: what percent of the foreign companies that operate in Finland are in Helsinki?
A: 75%
Q: what towns are a part of the metropolitan area?
A:

Some example contexts: Word scrambling

Please unscramble the letters into a word, and write that word:

asinoc =

Some example contexts: Language modeling/completion

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 . ->

Some example contexts: translation

- One- and few-shot:

Keinesfalls dürfen diese für den kommerziellen Gebrauch verwendet werden.

=

- Zero-shot:

Q: What is the {language} translation of {sentence} A:

Some example contexts: arithmetics

Q: What is (2 * 4) * 6?

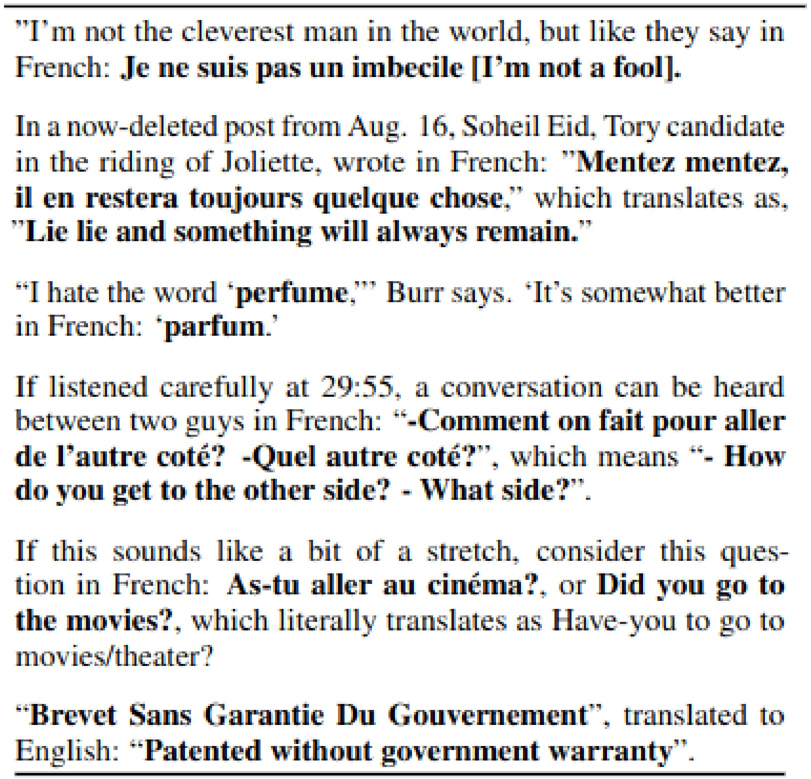
A:

Q: What is 17 minus 14?

A:

Is it Black Magic?

- (task, input, output) is sequence completion
- supervised objective = unsupervised objective on a subsequence
- sufficiently large LMs learn the tasks when seen in training for better prediction
- Inference:
  context = reminder



"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?**".

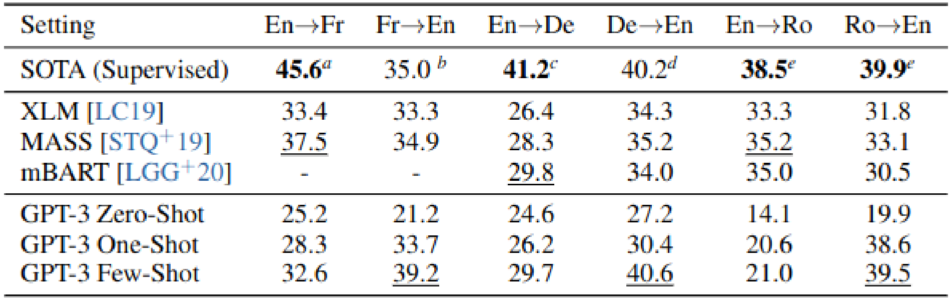
If this sounds like a bit of a stretch, consider this question in French: **As-tu aller au cinéma?**, or **Did you go to the movies?**, which literally translates as Have-you to go to movies/theater?

"**Brevet Sans Garantie Du Gouvernement**", translated to English: "**Patented without government warranty**".

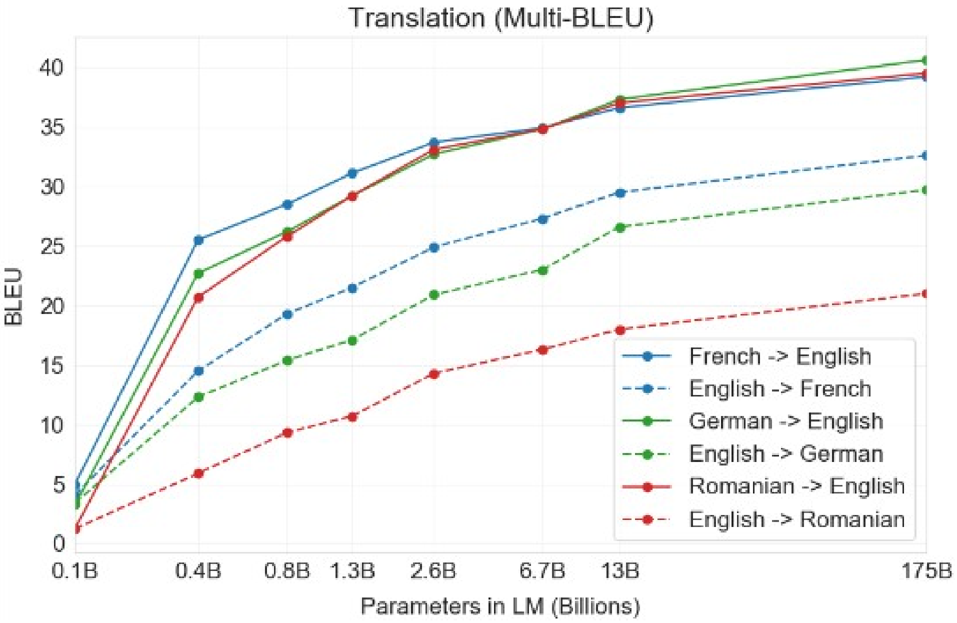
Upscaling language models

Model	#parameters	#layers	dimension	Context size	Batch size
GPT-2 (GPT-1)	117M	12	768	512 → 1024	512
GPT-2	345M	24	1024	1024	512
GPT-2	762M	36	1280	1024	512
GPT-2 (GPT-2)	1542M	48	1600	1024	512
GPT-3 small	125M	12	768	2048	0.5M
GPT-3 medium	350M	24	1024	2048	0.5M
GPT-3 large	760M	24	1536	2048	0.5M
GPT-3 XL	1.3B	24	2046	2048	1M
GPT-3 2.7B	2.7B	32	2560	2048	1M
GPT-3 6.7B	6.7B	32	4096	2048	2M
GPT-3 13B	13B	40	5140	2048	2M
GPT-3 (GPT-3)	175B	90	12288	2048	3.2M

Results: Translation

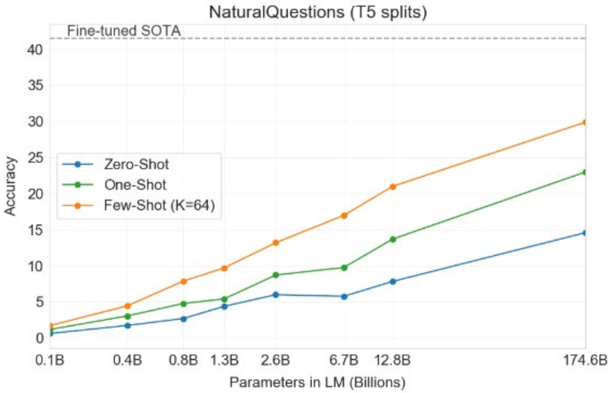


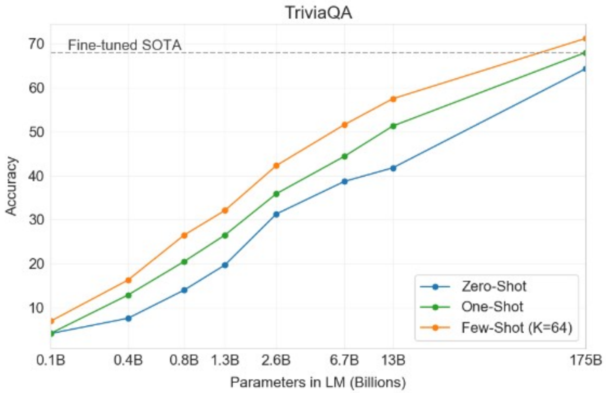
Setting	En→Fr	Fr→En	En→De	De→En	En→Ro	Ro→En
SOTA (Supervised)	**45.6**[a]	35.0 [b]	**41.2**[c]	40.2[d]	**38.5**[e]	**39.9**[e]
XLM [LC19]	33.4	33.3	26.4	34.3	33.3	31.8
MASS [STQ$^+$19]	37.5	34.9	28.3	35.2	35.2	33.1
mBART [LGG$^+$20]	-	-	29.8	34.0	35.0	30.5
GPT-3 Zero-Shot	25.2	21.2	24.6	27.2	14.1	19.9
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

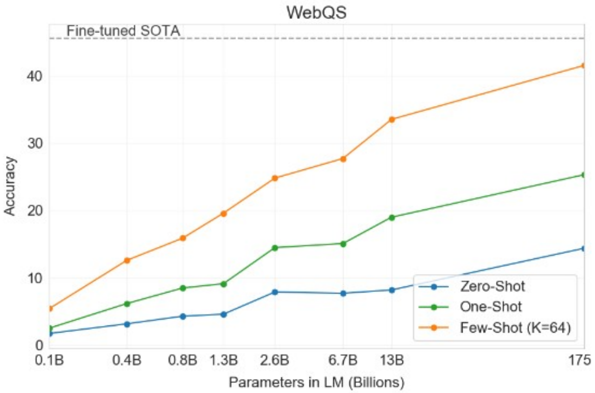


Results: Question Answering

- GPT-2 large: 4.1%
- GPT-2 small: 1%
- Learns answer style
- GPT-3:

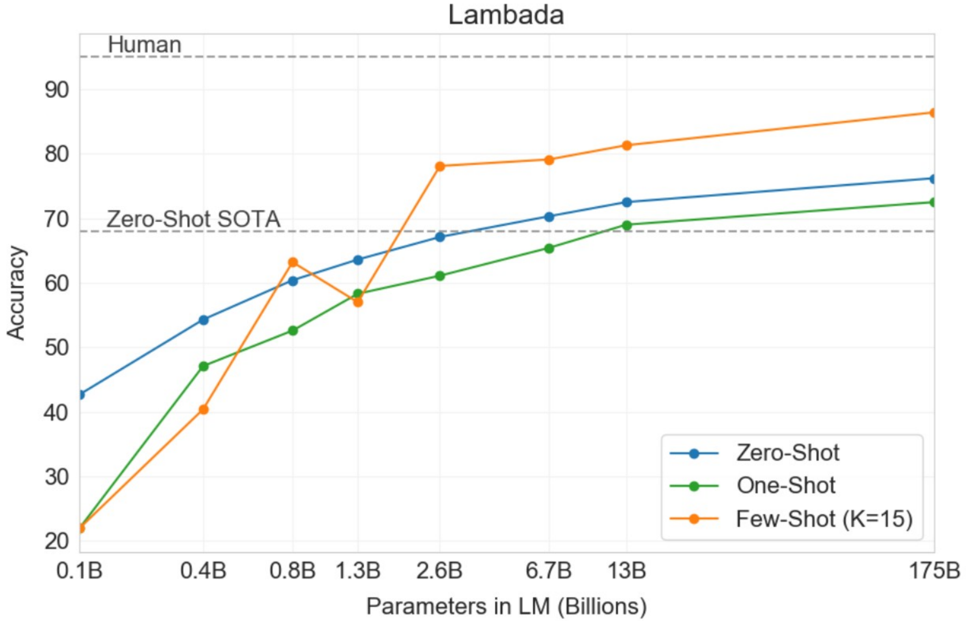

NaturalQuestions (T5 splits)


TriviaQA

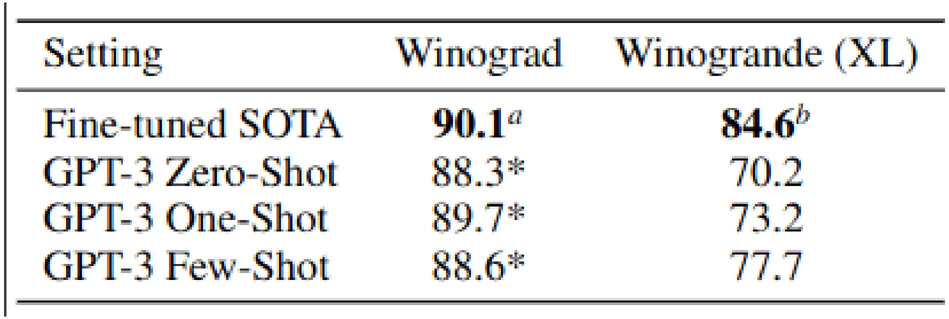

WebQS

Results: LAMBADA

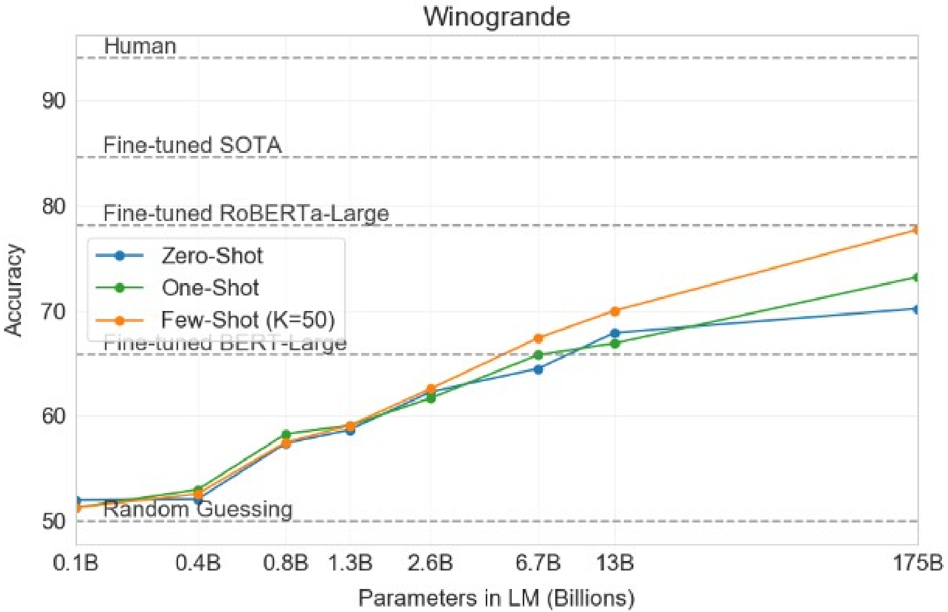
- GPT-2: filter for stop words
- GPT-3: fill in gaps (only works in few-shot case)



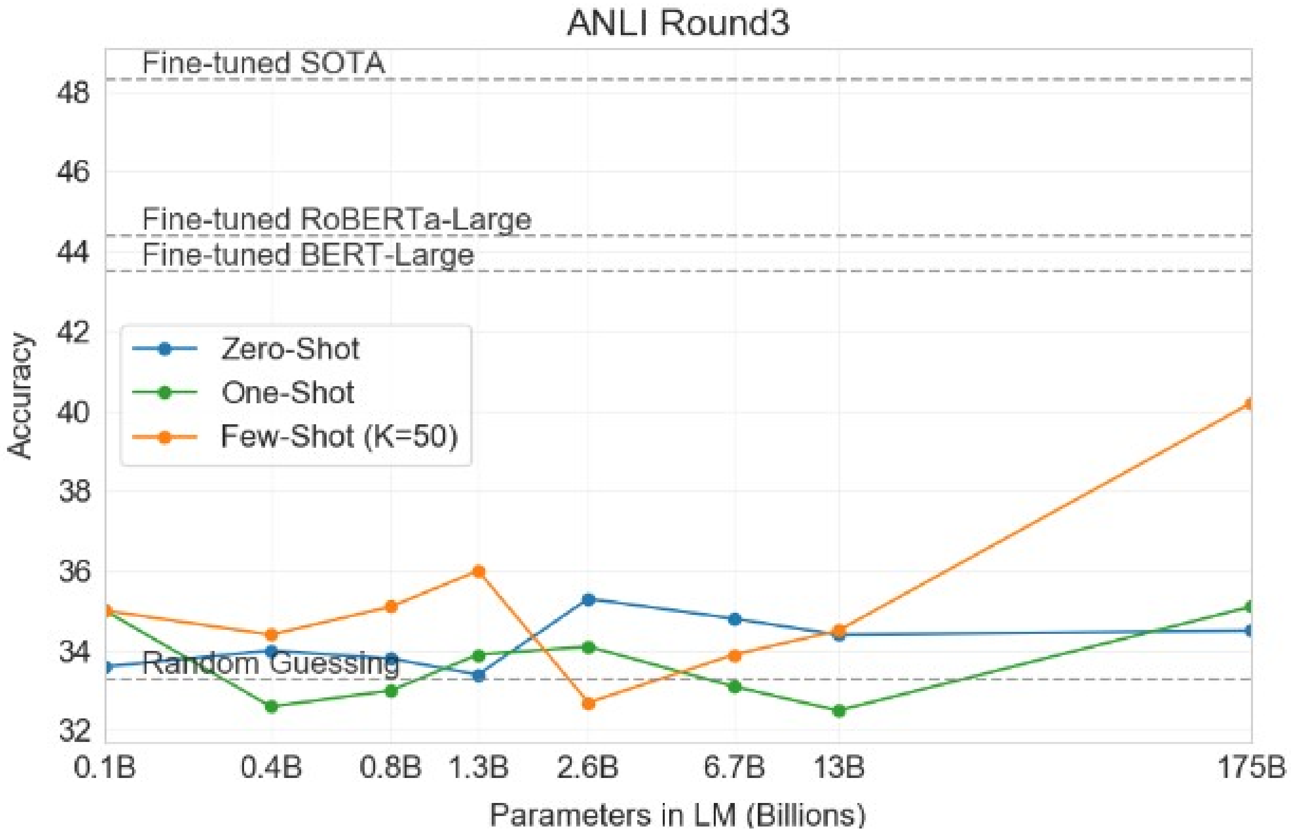
Results: Winograd



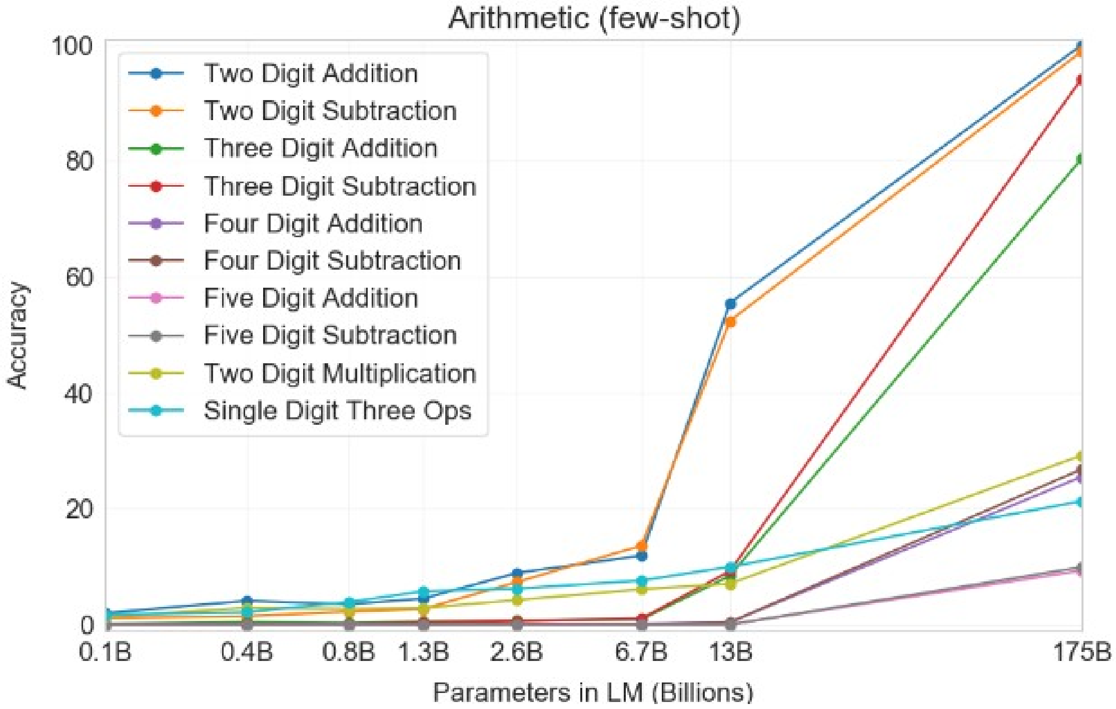
Setting	Winograd	Winogrande (XL)
Fine-tuned SOTA	**90.1**[a]	**84.6**[b]
GPT-3 Zero-Shot	88.3*	70.2
GPT-3 One-Shot	89.7*	73.2
GPT-3 Few-Shot	88.6*	77.7



Results: Natural language inference

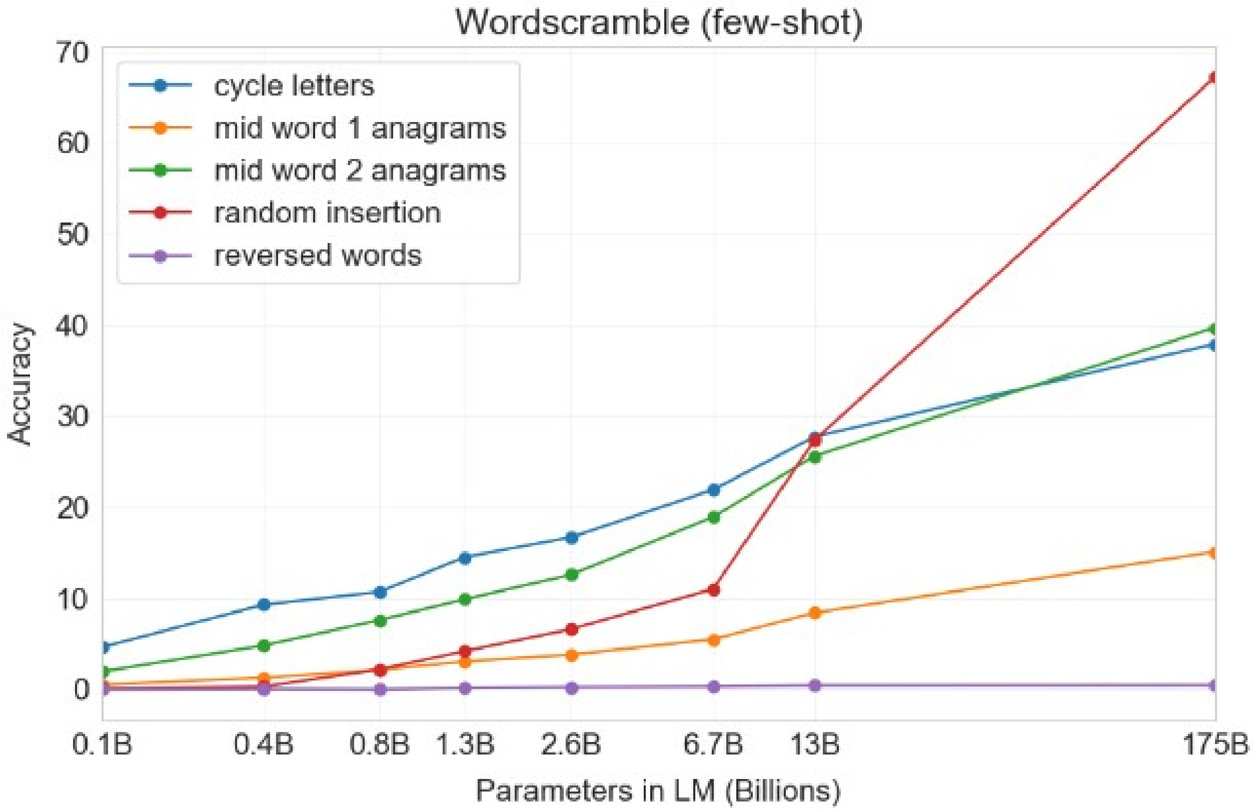


Results: Arithmetic

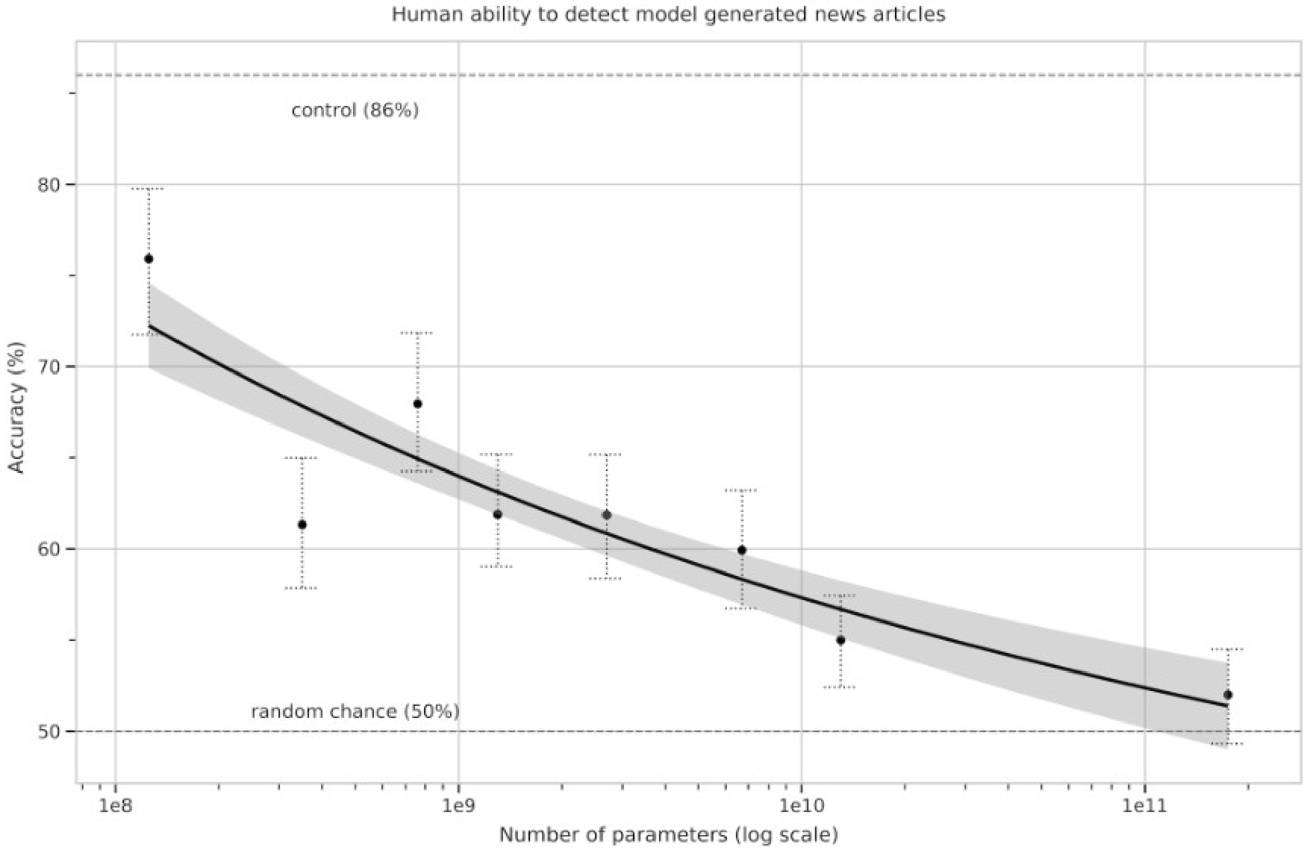


- No memorization, “typical” errors (e.g. carry)

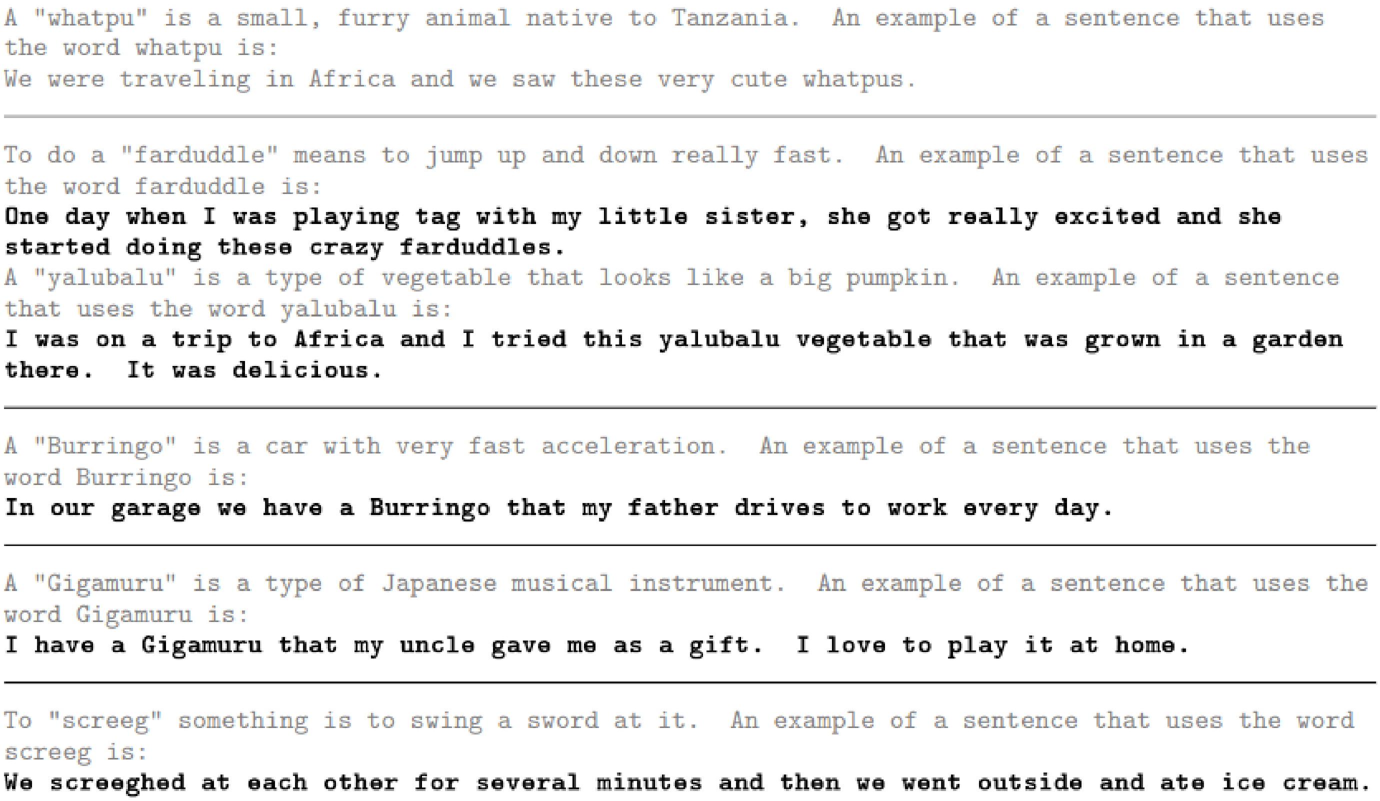
Results: Word Scrambling



Results: News article generation



Results: using novel words



A "whatpu" is a small, furry animal native to Tanzania. An example of a sentence that uses the word whatpu is:
We were traveling in Africa and we saw these very cute whatpus.

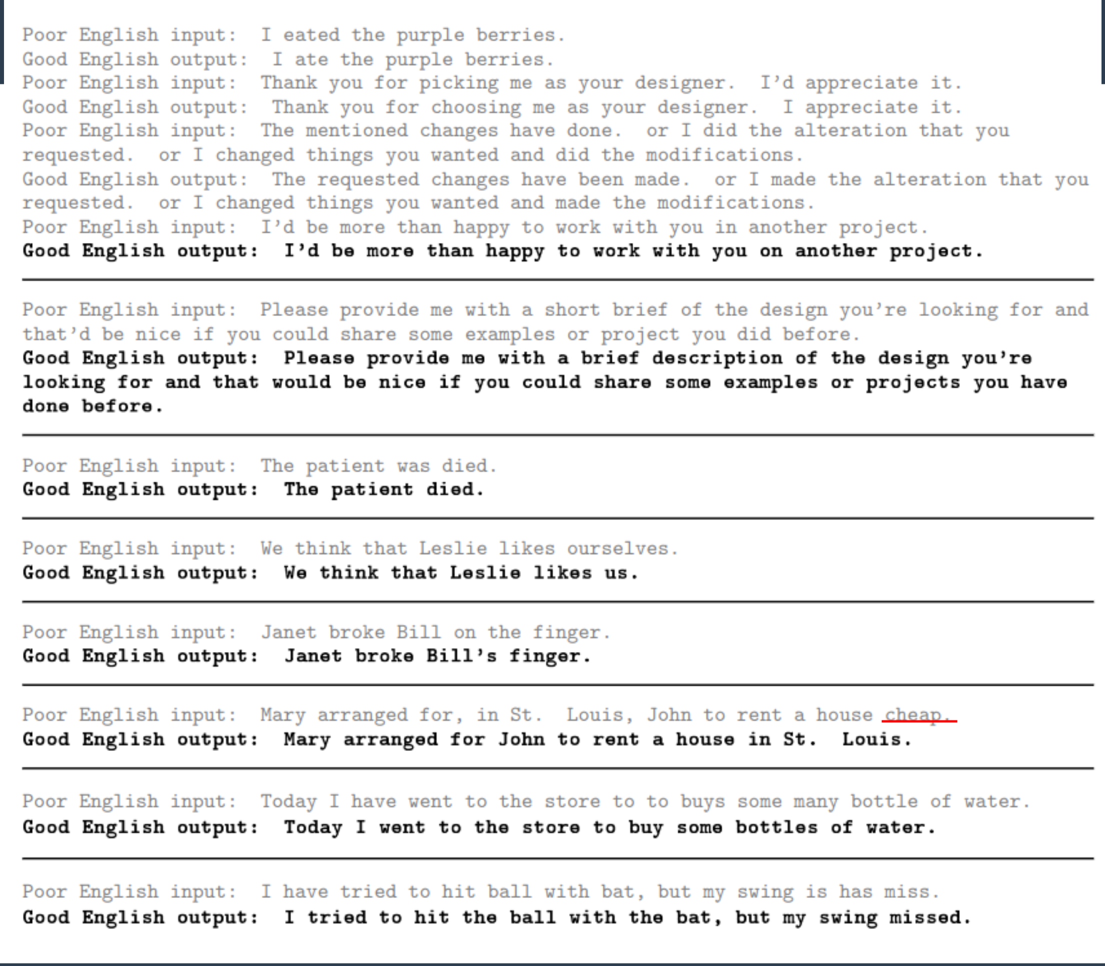
To do a "farduddle" means to jump up and down really fast. An example of a sentence that uses the word farduddle is:
One day when I was playing tag with my little sister, she got really excited and she started doing these crazy farduddles.
A "yalubalu" is a type of vegetable that looks like a big pumpkin. An example of a sentence that uses the word yalubalu is:
I was on a trip to Africa and I tried this yalubalu vegetable that was grown in a garden there. It was delicious.

A "Burringo" is a car with very fast acceleration. An example of a sentence that uses the word Burringo is:
In our garage we have a Burringo that my father drives to work every day.

A "Gigamuru" is a type of Japanese musical instrument. An example of a sentence that uses the word Gigamuru is:
I have a Gigamuru that my uncle gave me as a gift. I love to play it at home.

To "screeg" something is to swing a sword at it. An example of a sentence that uses the word screeg is:
We screeghed at each other for several minutes and then we went outside and ate ice cream.

Results: correcting English 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.

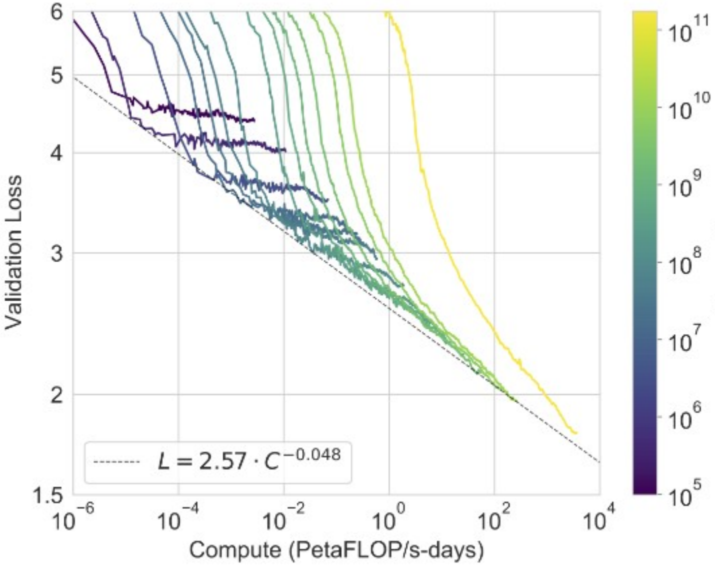
Poor English input: Janet broke Bill on the finger.
Good English output: Janet broke Bill's finger.

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.

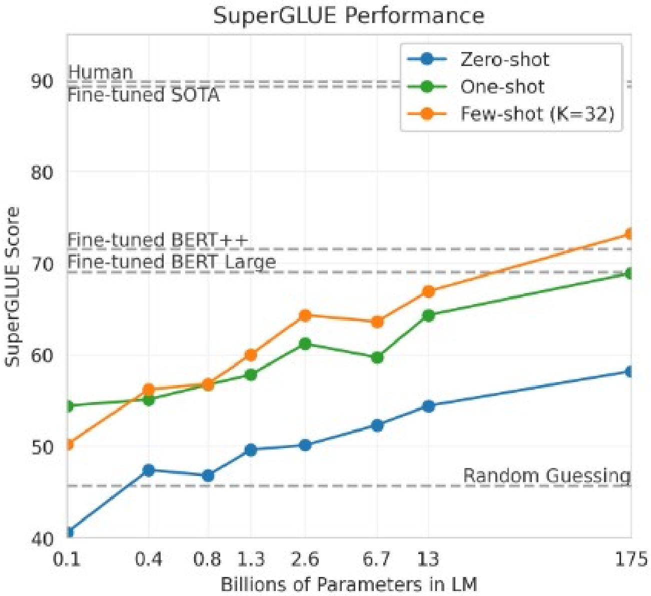
Poor English input: Today I have went to the store to to buys some many bottle of water.
Good English output: Today I went to the store to buy some bottles of water.

Poor English input: I have tried to hit ball with bat, but my swing is has miss.
Good English output: I tried to hit the ball with the bat, but my swing missed.

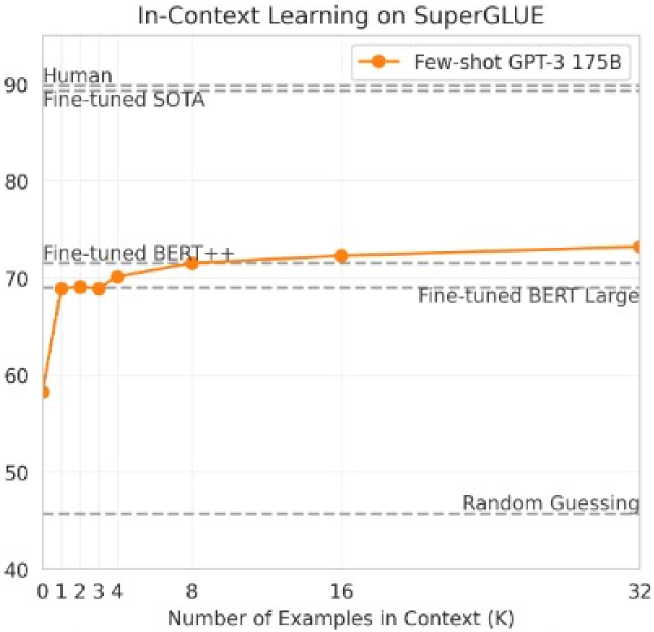
Learnings



Performance scales with compute



Performance increases with size



Performance (often) increases with context

Limitations

- text generation: semantic repetition, losing coherence, self-contradiction
- common-sense thinking (in particular, physics)
- reasoning about two sentences when comparison is involved
- lack of bidirectionality (could explain problems with two sentences)
- limits of pretraining objective (every token is weighted equally)
- lack of domain-knowledge
- low pre-training sample efficiency (sees more tetxt than humans in their lifetime)
- expensive inference due to size

Outlook

- Data contamination and memorization
- Ethical impacts