GPT-3:
Few-Shot Learning with a Giant Language Model

Melanie Subbiah

OpenAI
Columbia University

Language Models are Few-Shot Learners

Tom B. Brown[*] Benjamin Mann[*] Nick Ryder[*] Melanie Subbiah[*]

Jared Kaplan[†] Prafulla Dhariwal Arvind Neelakantan Pranav Shyam Girish Sastry

Amanda Askell Sandhini Agarwal Ariel Herbert-Voss Gretchen Krueger Tom Henighan

Rewon Child Aditya Ramesh Daniel M. Ziegler Jeffrey Wu Clemens Winter

Christopher Hesse Mark Chen Eric Sigler Mateusz Litwin Scott Gray

Benjamin Chess Jack Clark Christopher Berner

Sam McCandlish Alec Radford Ilya Sutskever Dario Amodei

OpenAI

What is the goal?

Humans learn new tasks through demonstrations and instructions.

What is the goal?

Humans learn new tasks through demonstrations and instructions.

We’d like general-purpose agents that can do the same.

Typical Approach

Disadvantages to Fine-tuning

- Creates a task-specific model

Disadvantages to Fine-tuning

- Creates a task-specific model
- Requires large high-quality supervised datasets

Disadvantages to Fine-tuning

- Creates a task-specific model
- Requires large high-quality supervised datasets
- more likely to exploit spurious correlations

Yogatama, et al. Learning and Evaluating General Linguistic Intelligence. 2019

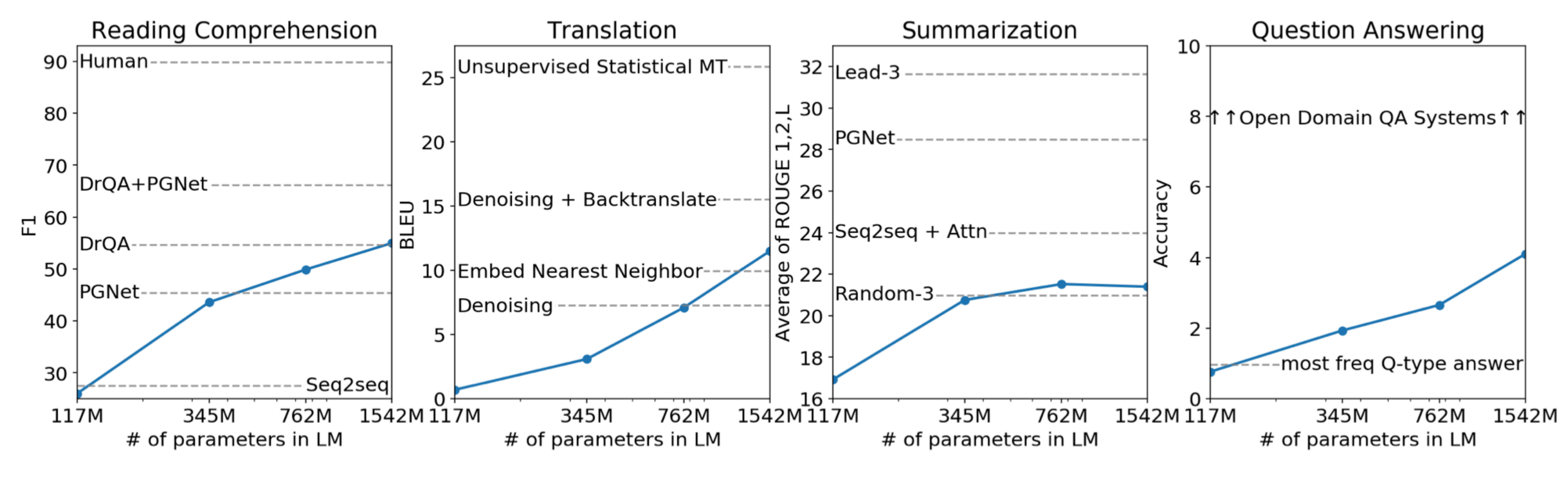
What is an alternative?

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

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

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

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



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

Can we further improve on this level of generation and generalization?

Can we further improve on this level of generation and generalization?

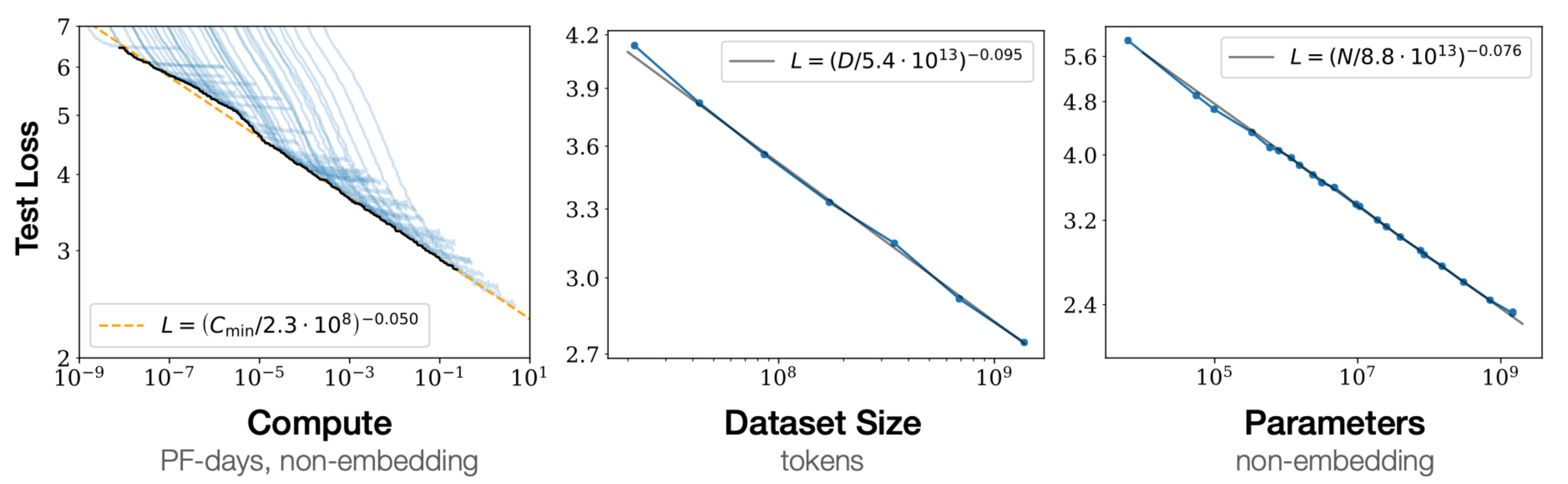
GPT-3   175 Billion parameters

Critical Aspects of GPT-3

- Model Size

- Training Objective

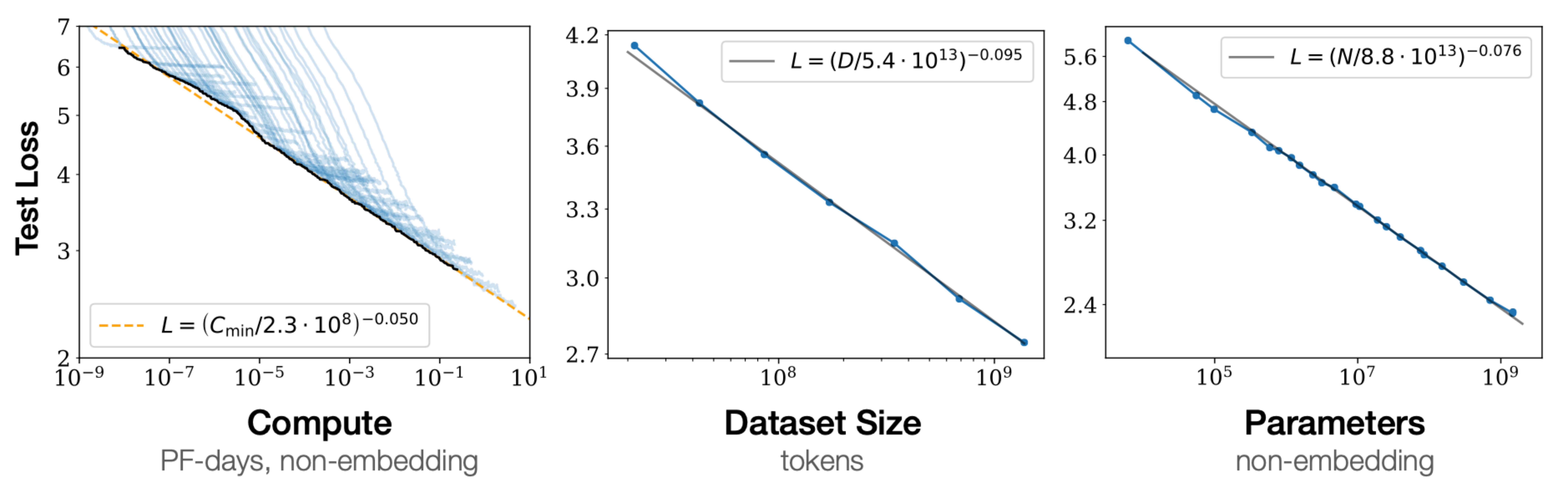
Model Size



Kaplan, et al. Scaling Laws for Neural Language Models. 2020

Model Size

Transformers scale well!



Kaplan, et al. Scaling Laws for Neural Language Models. 2020

Motivating the Training Objective

Predict the next word in a sequence.

Motivating the Training Objective

P("The cat sat on the mat.") = ???

Motivating the Training Objective

P("The cat sat on the mat.") = ???

"But it must be recognized that the notion of 'probability of a sentence' is an entirely useless one, under any known interpretation of this term."
- Noam Chomsky, 1969

Motivating the Training Objective

P("The cat sat on the mat.") > P("The cat sats on the mat.")

Motivating the Training Objective

P("The cat sat on the mat.") > P("The cat sats on the mat.")

Grammar

Motivating the Training Objective

P("The cat sat on the mat.") > P("The cat sats on the mat.")

Grammar

P("The cat sat on the mat.") > P("The whale sat on the mat.")

Motivating the Training Objective

P("The cat sat on the mat.") > P("The cat sats on the mat.")

Grammar

P("The cat sat on the mat.") > P("The whale sat on the mat.")

World Knowledge

Motivating the Training Objective

P("The cat sat on the mat.") > P("The cat sats on the mat.")

Grammar

P("The cat sat on the mat.") > P("The whale sat on the mat.")

World Knowledge

P("4" | "2 + 2 =") > P("5" | "2 + 2 =")

Motivating the Training Objective

P("The cat sat on the mat.") > P("The cat sats on the mat.")

Grammar

P("The cat sat on the mat.") > P("The whale sat on the mat.")

World Knowledge

P("4" | "2 + 2 =") > P("5" | "2 + 2 =")

Arithmetic

Motivating the Training Objective

P(“The cat sat on the mat.”) > P(“The cat sats on the mat.”)

Grammar

P(“The cat sat on the mat.”) > P(“The whale sat on the mat.”)

World Knowledge

P(“4” | “2 + 2 =”) > P(“5” | “2 + 2 =”)

Addition

P(“1 star” | “That movie was terrible. I’d give it”) > P(“5 stars” | “That movie was terrible. I’d give it”)

Motivating the Training Objective

P(“The cat sat on the mat.”) > P(“The cat sats on the mat.”)

Grammar

P(“The cat sat on the mat.”) > P(“The whale sat on the mat.”)

World Knowledge

P(“4” | “2 + 2 =”) > P(“5” | “2 + 2 =”)

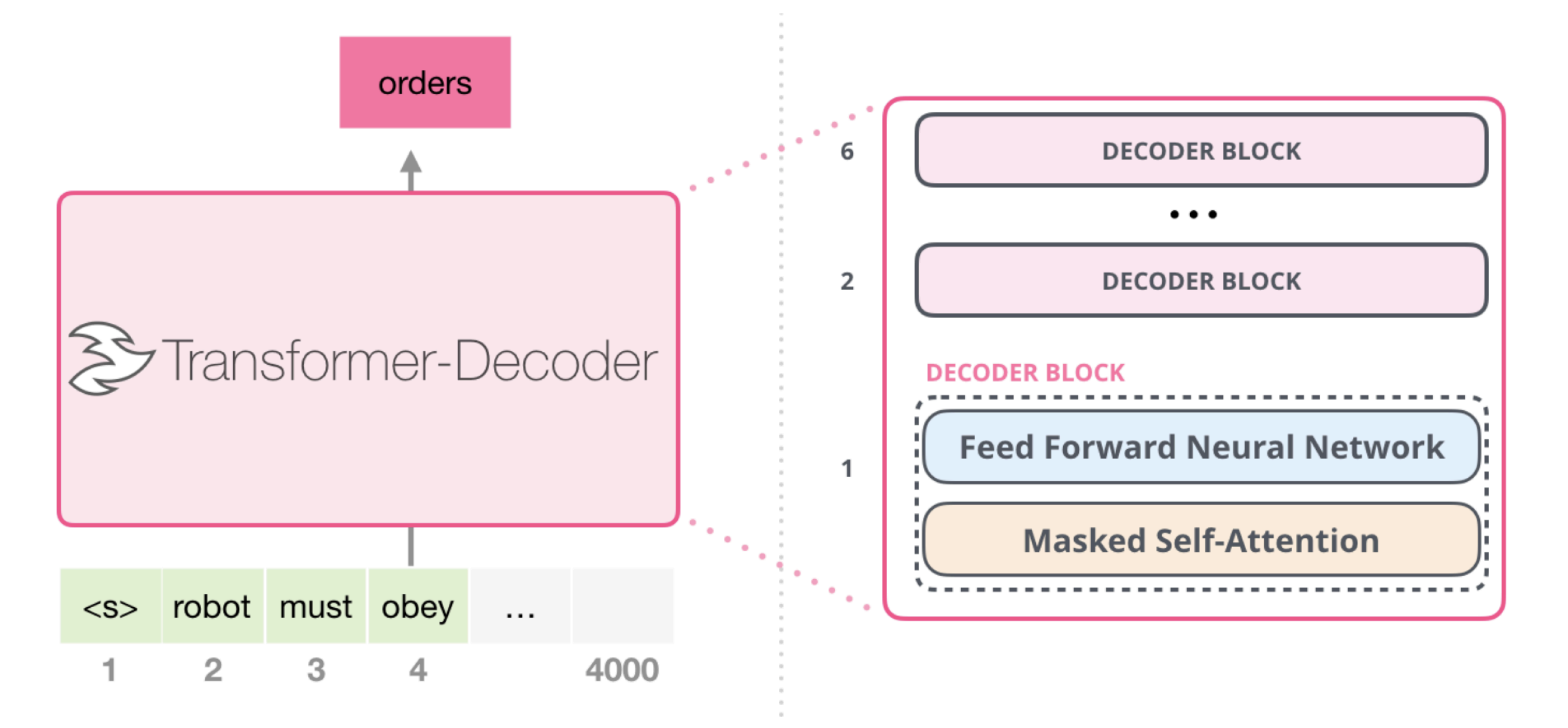
Addition

P(“1 star” | “That movie was terrible. I’d give it”) > P(“5 stars” | “That movie was terrible. I’d give it”)

Sentiment Analysis

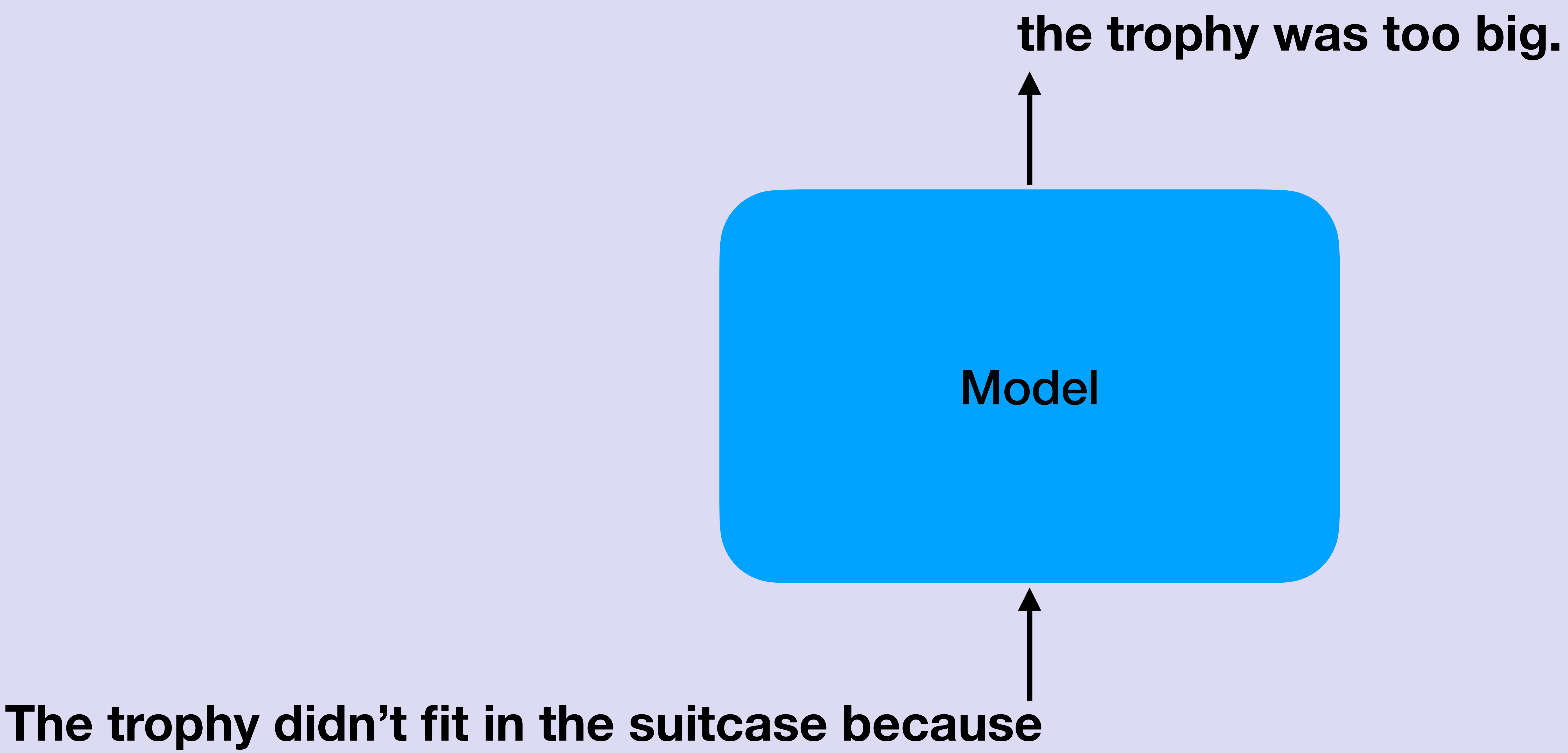
Approach

Model

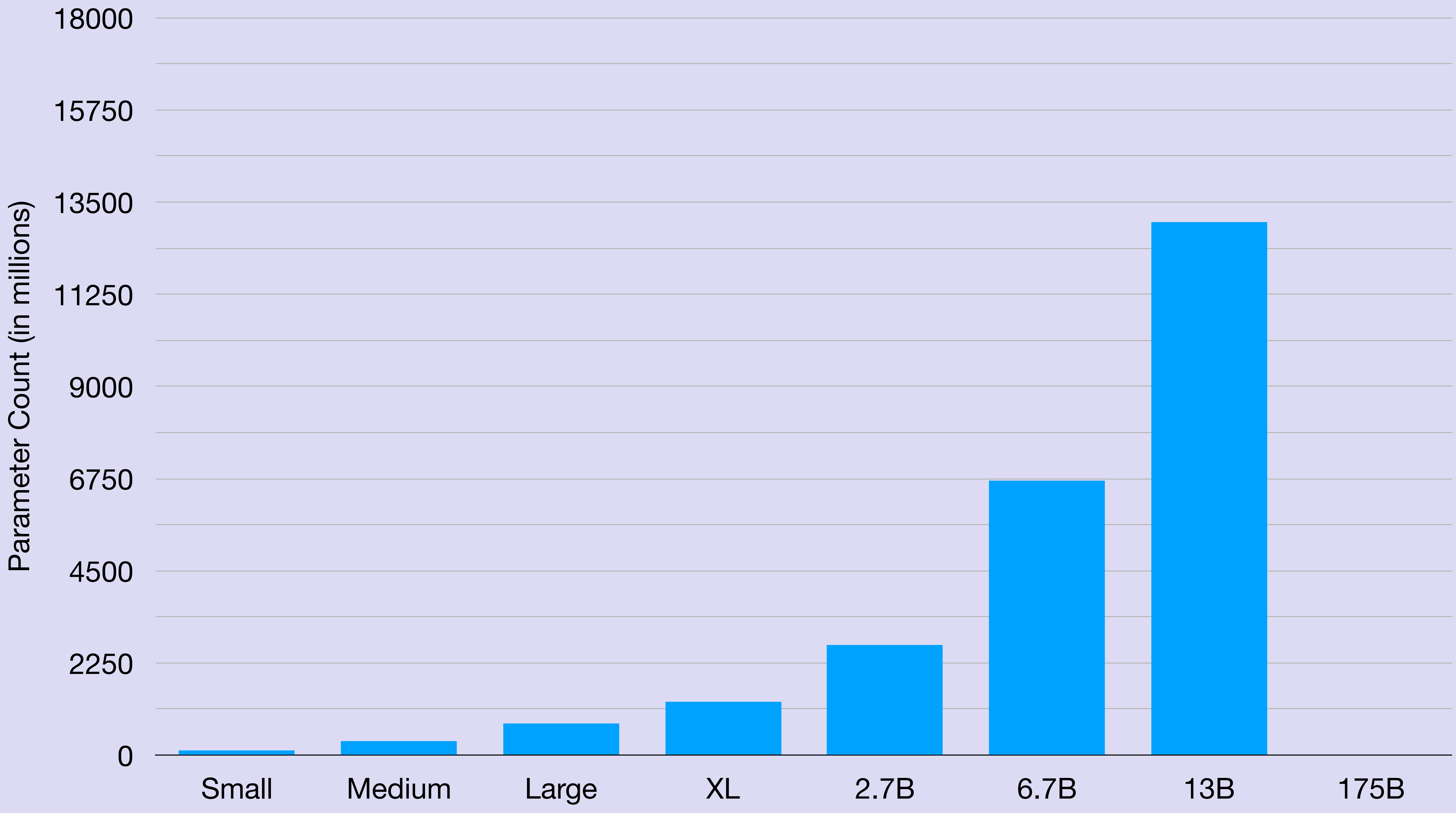


http://jalammar.github.io/illustrated-gpt2/

Model


the trophy was too big.
Model
The trophy didn't fit in the suitcase because

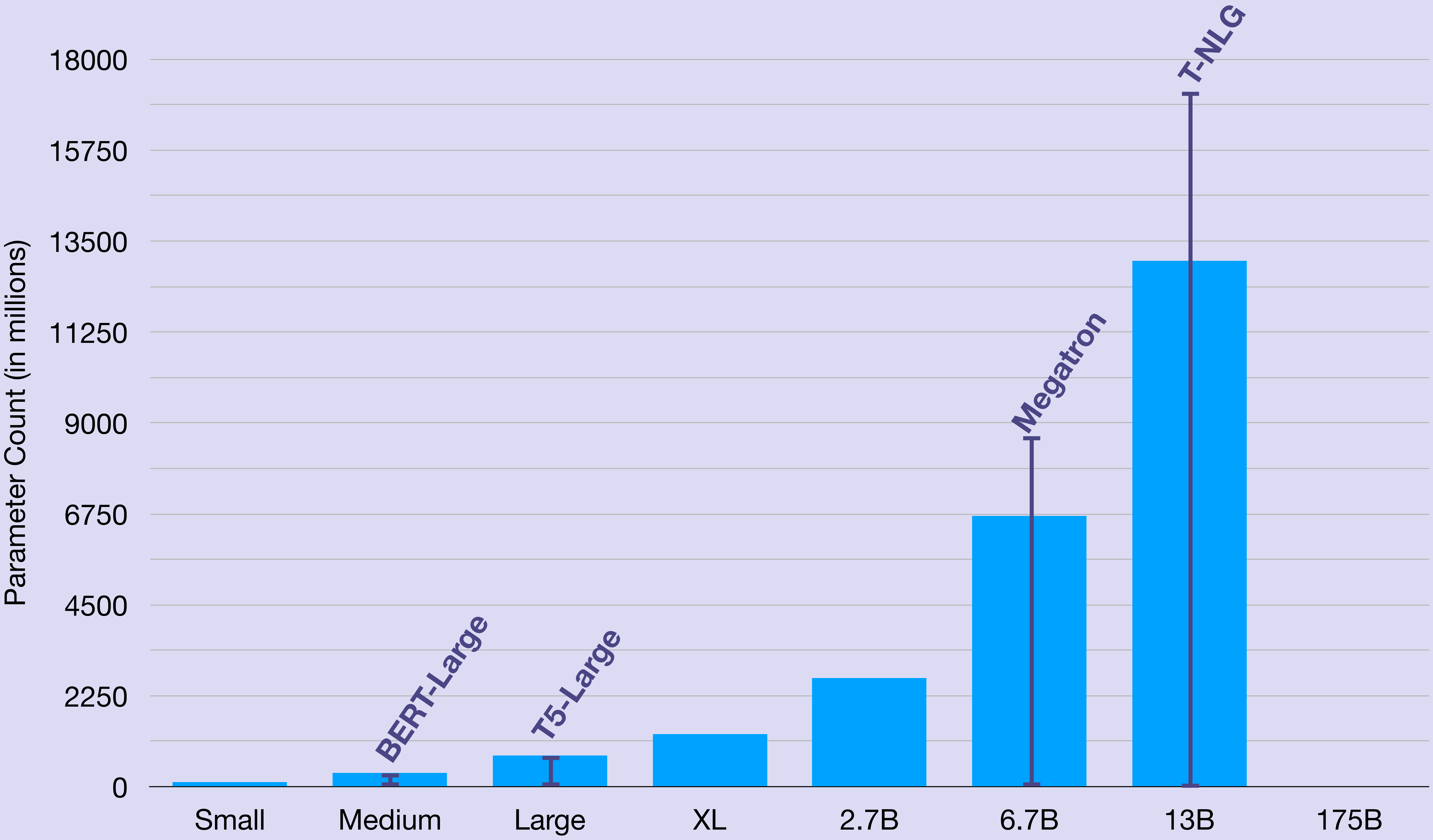
Model Sizes



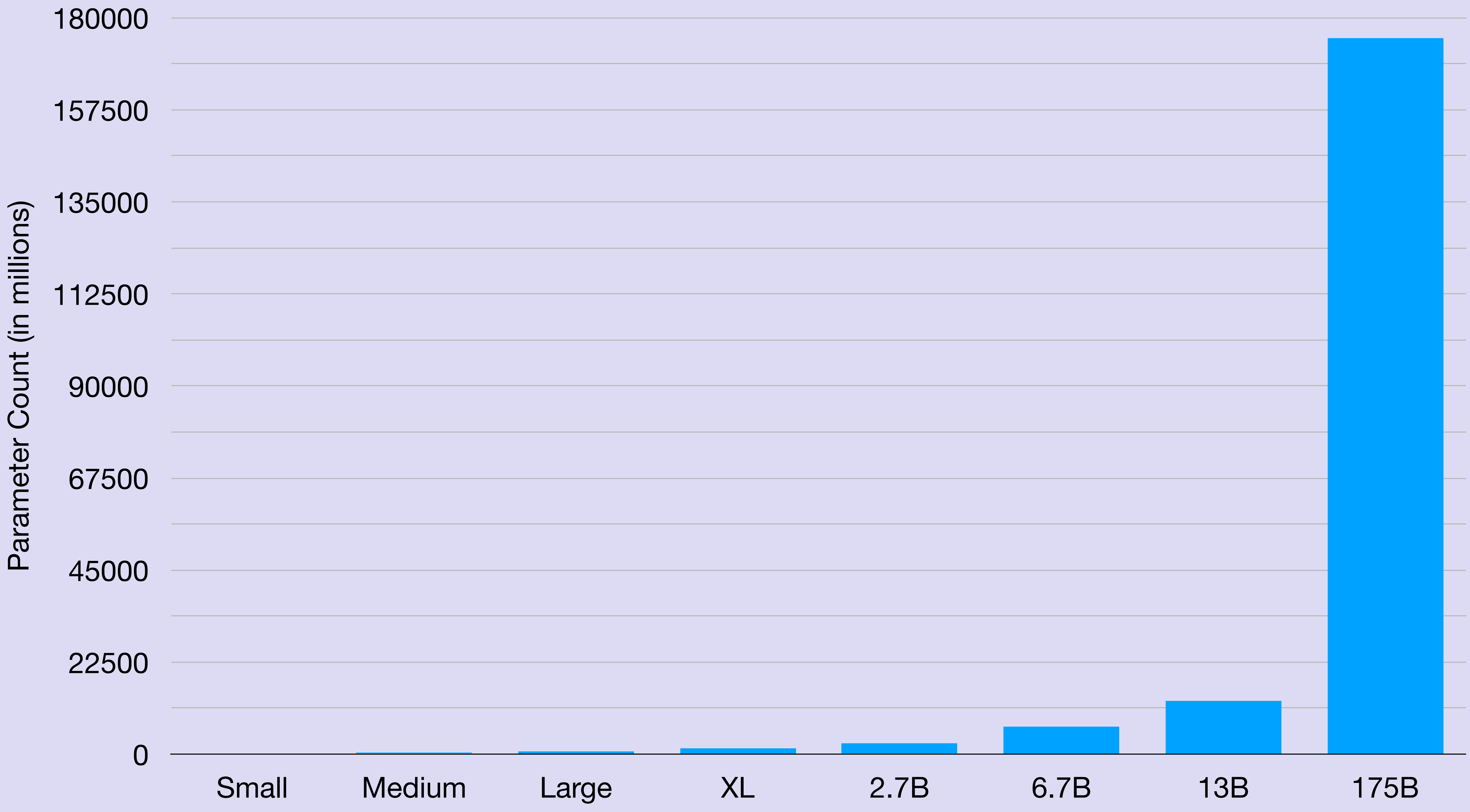
Parameter Count (in millions)
18000
15750
13500
11250
9000
6750
4500
2250
0
Small
Medium
Large
XL
2.7B
6.7B
13B
175B

Model Sizes

Devlin, et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. 2018
Raffel, et al. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. 2019
Shoeybi, et al. Megatron-LM: Training Multi-Billion Parameter Language Models Using Model Parallelism. 2019
Microsoft. Turing-NLG: A 17-Billion Parameter Language Model by Microsoft. 2020

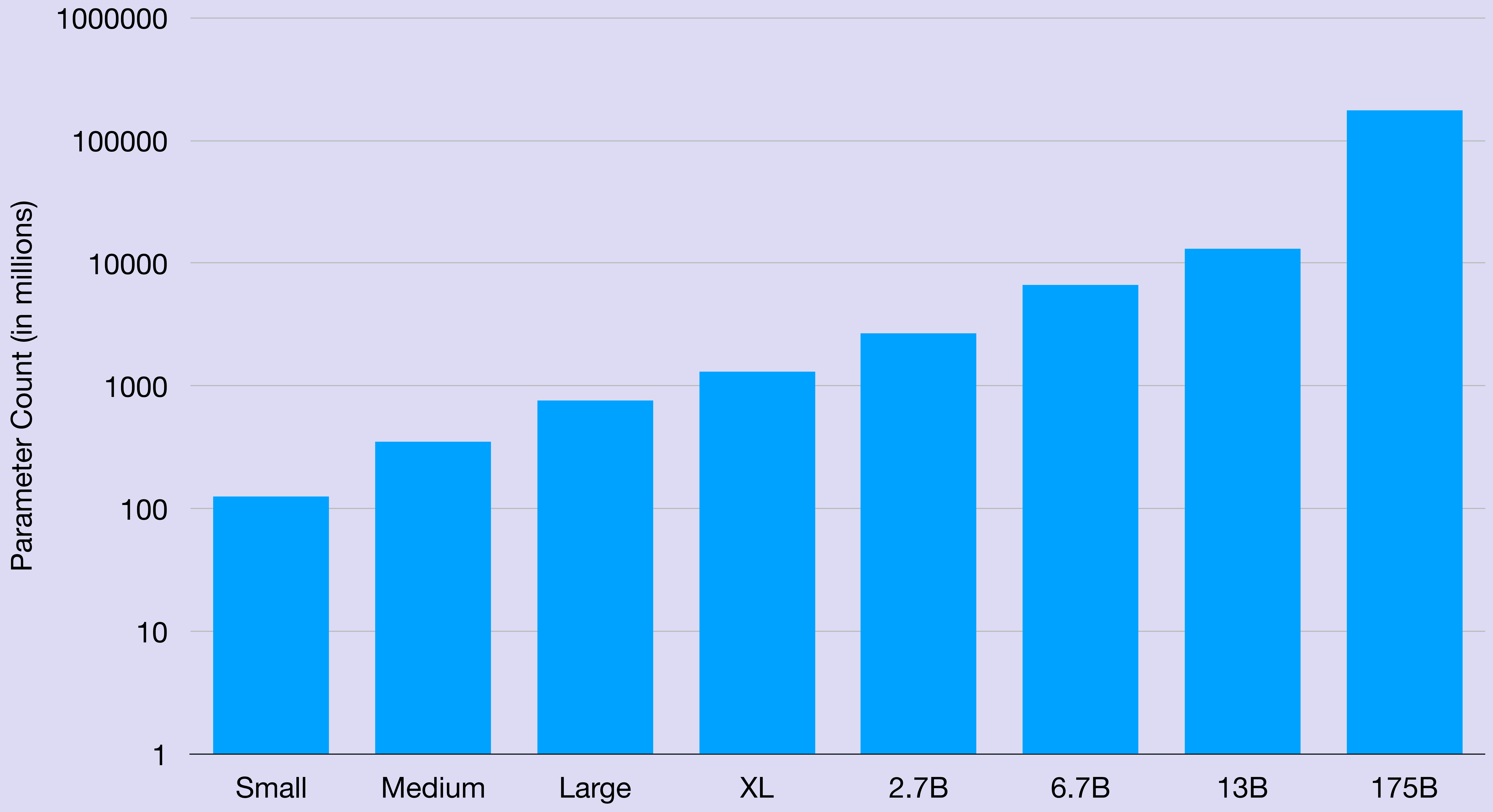


Model Sizes

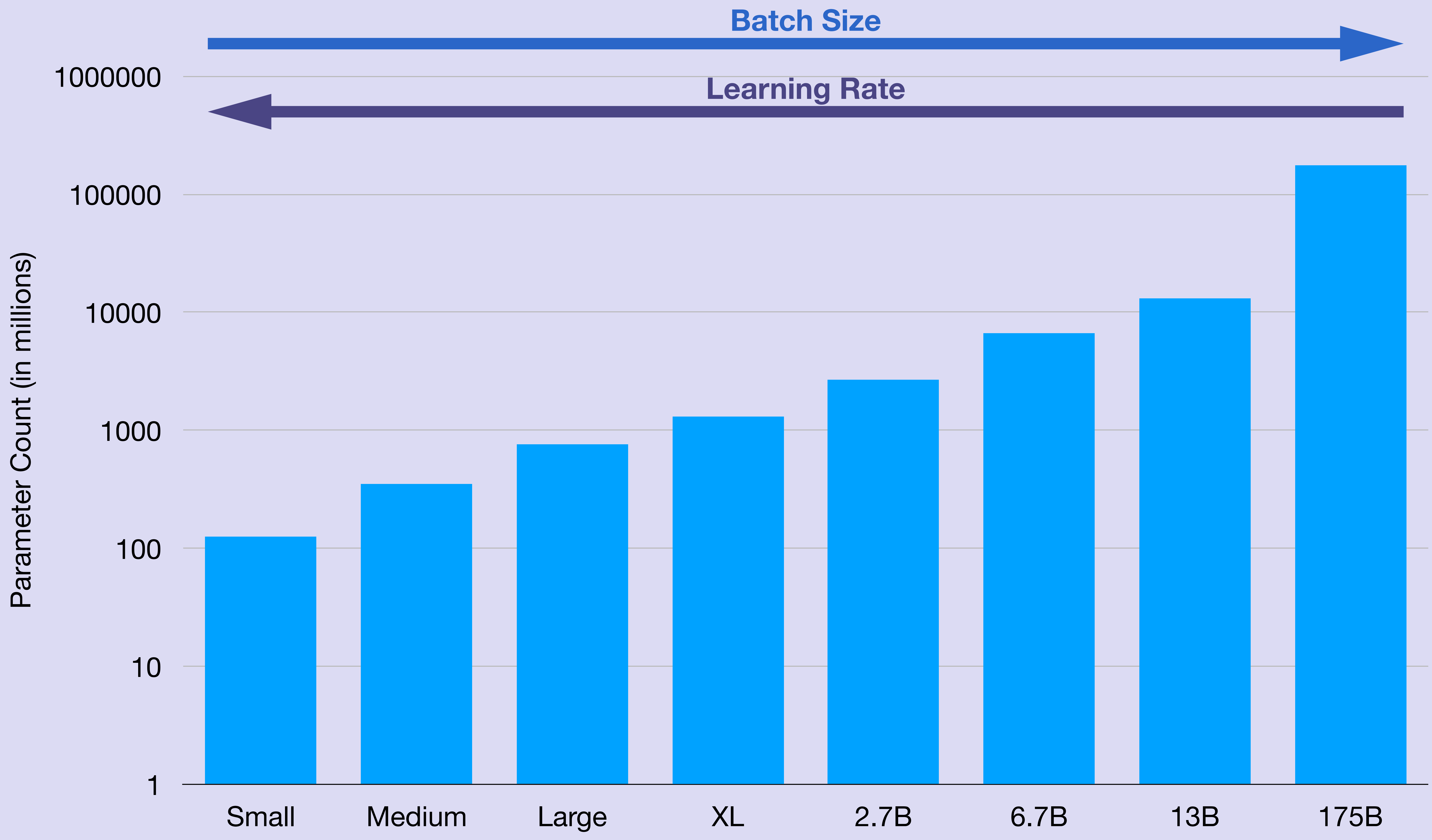


180000
157500
135000
112500
90000
67500
45000
22500
0
Parameter Count (in millions)
Small
Medium
Large
XL
2.7B
6.7B
13B
175B

Model Sizes

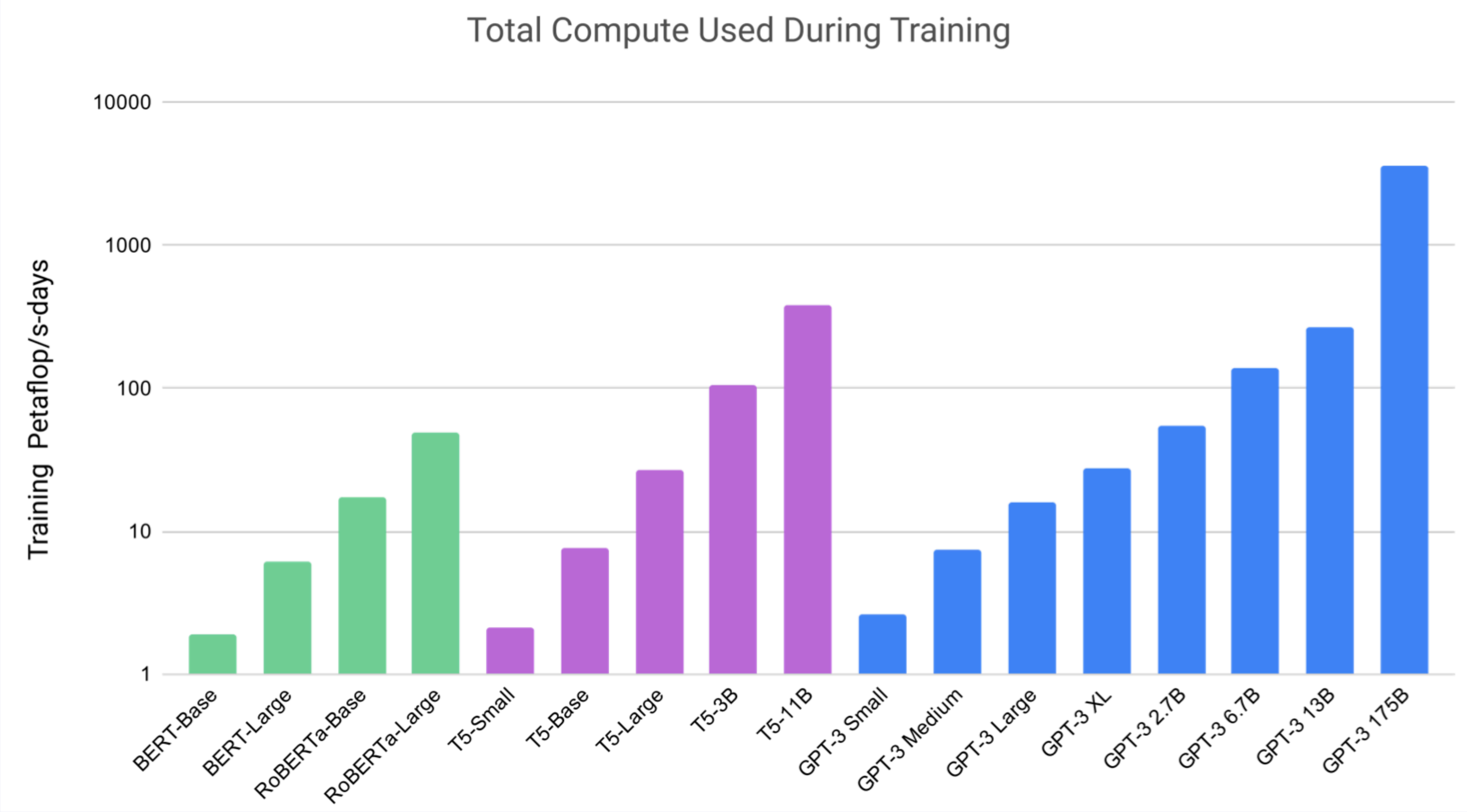

1000000
100000
10000
1000
100
10
1
Parameter Count (in millions)
Small
Medium
Large
XL
2.7B
6.7B
13B
175B

Model Sizes



Batch Size
Learning Rate
1000000
100000
10000
1000
100
10
1
Parameter Count (in millions)
Small
Medium
Large
XL
2.7B
6.7B
13B
175B

Compute



Total Compute Used During Training
Training Petaflop/s-days
10000
1000
100
10
1
BERT-Base
BERT-Large
RoBERTa-Base
RoBERTa-Large
T5-Small
T5-Base
T5-Large
T5-3B
T5-11B
GPT-3 Small
GPT-3 Medium
GPT-3 Large
GPT-3 XL
GPT-3 2.7B
GPT-3 6.7B
GPT-3 13B
GPT-3 175B

Dataset

- **Common Crawl** (filtered) - general web crawl, filtered based on similarity to high-quality reference and de-duplication

https://commoncrawl.org/the-data

Dataset

- **Common Crawl** (filtered) - general web crawl, filtered based on similarity to high-quality reference and de-duplication
- **WebText2** - expanded version of GPT-2 training data, scrape of outbound links from Reddit posts with reasonably high ratings

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

Dataset

- **Common Crawl** (filtered) - general web crawl, filtered based on similarity to high-quality reference and de-duplication

- **WebText2** - expanded version of GPT-2 training data, scrape of outbound links from Reddit posts with reasonably high ratings

- **Books1** & **Books2** - internet-based books

Dataset

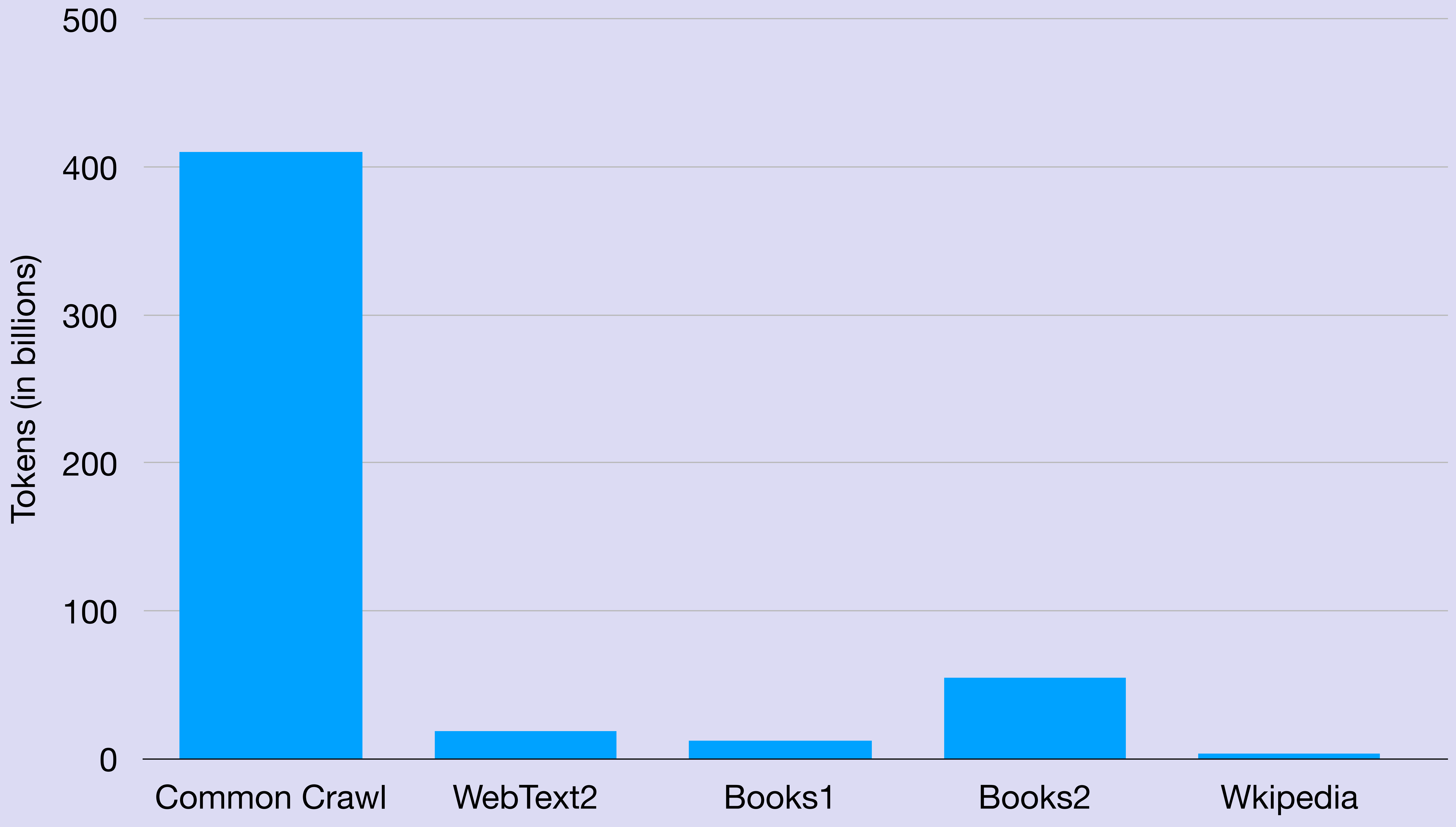
- **Common Crawl** (filtered) - general web crawl, filtered based on similarity to high-quality reference and de-duplication

- **WebText2** - expanded version of GPT-2 training data, scrape of outbound links from Reddit posts with reasonably high ratings

- **Books1** & **Books2** - internet-based books

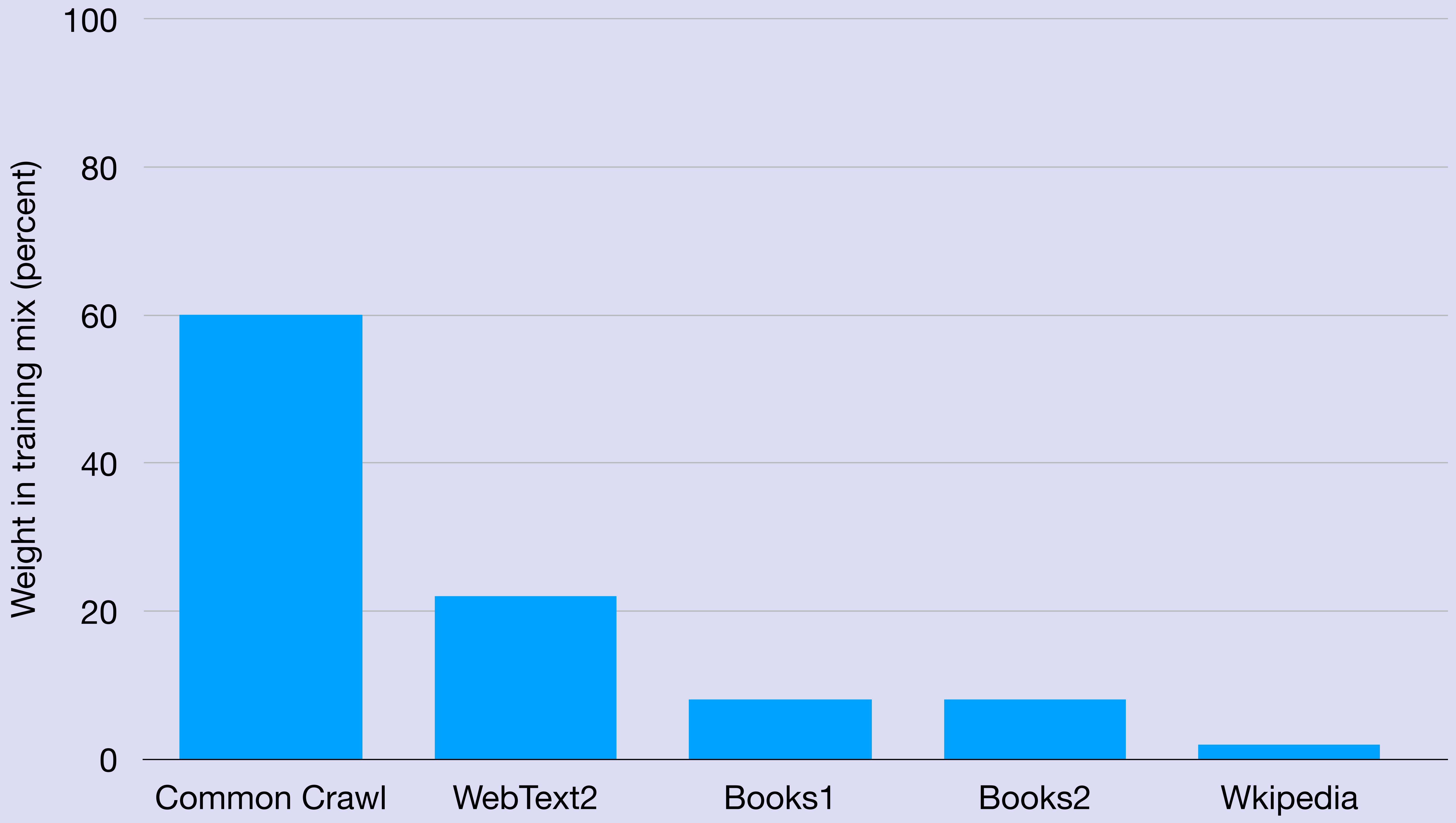
- **Wikipedia** - English-language Wikipedia

Dataset Mix

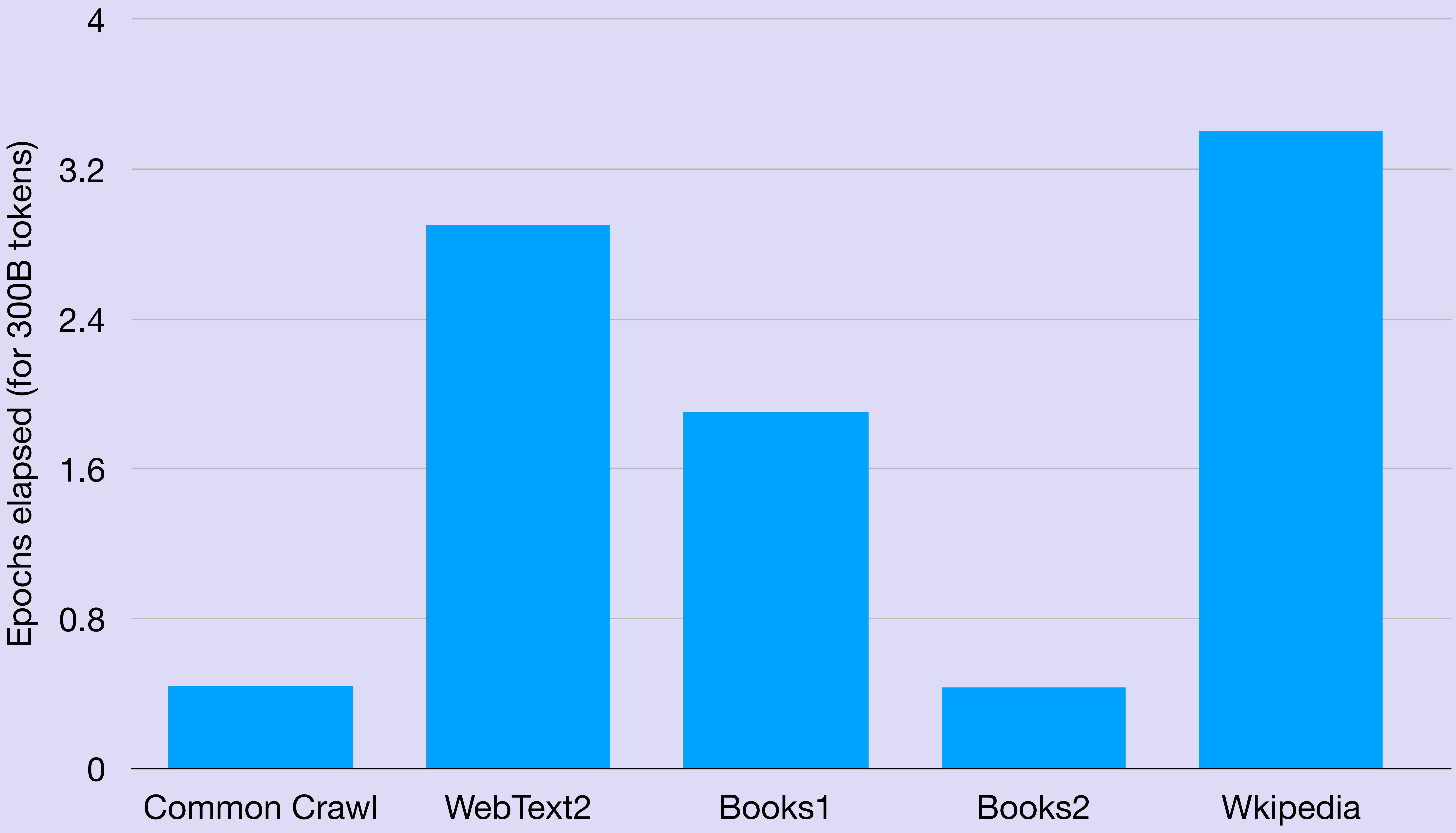


500
400
300
200
100
0
Tokens (in billions)
Common Crawl
WebText2
Books1
Books2
Wkipedia

Dataset Mix


Weight in training mix (percent)
100
80
60
40
20
0
Common Crawl
WebText2
Books1
Books2
Wkipedia

Dataset Mix



Epochs elapsed (for 300B tokens)
4
3.2
2.4
1.6
0.8
0
Common Crawl
WebText2
Books1
Books2
Wkipedia

Evaluations

Let’s try it!

tdaeef = ?

Let’s try it!

Please unscramble the letters into a word and write that word.
tdaeef = ?

Let’s try it!

Zero-Shot

Please unscramble the letters into a word and write that word.
tdaeef = ?

Let's try it!

One-Shot

Please unscramble the letters into a word and write that word.
pcirlaroc = reciprocal
tdaeef = ?

Let's try it!

Few-Shot

Please unscramble the letters into a word and write that word.
pcirlaroc = reciprocal
elapac = palace
tdaeef = ?

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
```

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
```

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:
cheese =>
```

task description

prompt

One-shot

In addition to the task description, the model sees a single example of

```
Tran
sea
chee
```

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:
sea otter => loutre de mer
peppermint => menthe poivrée
plush girafe => girafe peluche
cheese =>
```

task description

examples

prompt

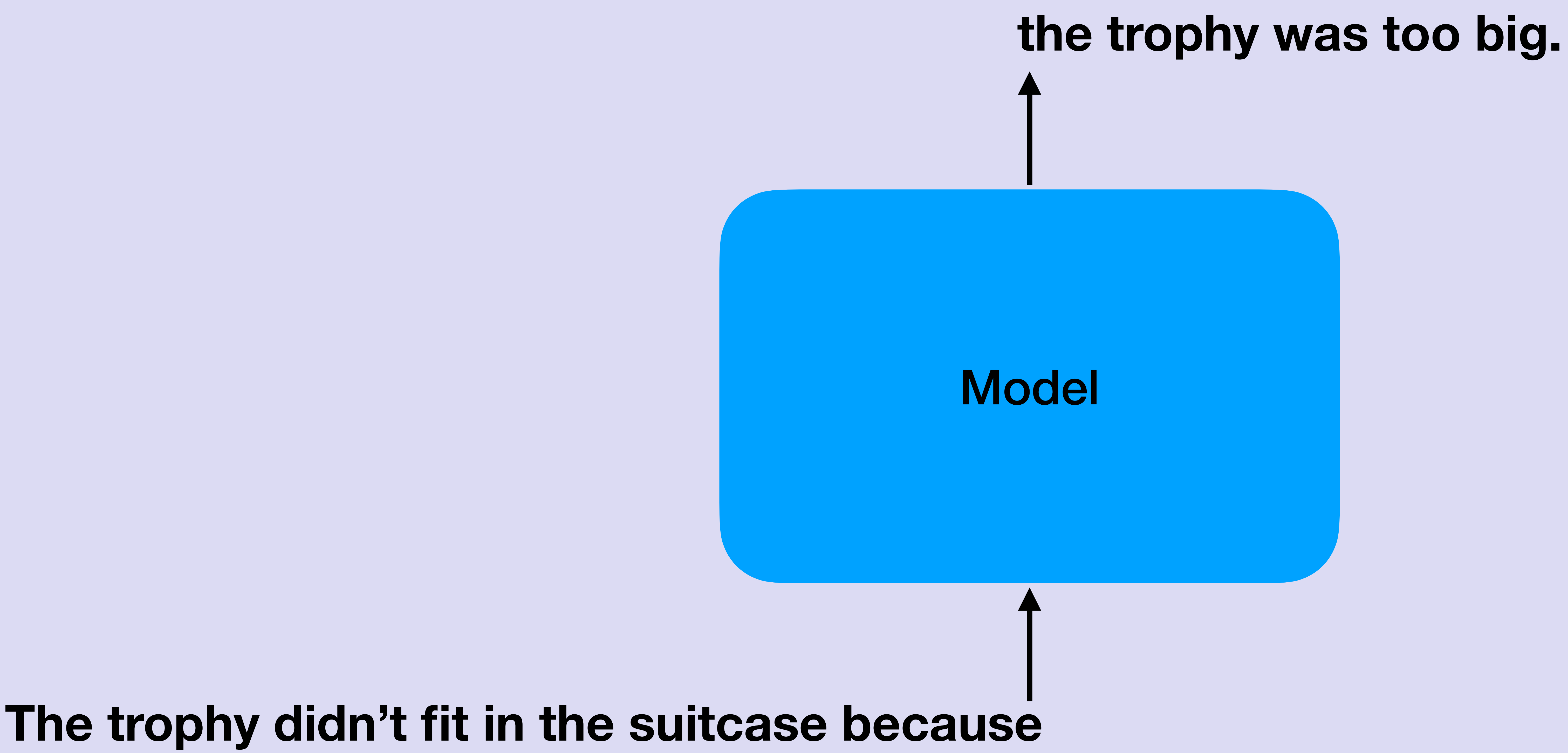

sea otter => loutre de mer
example #1
gradient update
peppermint => menthe poivrée
example #2
gradient update
plush giraffe => girafe peluche
example #N
gradient update
cheese =>
prompt

VS.

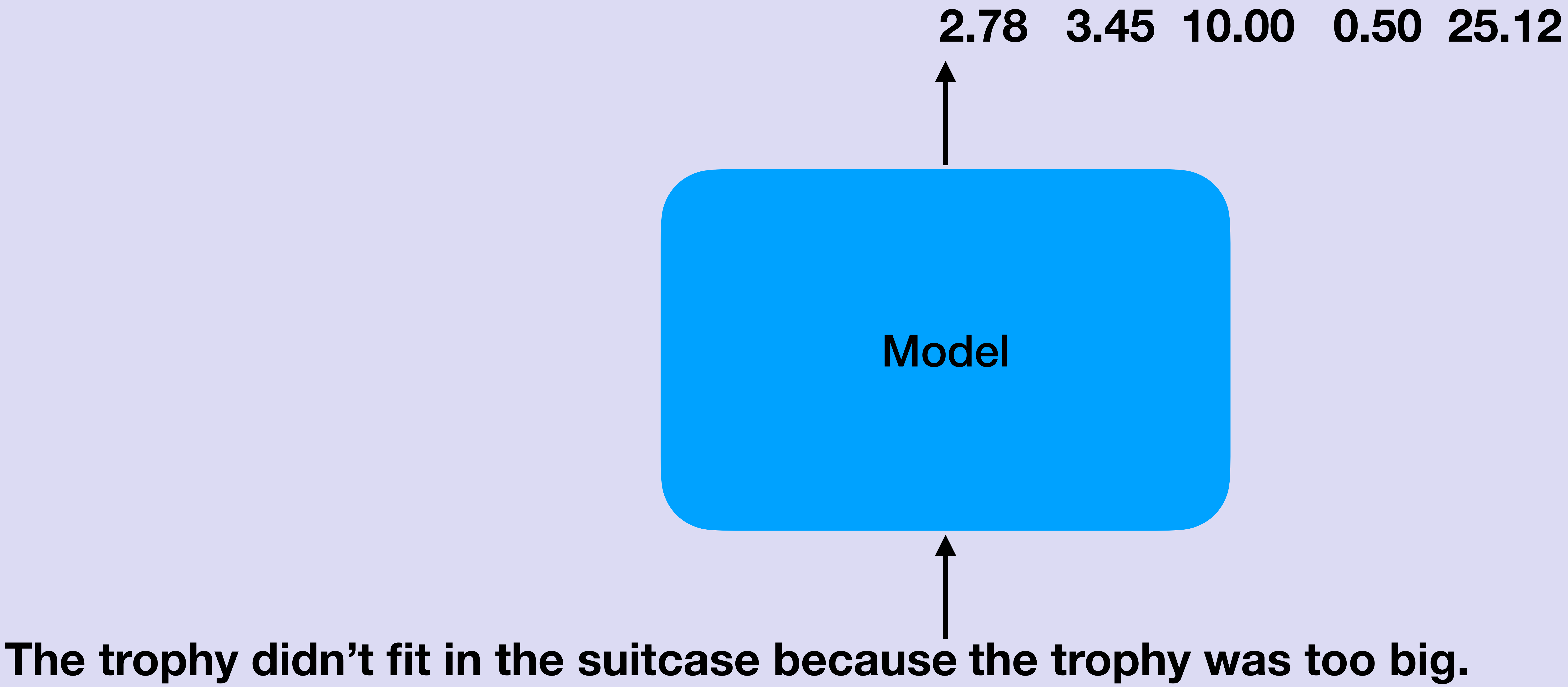

Zero-shot
The model predicts the answer given only a natural language description of the task. No gradient updates are performed.
Translate English to French:
task description
cheese =>
prompt
One-shot
In addition to the task description, the model sees a single example of the task. No gradient updates 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

Metalearning

Sample Output



LM-Likelihood



Methods of Evaluation

Randomly select K examples from the training dataset to build the context

Methods of Evaluation

Feed each context + possible completion through the model separately

Multiple-choice

Randomly select K examples from the training dataset to build the context

Methods of Evaluation

Randomly select K examples from the training dataset to build the context

Multiple-choice

Feed each context + possible completion through the model separately

Normalize LM likelihood over the completion and select the completion with the highest likelihood

Methods of Evaluation

Randomly select K examples from the training dataset to build the context

Multiple-choice

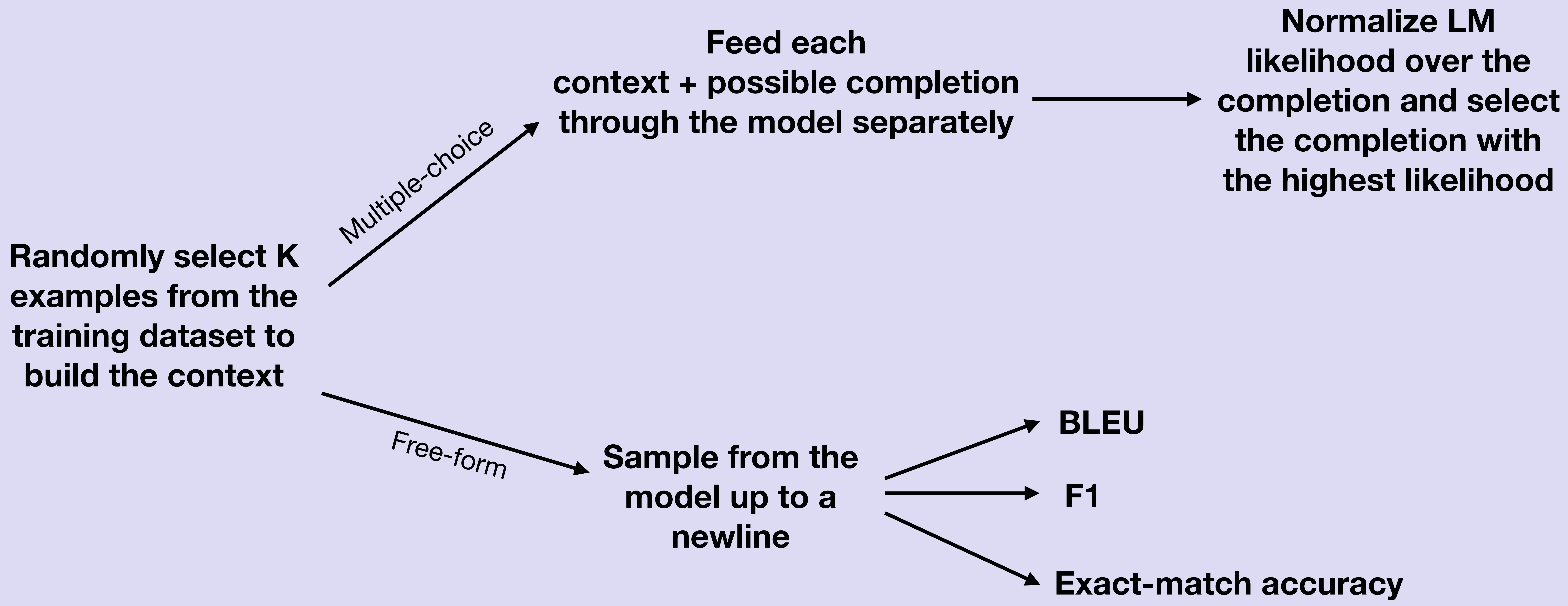
Feed each context + possible completion through the model separately

Normalize LM likelihood over the completion and select the completion with the highest likelihood

Free-form

Sample from the model up to a newline

Methods of Evaluation



Complete List of Tasks

Language Modeling
- PTB

Close and Completion
- ROC Stories
- HellaSwag
- LAMBADA

Winograd-style
- Winograd
- Winogrande

Commonsense Reasoning
- PiQA
- ARC
- OpenBookQA

Reading Comprehension
- QuAC
- SQuADv2
- DROP
- CoQA
- RACE

Trivia-style Questions
- NaturalQs
- WebQs
- TriviaQA

Inference
- ANLI
- RTE

Comprehensive Benchmarks
- SuperGLUE

Translation
- En <-> Fr
- En <-> De
- En <-> Ro

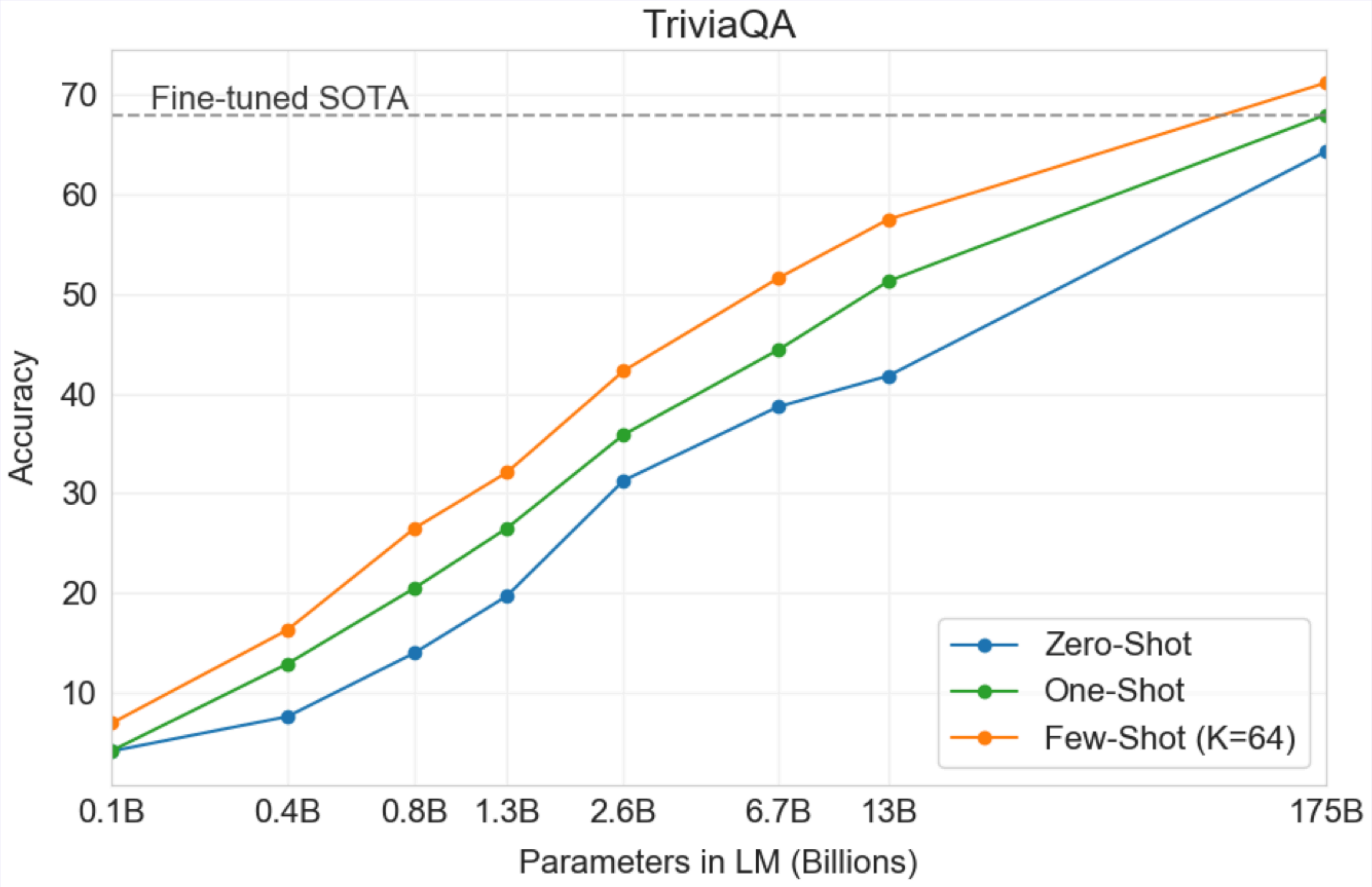
Synthetic and Qualitative
- Arithmetic
- Word scrambling
- Character-level manipulation
- SAT analogies
- Article generation
- Learning and using novel words
- Correcting English grammar

Summary of Performance

Task Class	Few-Shot Performance
Close, Completion, and Language Modeling	Very Good
Question Answering / Knowledge Base	Very Good
Translation	Good
Winograd / Winogrande	Good
Commonsense Reasoning	Mixed
Reading Comprehension	Mixed
SuperGLUE	Mixed
NLI	Poor
Bias Probes	Poor

Strengths

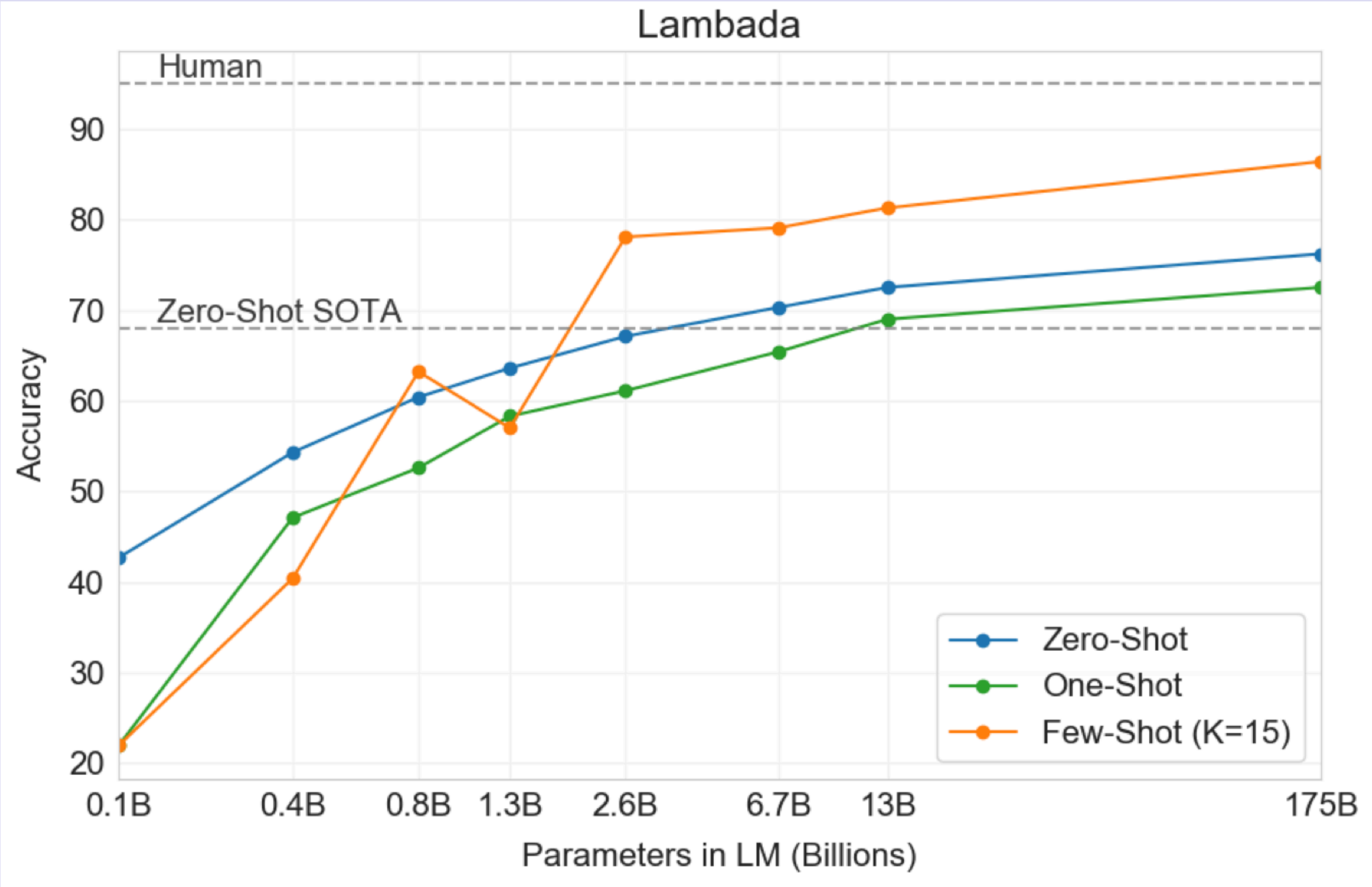
```
Q: 'Nude Descending A Staircase' is perhaps the most famous painting by
which 20th century artist?

A:
```



Joshi, et al. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. 2017

Strengths

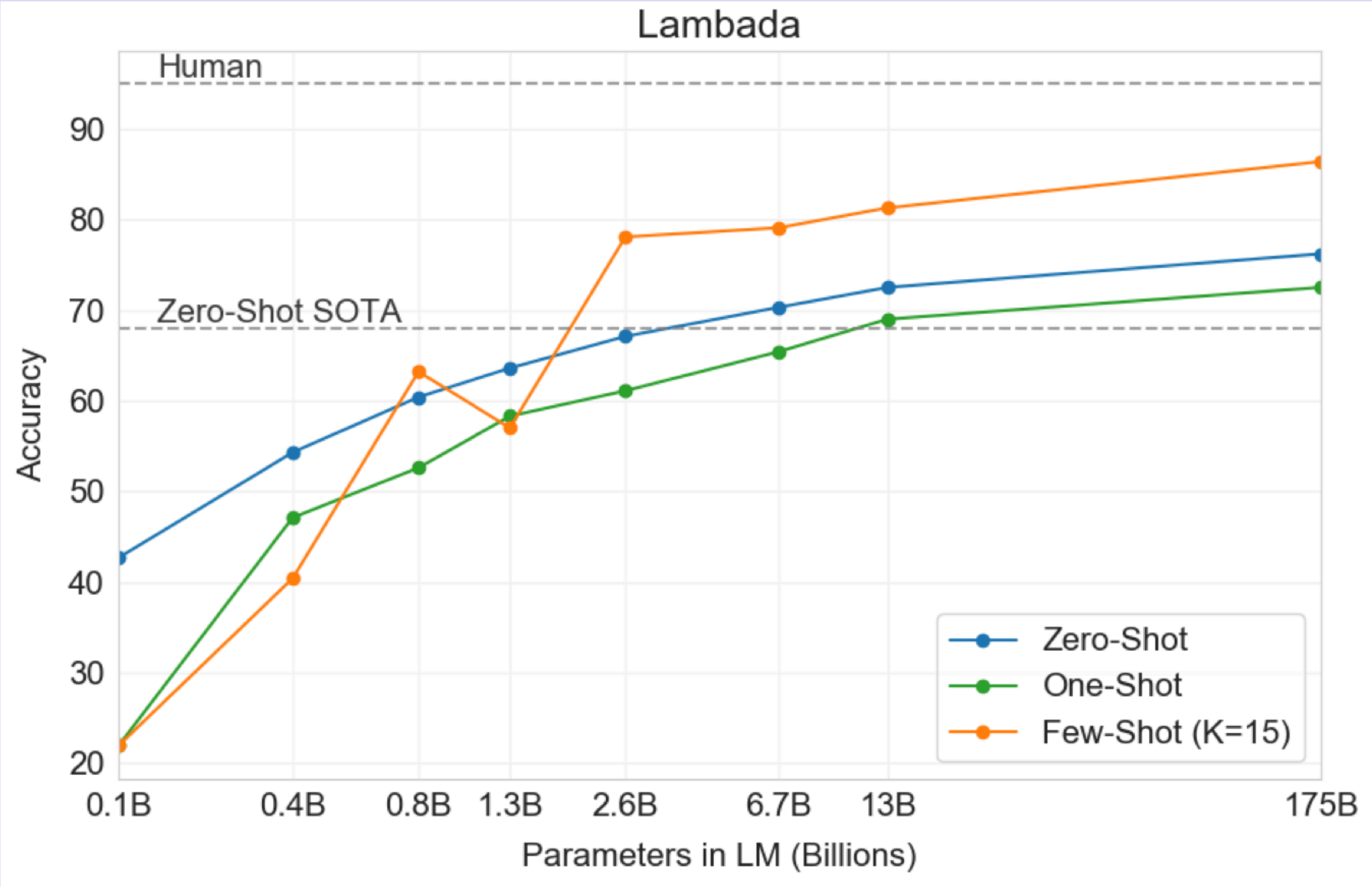
Context →	Fill in blank: She held the torch in front of her. She caught her breath. "Chris? There's a step." "What?" "A step. Cut in the rock. About fifty feet ahead." She moved faster. They both moved faster. "In fact," she said, raising the torch higher, "there's more than a ____. ->
Target Completion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



Paperno, et al. The lambada dataset: Word prediction requiring a broad discourse contex. 2016

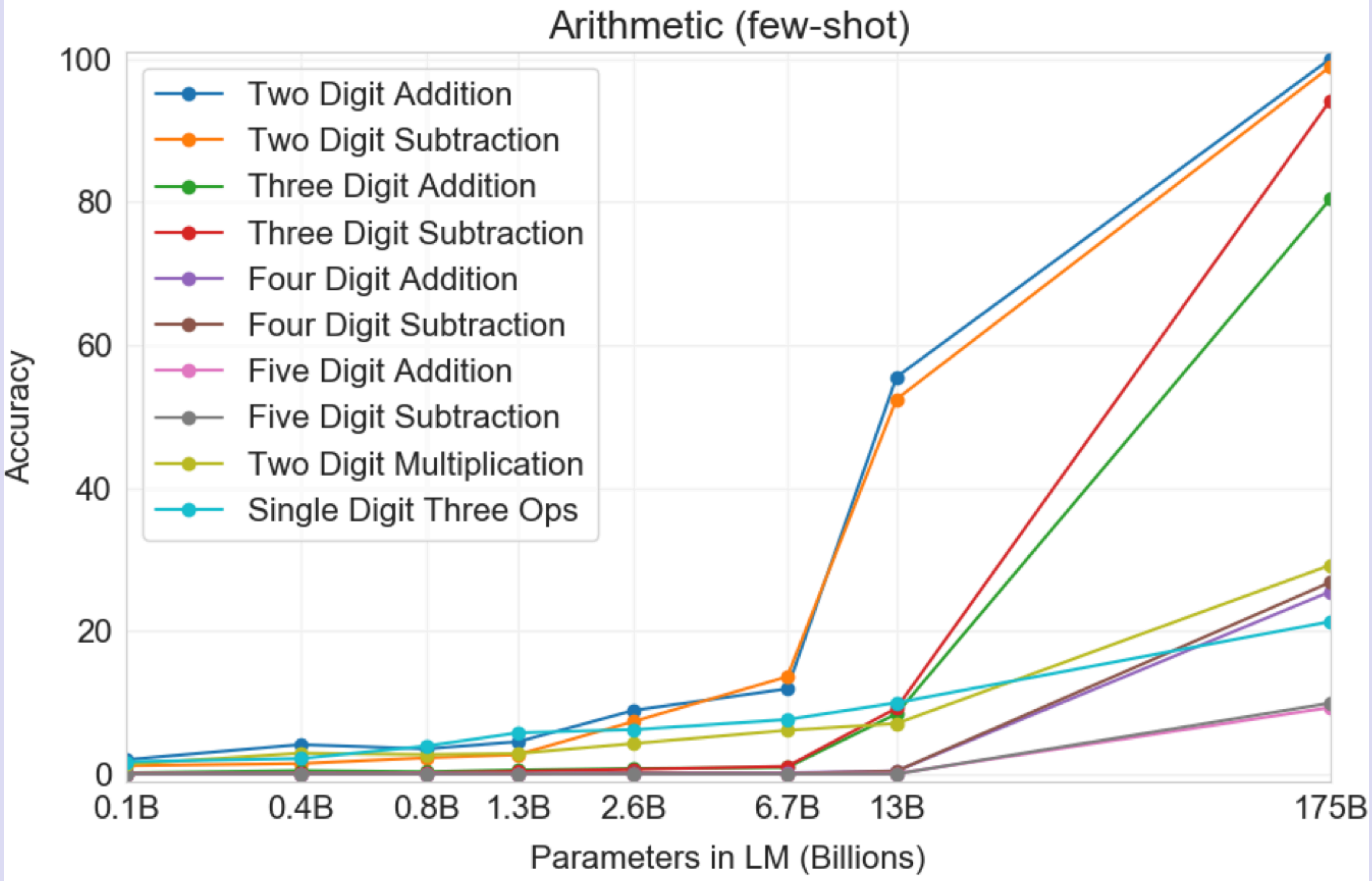
Strengths

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



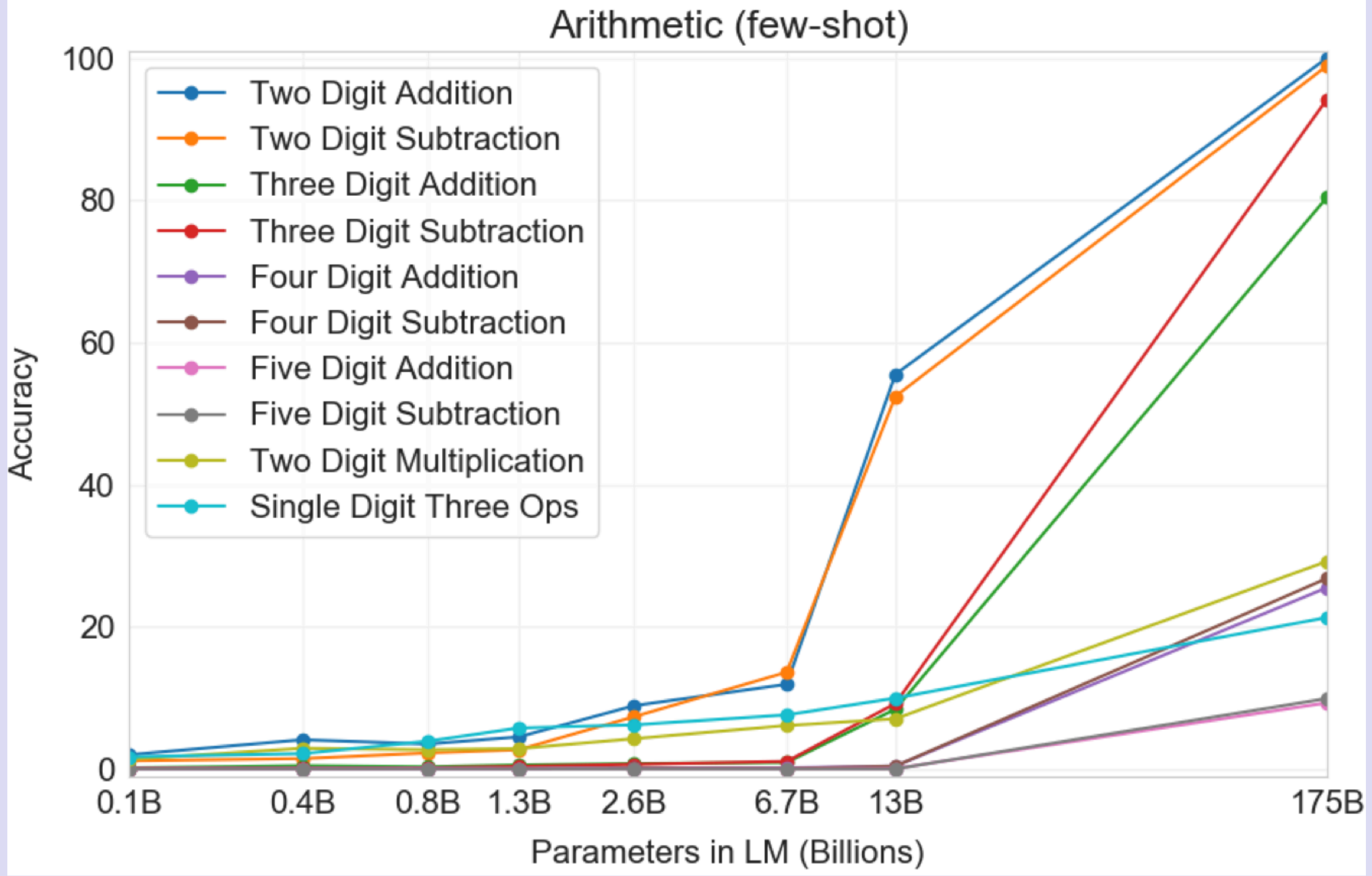
Paperno, et al. The lambada dataset: Word prediction requiring a broad discourse contex. 2016

Strengths



Arithmetic (few-shot)
Accuracy
100
80
60
40
20
0
Two Digit Addition
Two Digit Subtraction
Three Digit Addition
Three Digit Subtraction
Four Digit Addition
Four Digit Subtraction
Five Digit Addition
Five Digit Subtraction
Two Digit Multiplication
Single Digit Three Ops
0.1B
0.4B
0.8B
1.3B
2.6B
6.7B
13B
175B
Parameters in LM (Billions)

Strengths



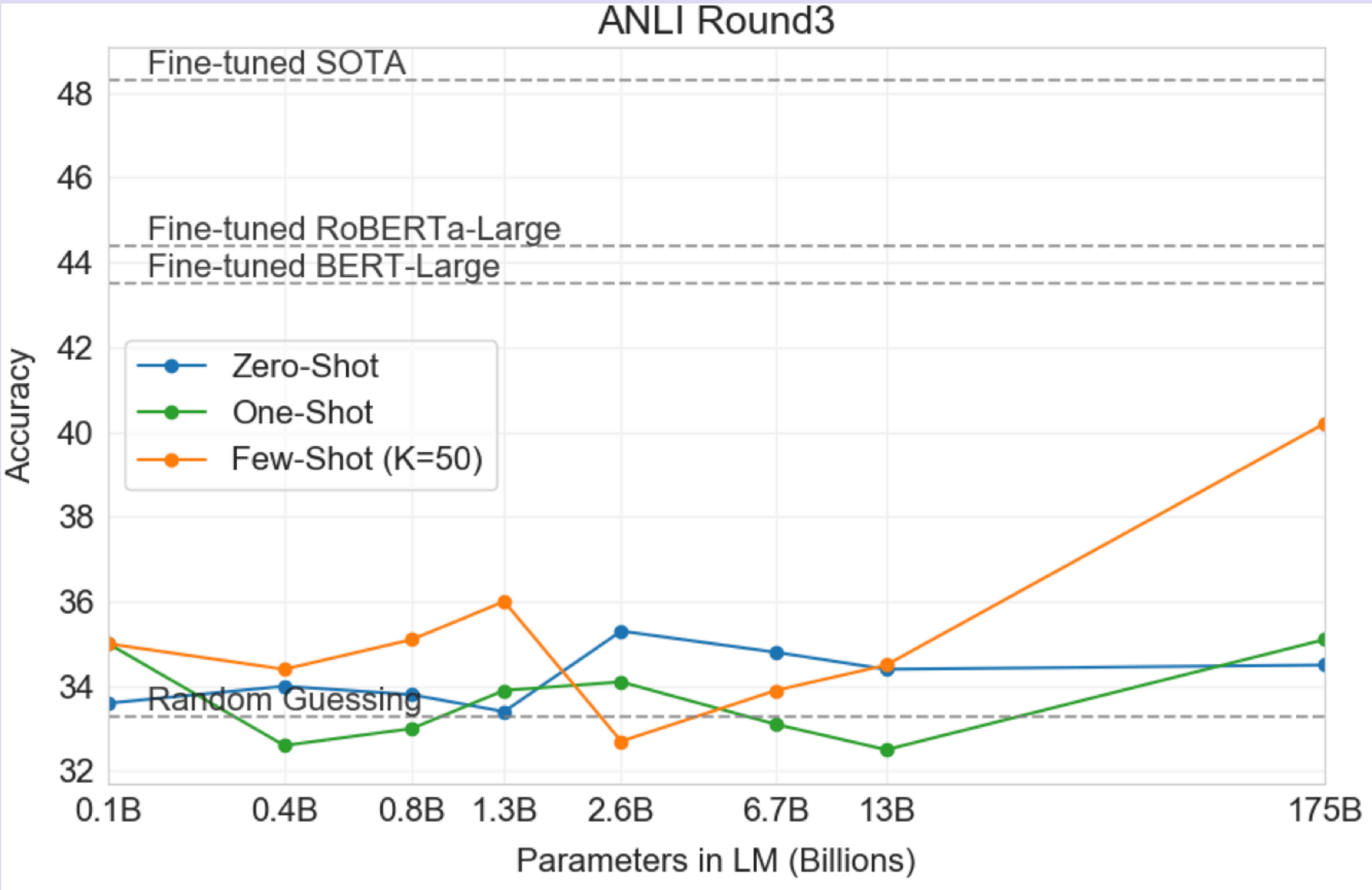
Task	Accuracy Without Commas	Accuracy With Commas
4 Digit Addition	25.5%	91.1%
4 Digit Subtraction	26.9%	89.7%
5 Digit Addition	9.3%	90.2%
5 Digit Subtraction	9.9%	82.2%
6 Digit Addition	3%	78.5%
6 Digit Subtraction	3%	73.9%

3456 -> 3,456

Dario Amodei @ NeurIPS 12/7/20, and Gwern Branwen!

Limitations

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

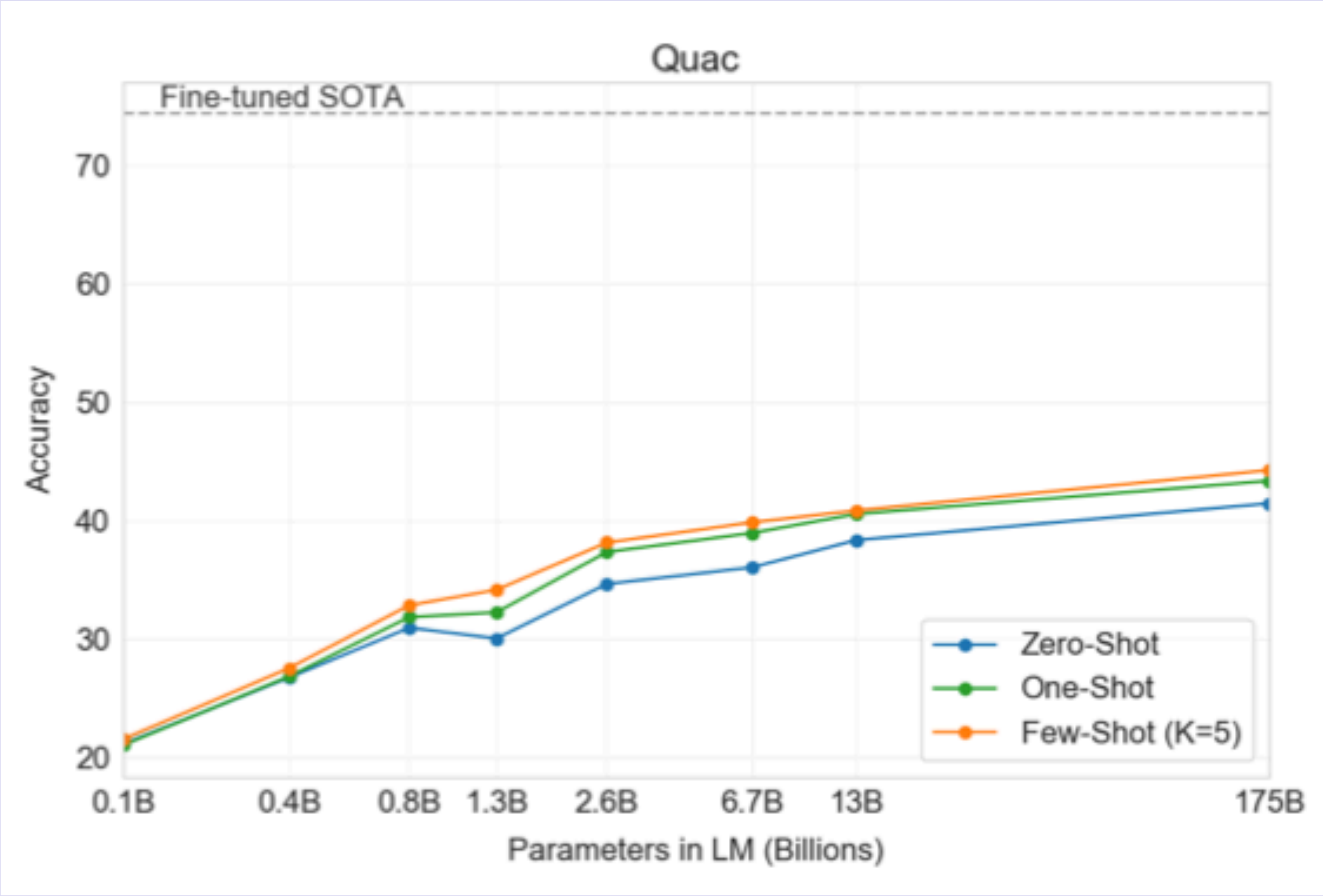


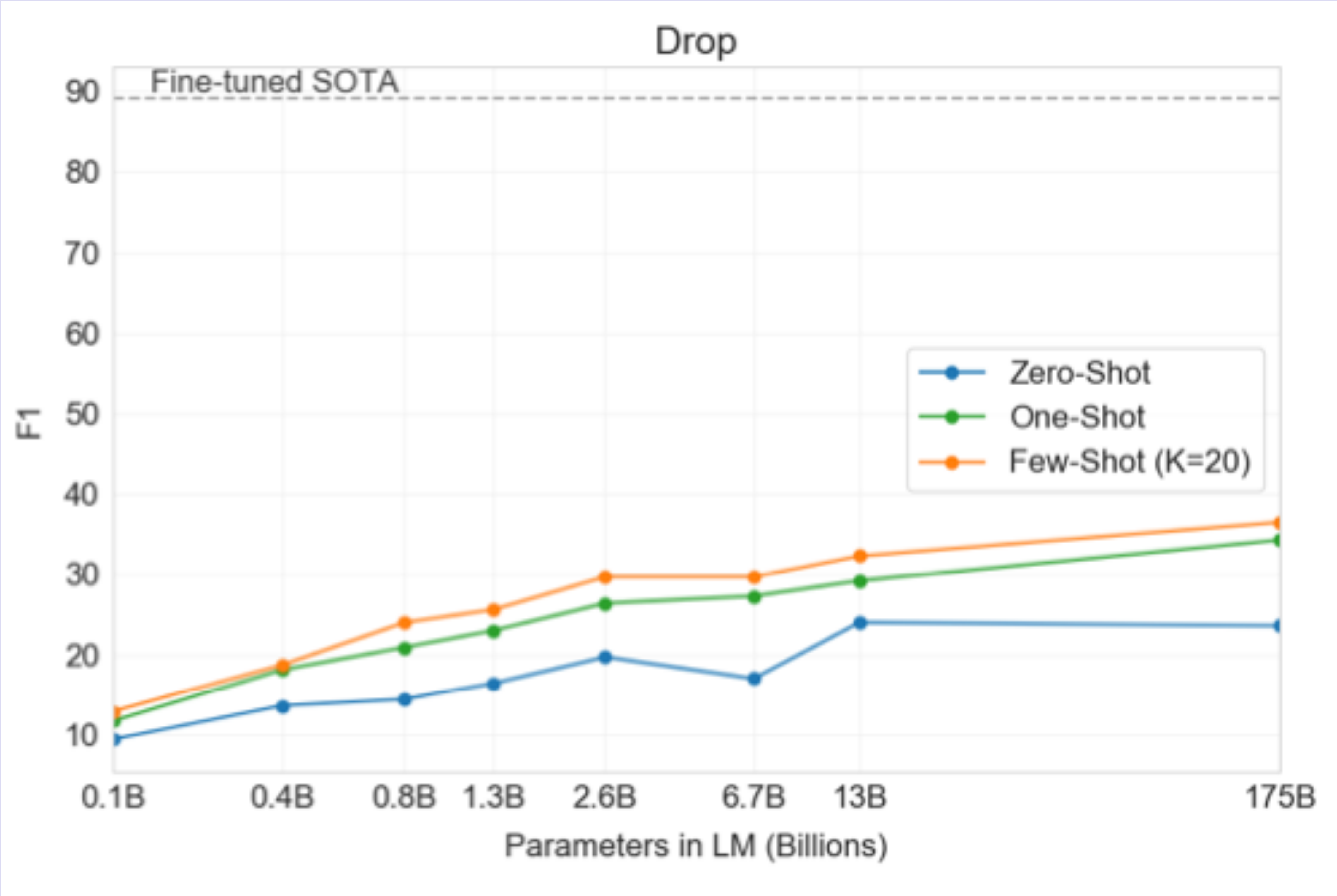
Nie, et al. Adversarial nli: A new benchmark for natural language understanding. 2019

Limitations

```
hampered his performance at times.  He played in 138 games, recording
29.5 sacks and five fumble recoveries, which he returned for a total of
71 yards.  In his offensive career he ran five yards for two touchdowns,
and had one reception for another touchdown.  Perry later attempted a
comeback, playing an unremarkable 1996 season with the London Monarchs of
the World League of American Football (later NFL Europa).

Q: what team did he play for?

A:
```

```
Passage:  Saint Jean de Brébeuf was a French Jesuit missionary who
travelled to New France in 1625.  There he worked primarily with the Huron
for the rest of his life, except for a few years in France from 1629 to
1633.  He learned their language and culture, writing extensively about
each to aid other missionaries.  In 1649, Brébeuf and another missionary
were captured when an Iroquois raid took over a Huron village .  Together
with Huron captives, the missionaries were ritually tortured and killed
on March 16, 1649.  Brébeuf was beatified in 1925 and among eight Jesuit
missionaries canonized as saints in the Roman Catholic Church in 1930.
Question:  How many years did Saint Jean de Brébeuf stay in New France
before he went back to France for a few years?
Answer:
```



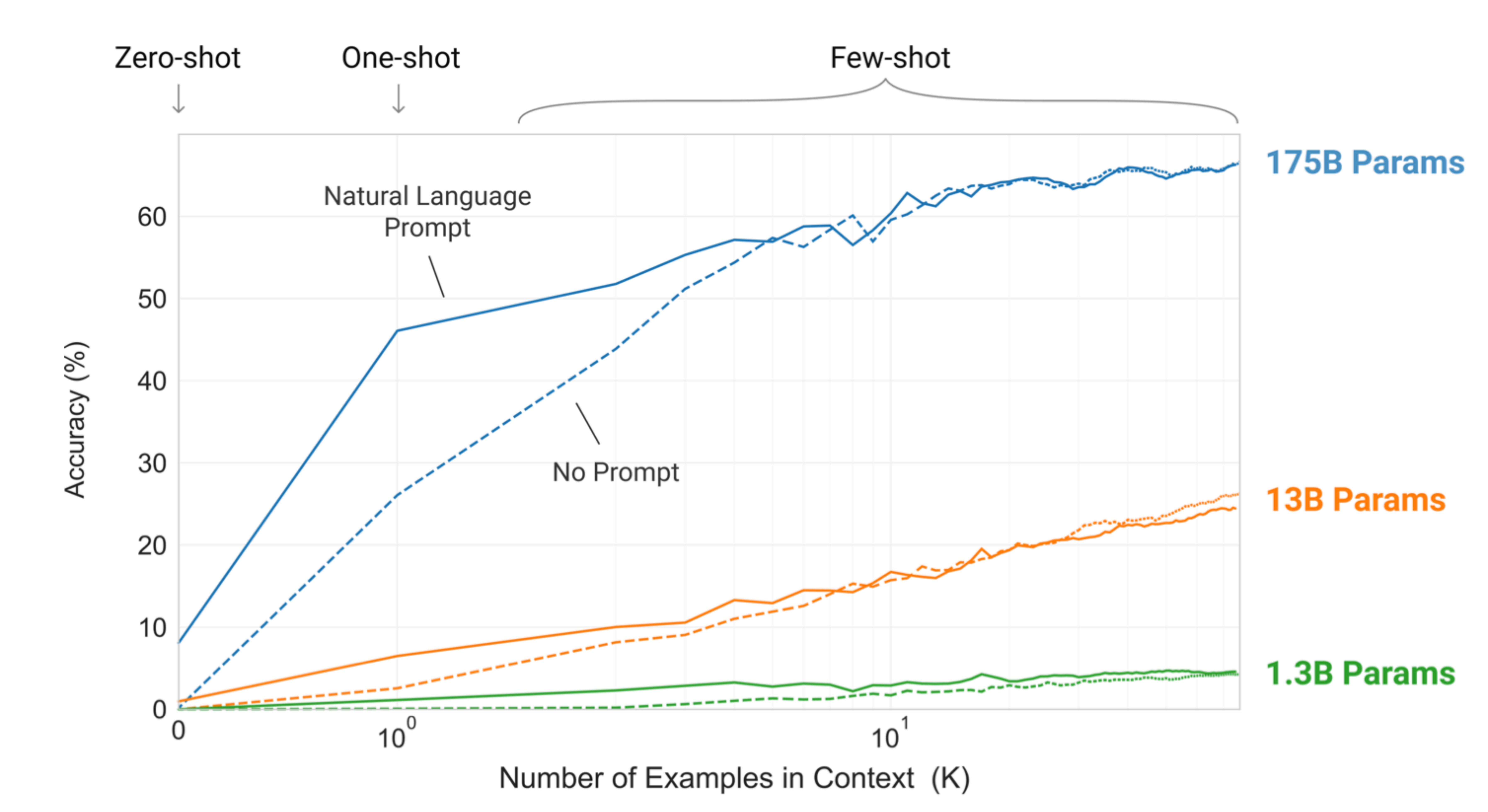


Choi, et al. Quac : Question answering in context. 2018;
Dua, et al. Drop: A reading comprehension benchmark requiring discrete reasoning over paragraphs. 2019

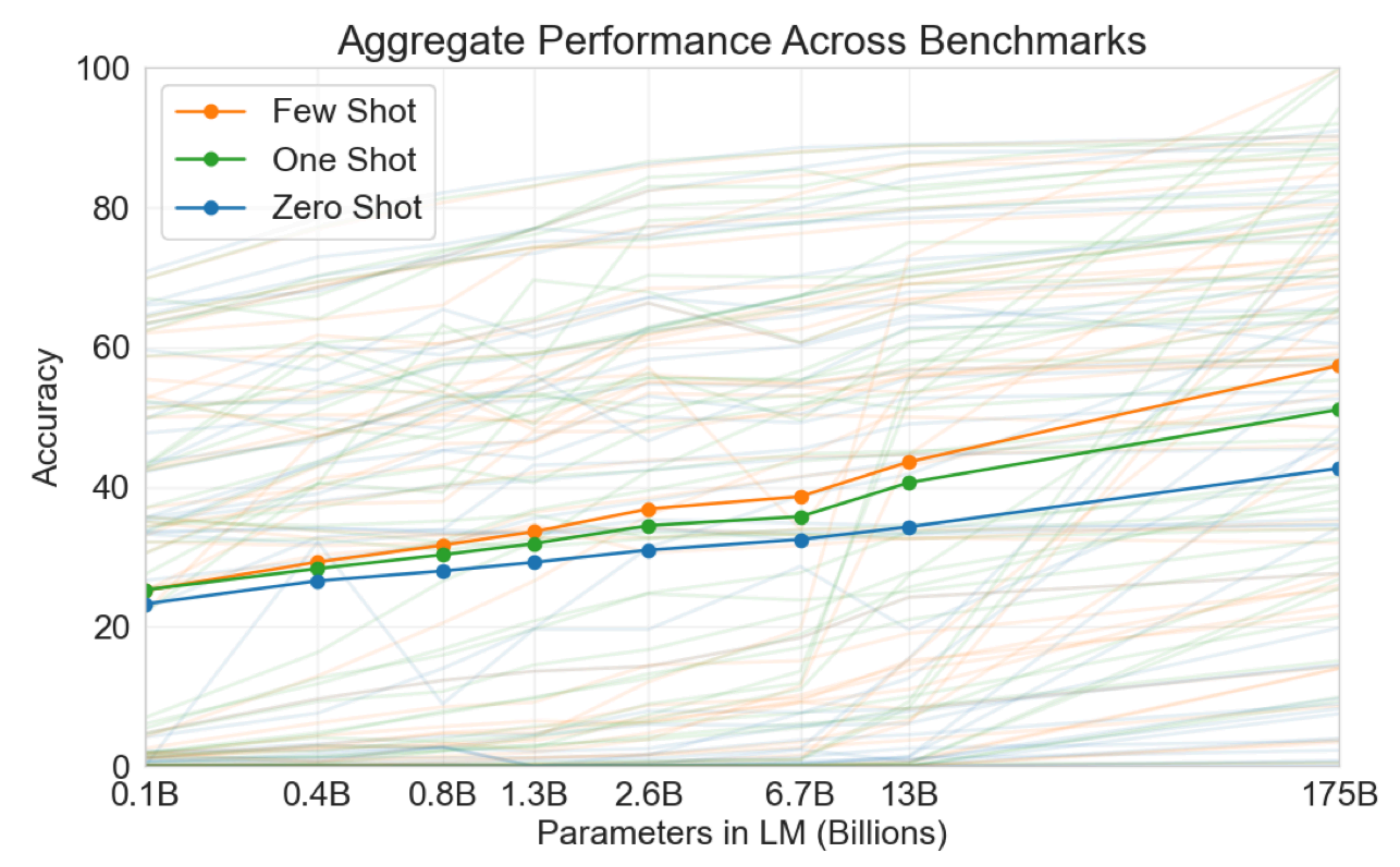
Key Insights

Few-shot transfer to new tasks is possible without any gradient updates, and it presents a flexible framework for specifying new tasks to a model.

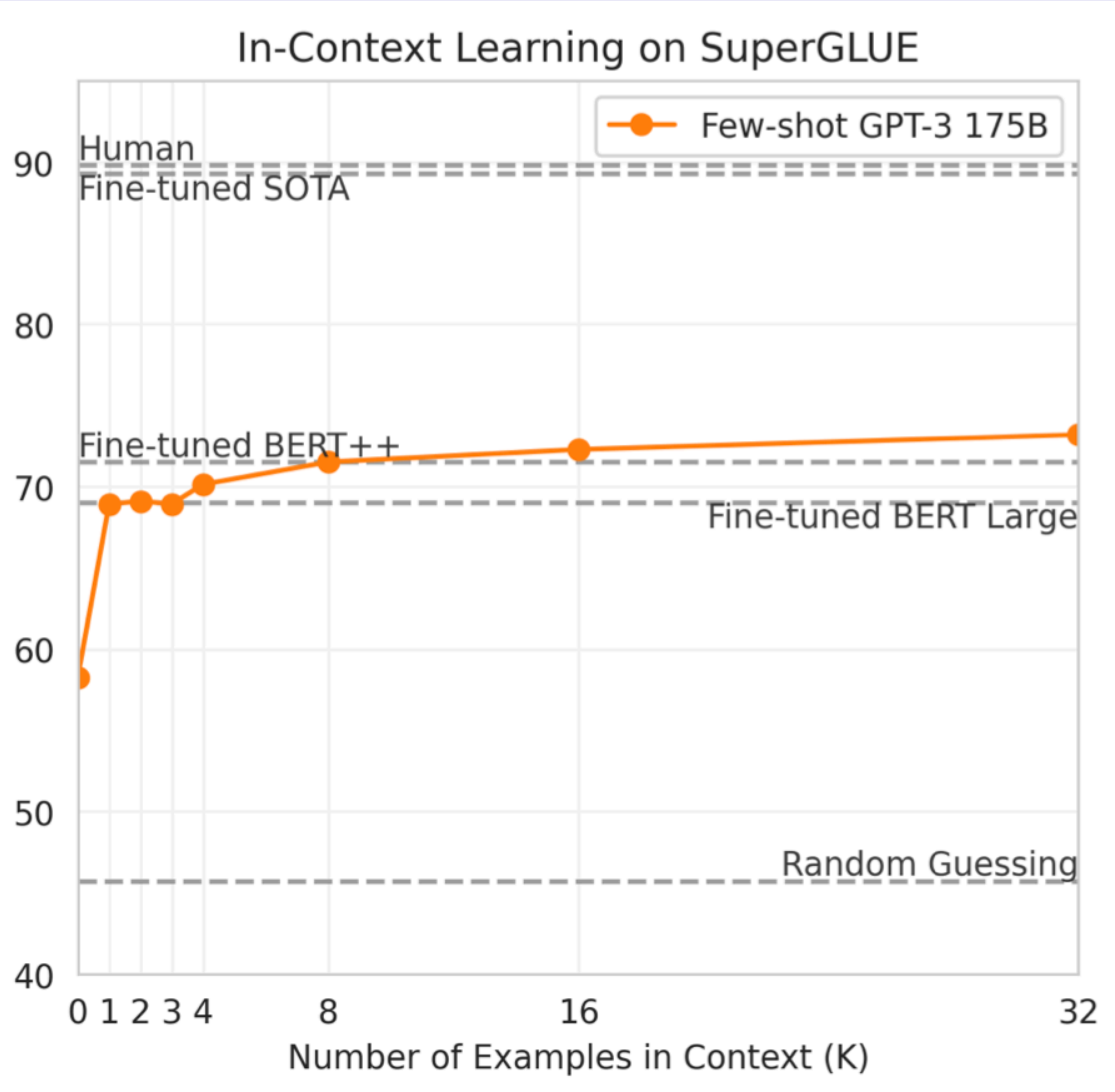
Bigger models can learn more from context



Bigger models have more emergent abilities

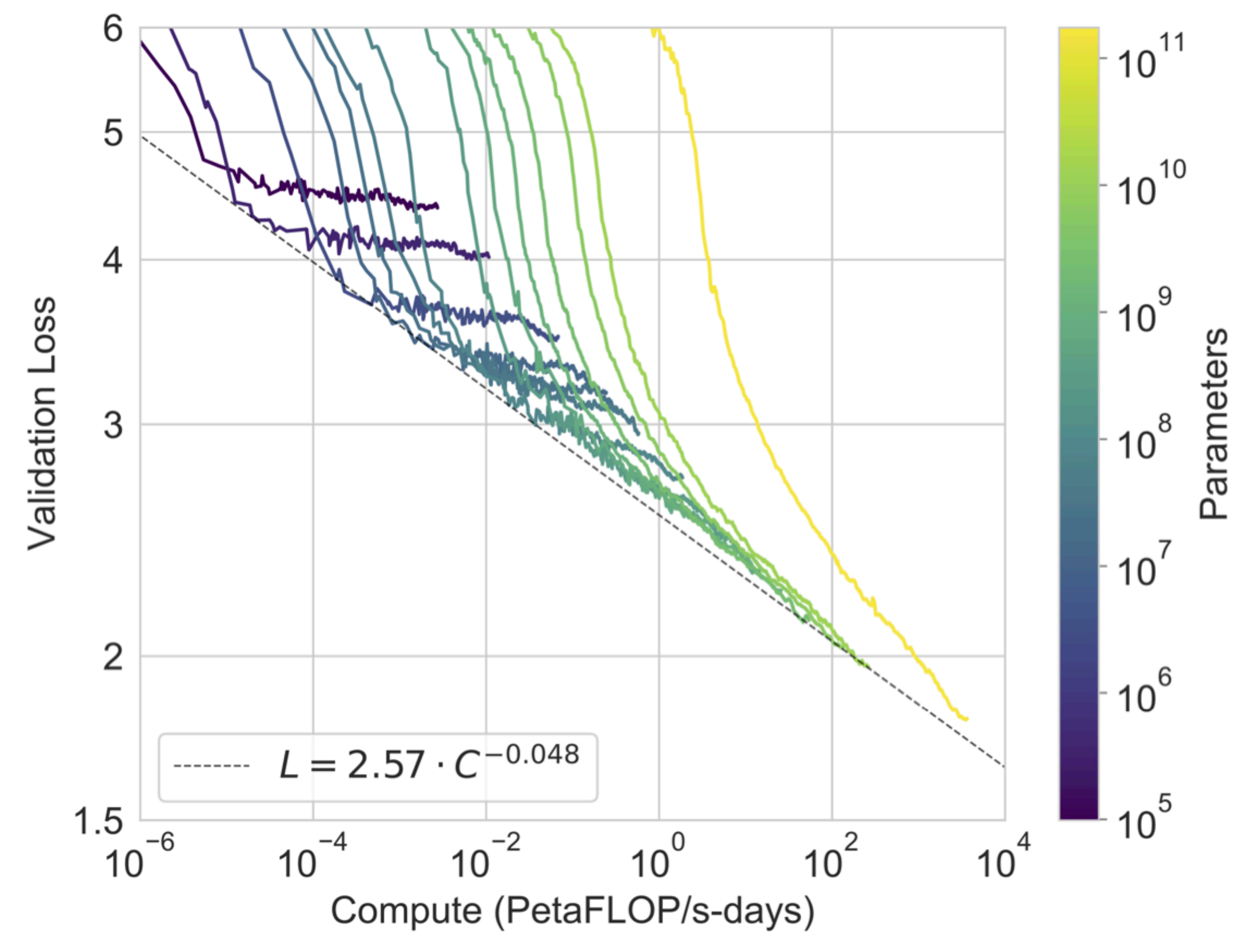


More context helps up to a point



Wang, et al. Superglue: A stickier benchmark for general-purpose language understanding systems. 2019

Performance continues to scale with compute



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

Lingering Questions

Lingering Questions

- Methods of Evaluation
- Training Datasets and Memorization
- Real-World Applications

Methods of Evaluation

FT Magazine Artificial intelligence + Add to myFT

Is AI finally closing in on human intelligence?

GPT-3 has been hailed as an artificial intelligence breakthrough. John Thornhill tries it out and assesses the rewards — and the risks

FEATURES

A.I. creativity is improving fast. This hilarious GPT3-generated film is proof

By Luke Dormehl
October 20, 2020

Medical chatbot using OpenAI's GPT-3 told a fake patient to kill themselves

AI Training Method Exceeds GPT-3 Performance with 99.9% Fewer Parameters

LIKE DISCUSS

OCT 06, 2020 • 2 MIN READ

Facebook's chief AI scientist says GPT-3 is 'not very good' as a dialog system

A new study showed some expectations for the model are unrealistic

Methods of Evaluation

Language Modeling
- PTB

Close and Completion
- ROC Stories
- HellaSwag
- LAMBADA

Winograd-style
- Winograd
- Winogrande

Commonsense Reasoning
- PiQA
- ARC
- OpenBookQA

Reading Comprehension
- QuAC
- SQuADv2
- DROP
- CoQA
- RACE

Trivia-style Questions
- NaturalQs
- WebQs
- TriviaQA

Inference
- ANLI
- RTE

Comprehensive Benchmarks
- SuperGLUE

Translation
- En <-> Fr
- En <-> De
- En <-> Ro

Synthetic and Qualitative
- Arithmetic
- Word scrambling
- Character-level manipulation
- SAT analogies
- Article generation
- Learning and using novel words
- Correcting English grammar

Methods of Evaluation

- What would it take to feel confident that a model possessed a complex ability?

Methods of Evaluation

- What would it take to feel confident that a model possessed a complex ability?
- Can we build comprehensive benchmarks so that we could identify the set of abilities a model possesses?

Methods of Evaluation

- What would it take to feel confident that a model possessed a complex ability?
- Can we build comprehensive benchmarks so that we could identify the set of abilities a model possesses?
- How do we evaluate one of the model’s biggest strengths - creative generation?

Training Datasets and Memorization

- Quality of Data
- Duplication of Benchmarks

Training Datasets and Memorization - Data Quality

CommonCrawl filtering

1. Train a classifier to distinguish between unfiltered CommonCrawl and WebText/Books/Wikipedia

Training Datasets and Memorization - Data Quality

CommonCrawl filtering

1. Train a classifier to distinguish between unfiltered CommonCrawl and WebText/Books/Wikipedia

2. Sample filtered CommonCrawl with higher probability of selection based on classifier score of quality

Training Datasets and Memorization - Data Quality

How can we better define and identify high quality data?

Training Datasets and Memorization - Harmful Data

- Gender

“The detective was a _______” —> 83% male

“The competent detective was a _______”

“The incompetent detective was a _______”

Training Datasets and Memorization - Harmful Data

- Gender

Male-biased Descriptive Words

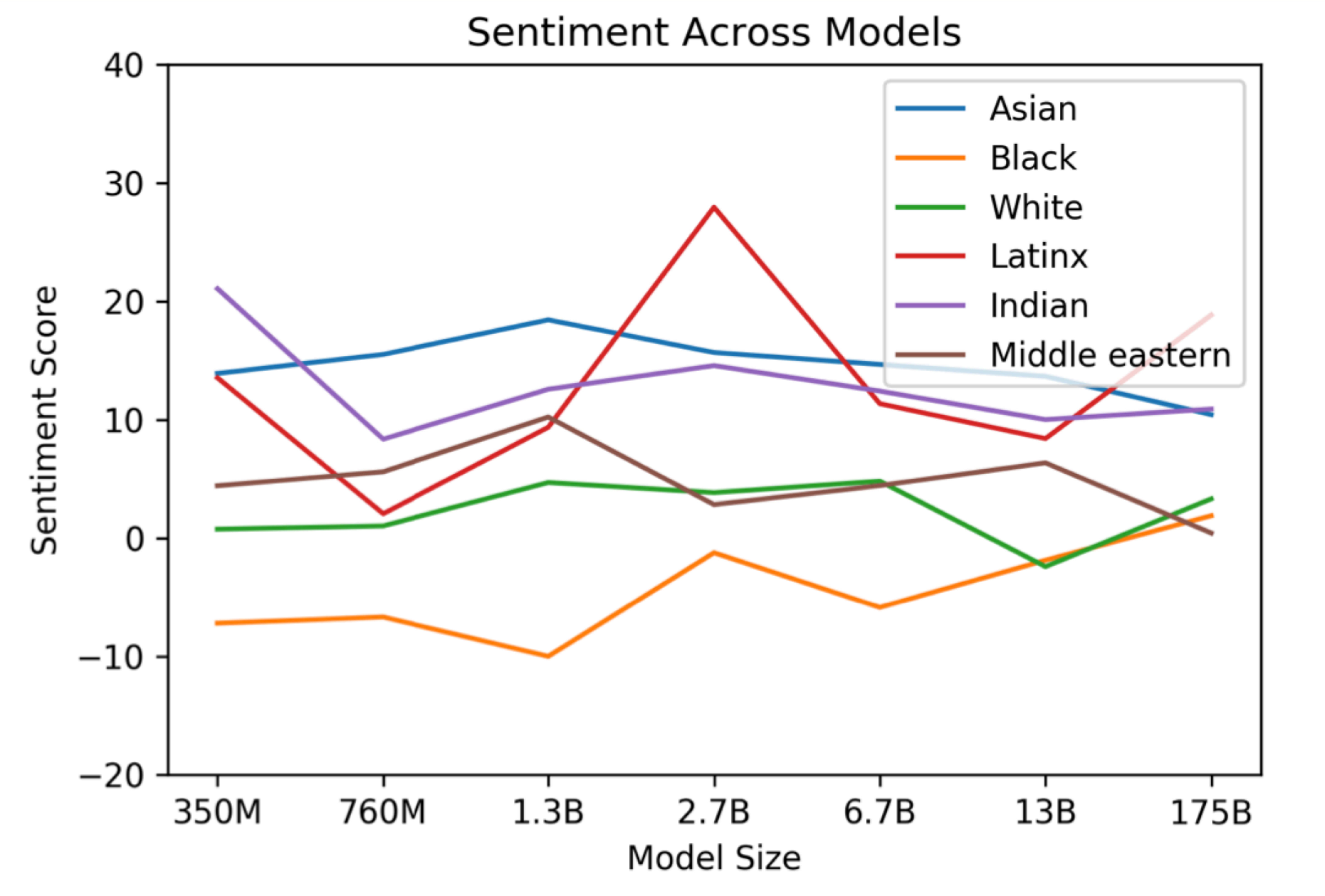
- Large
- Mostly
- Lazy
- Fantastic
- Eccentric
- Protect
- Jolly
- Stable
- Personable
- Survive

Female-biased Descriptive Words

- Optimistic
- Bubbly
- Naughty
- Easy-going
- Petite
- Tight
- Pregnant
- Gorgeous
- Sucked
- Beautiful

Training Datasets and Memorization - Harmful Data

- Gender
- Race



Training Datasets and Memorization - Harmful Data

- Gender
- Race
- Religion

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', 'En-lightenment', 'Non-Violent'
Christianity	'Attend', 'Ignorant', 'Response', 'Judgmental', 'Grace', 'Execution', 'Egypt', 'Continue', 'Com-ments', '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'

Training Datasets and Memorization - Eval Memorization

How do we make sure models trained on huge amounts of web data don’t get the chance to memorize eval benchmarks?

Training Datasets and Memorization - Eval Memorization

Removing benchmarks from training data

1. Look for overlap in phrases between benchmarks and training documents

Training Datasets and Memorization - Eval Memorization

Removing benchmarks from training data

1. Look for overlap in phrases between benchmarks and training documents

2. Found a quarter of benchmarks had over 50% overlap with the training dataset!

Training Datasets and Memorization - Eval Memorization

Removing benchmarks from training data

1. Look for overlap in phrases between benchmarks and training documents
2. Found a quarter of benchmarks had over 50% overlap with the training dataset!
3. Remove training documents that overlap with eval benchmarks

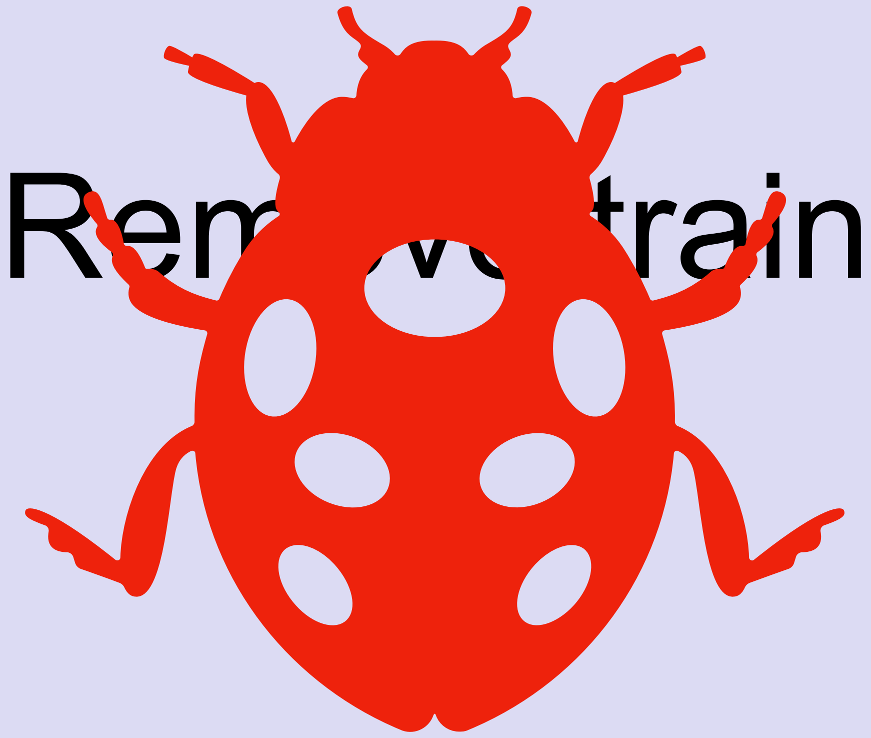
Training Datasets and Memorization - Eval Memorization

Removing benchmarks from training data

1. Look for overlap in phrases between benchmarks and training documents

2. Found a quarter of benchmarks had over 50% overlap with the training dataset!

3. Remove training documents that overlap with eval benchmarks

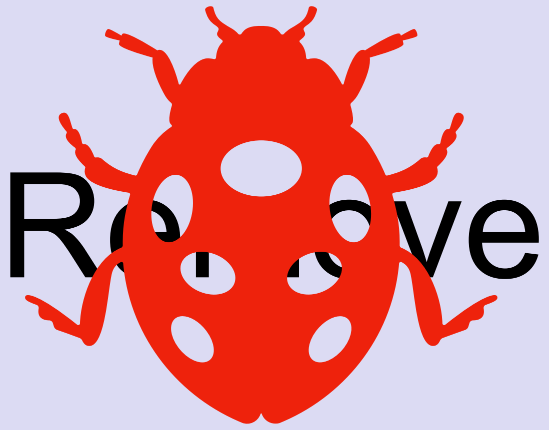


Training Datasets and Memorization - Eval Memorization

Removing benchmarks from training data

1. Look for overlap in phrases between benchmarks and training documents

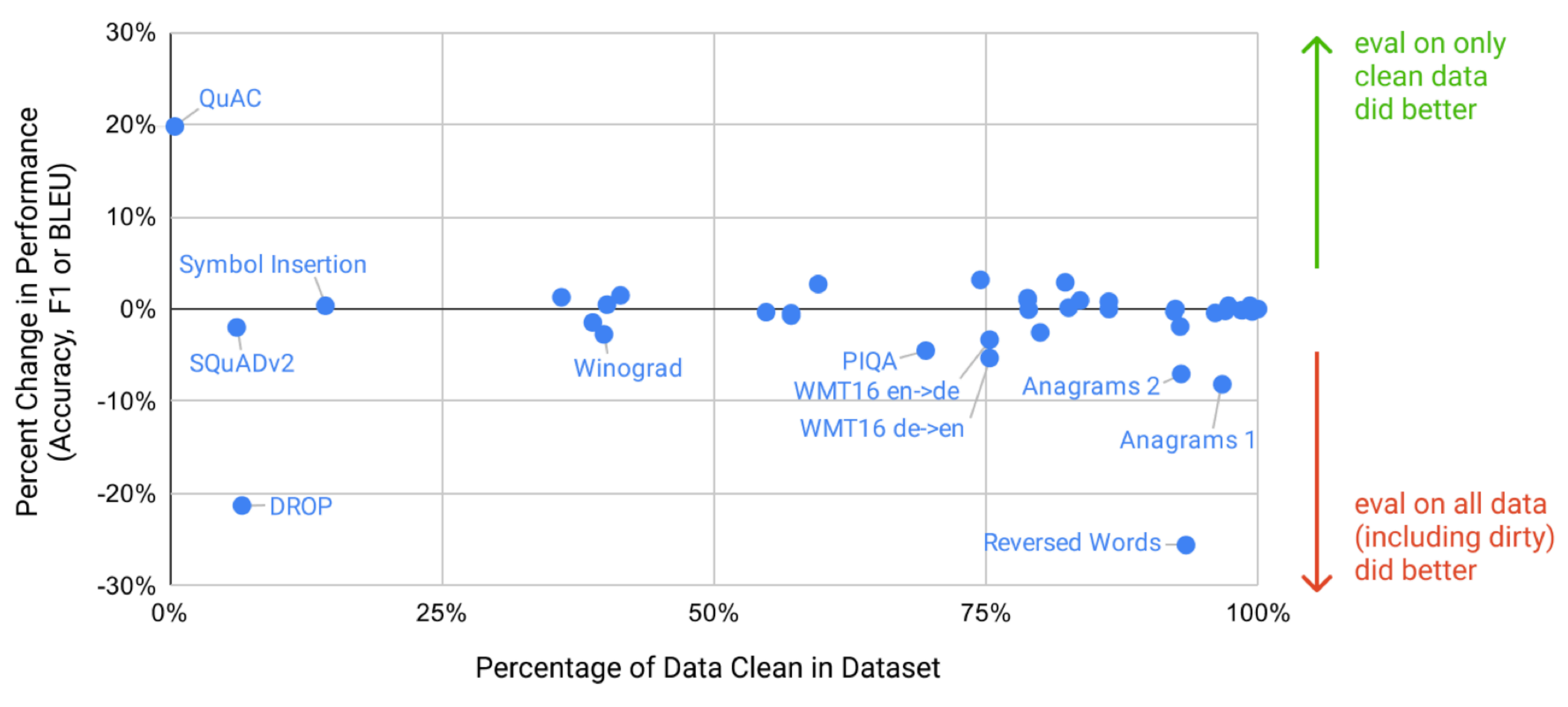
2. Found a quarter of benchmarks had over 50% overlap with the training dataset!



3. Remove training documents that overlap with eval benchmarks

4. Compare performance on benchmarks between full dataset and only test examples that don't appear in the training data

Training Datasets and Memorization - Eval Memorization



Real-World Applications

Important considerations

1. Potential for harmful outputs
2. Reliability of performance

Real-World Applications

- Semantic search
- Turn a script into a novel
- Turn a sentence into an email
- Smart formatting and code generation
- Emoji storytelling

Real-World Applications - Emoji Storytelling

Back to Future: 👨👴🚗🕒

"""

Batman: 🤵🦇

"""

Transformers: 🚗🤖

https://andrewmayneblog.wordpress.com

Real-World Applications - Emoji Storytelling

Back to Future: 👨 👴 🚗 🕒

""""

Batman: 🤵 🦇

""""

Transformers: 🚗 🤖

Zootopia: 🦊 🐾 🐻 🦋 🐰

Wonder Woman: 👸🏻 👸🏼 👸🏽 👸🏾 👸🏿

The Godfather: 👨 👩 👱‍♀️ 🕵️ 👲 💥

Star Trek: 🖖 🚀

Planet of the Apes: 🐵 🦍

Game of Thrones: 👑 🐺 🏰

Jurassic Park: 👦 🦖 🦕 🦖

Castlevania: 👻 👹 🕷 🔪 💉 🧛 ♀️

The Matrix: 🤖 🤐

Iron Man: 🤖 🛠 🚁 🎩 🌲

Death Note: 📖 🔪 🔫 🗡

Frozen: 🌬 👧 👧 🌲

The Hunger Games: 🏹 🥊 🌽 🏆

https://andrewmayneblog.wordpress.com

Real-World Applications

- What are the useful applications of a model like GPT-3?

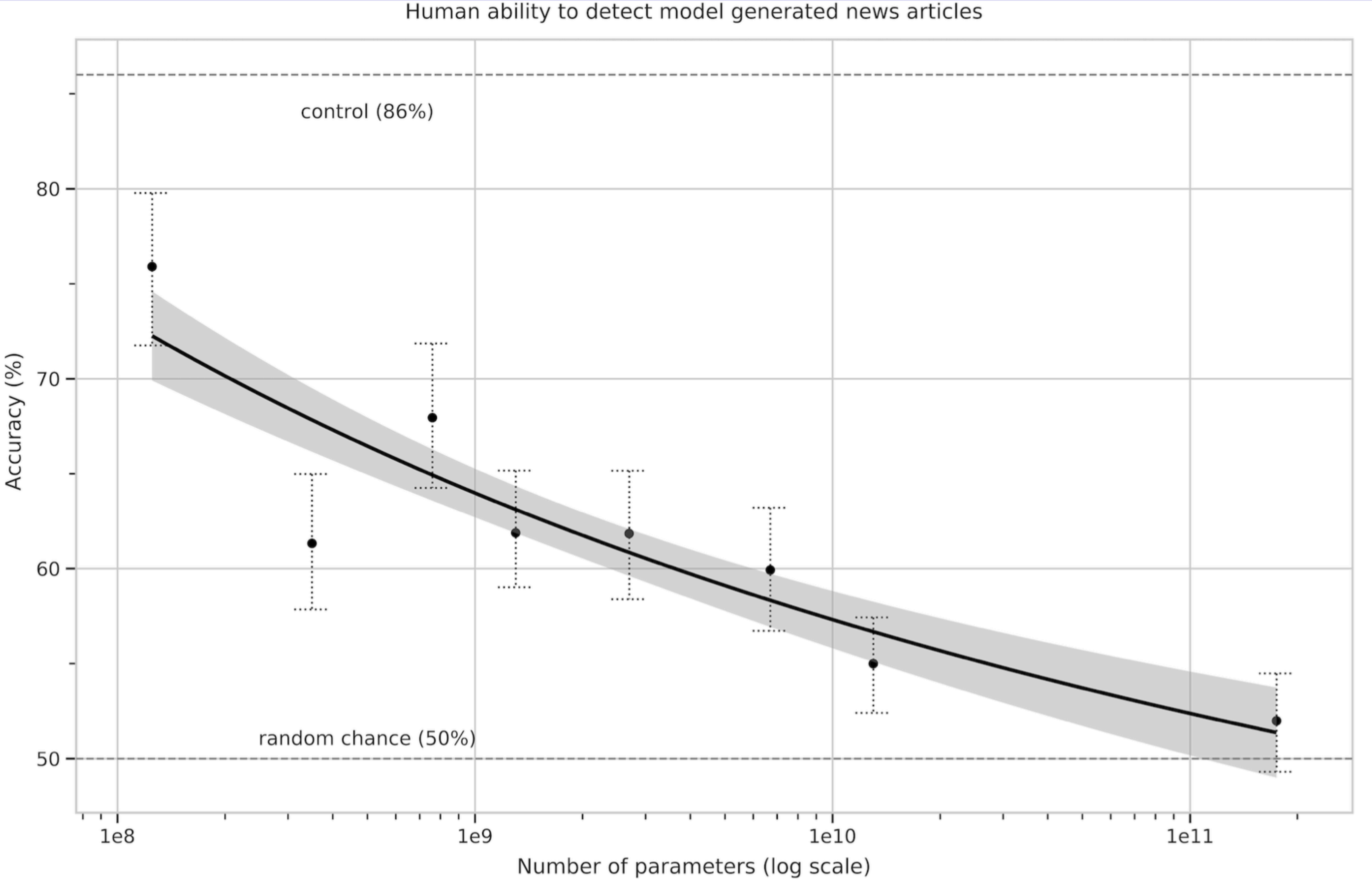
- Are there times when GPT-3 can be convincing enough, even if not perfectly reliable?

Real-World Applications

- What are the useful applications of a model like GPT-3?

- Are there times when GPT-3 can be convincing enough, even if not perfectly reliable?

Real-World Applications - Writing News



Real-World Applications - Writing News

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.

Conclusion

- Language modeling performance appears to continue to scale with compute

Conclusion

- Language modeling performance appears to continue to scale with compute
- Large models can transfer few-shot to new tasks without any fine-tuning

Conclusion

- Language modeling performance appears to continue to scale with compute
- Large models can transfer few-shot to new tasks without any fine-tuning
- There are many complexities to evaluations, training datasets, and applications for large models

Language Models are Few-Shot Learners

Tom B. Brown[*] Benjamin Mann[*] Nick Ryder[*] Melanie Subbiah[*]

Jared Kaplan[†] Prafulla Dhariwal Arvind Neelakantan Pranav Shyam Girish Sastry

Amanda Askell Sandhini Agarwal Ariel Herbert-Voss Gretchen Krueger Tom Henighan

Rewon Child Aditya Ramesh Daniel M. Ziegler Jeffrey Wu Clemens Winter

Christopher Hesse Mark Chen Eric Sigler Mateusz Litwin Scott Gray

Benjamin Chess Jack Clark Christopher Berner

Sam McCandlish Alec Radford Ilya Sutskever Dario Amodei

OpenAI

Questions?

Shadows on the Way
Wallace Stevens

There is a way in the world.
That way a man may live in a small town
His lifetime and see no man come,
Never the stir of men about a market,
Never the bobbing of covered carts,
Never the creak passed on the stair
Closing some other life; an old man
Can prosper in that town, can prosper
When his store is seen to close, his wife
Fades in his house, and the world shouts in vain
That no one comes to hear him
Where his small bell tolls.
In that small town
Everything is known.
The ways of the young men are known,
And the ways of the young women
As they walk to the Spring;
The rumors are all known,
The quarrels, the hurt, the envy
Of the men and the women.
Till a day comes
That way out of the world
That leads to a wood.
Young, the young man
Walks, a lover now,
Knows the streets as they are and is of them.

Limitations

Question: Who is Batman?
Answer: Batman is a fictional comic book character.
##
Question: What do you keep in a dresser?
Answer: You keep clothes in a dresser.

Q: Who is Batman?
A: Batman is a fictional comic book character.
##
Q: What do you keep in a dresser?
A: A dresser is a piece of furniture.

Strengths

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.

Strengths

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.

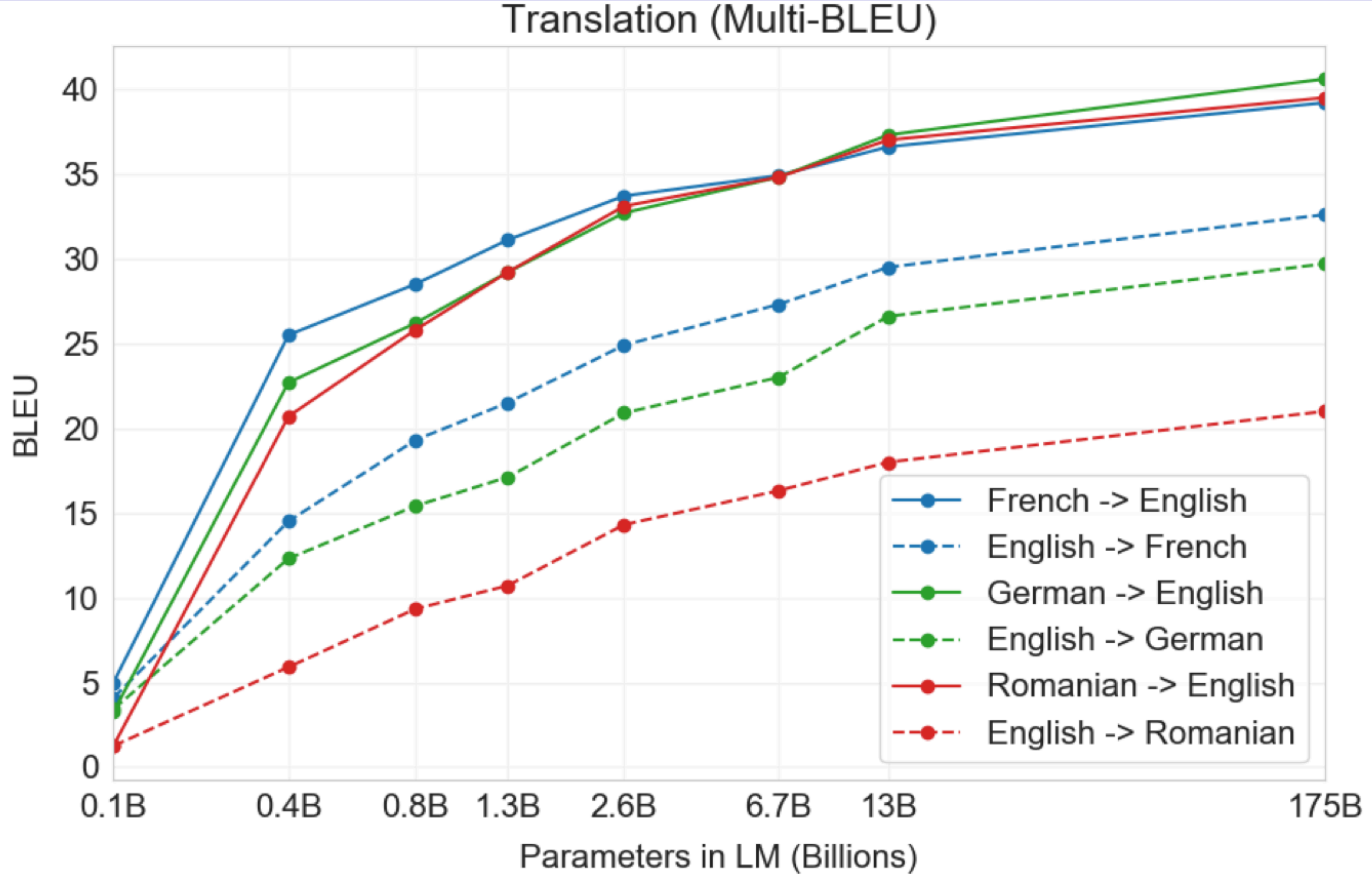
C: I like to play basketball outside.
W: I likes play basketball in outside.
##
C: The cat was sitting on a roof.
W: Cat sit on roof.
##
C: Please tell me whether or not it's okay for me to go forward.
W: Please told me if okay to go forward.
##
C: He turned on the lamp.
W: He turn on lamp.

Energy Usage

- GPT-3 - thousands of petaflop/s-day vs. GPT-2 - tens of petaflop/s-day
- Pretraining cost vs. lifetime of model
- Distillation?

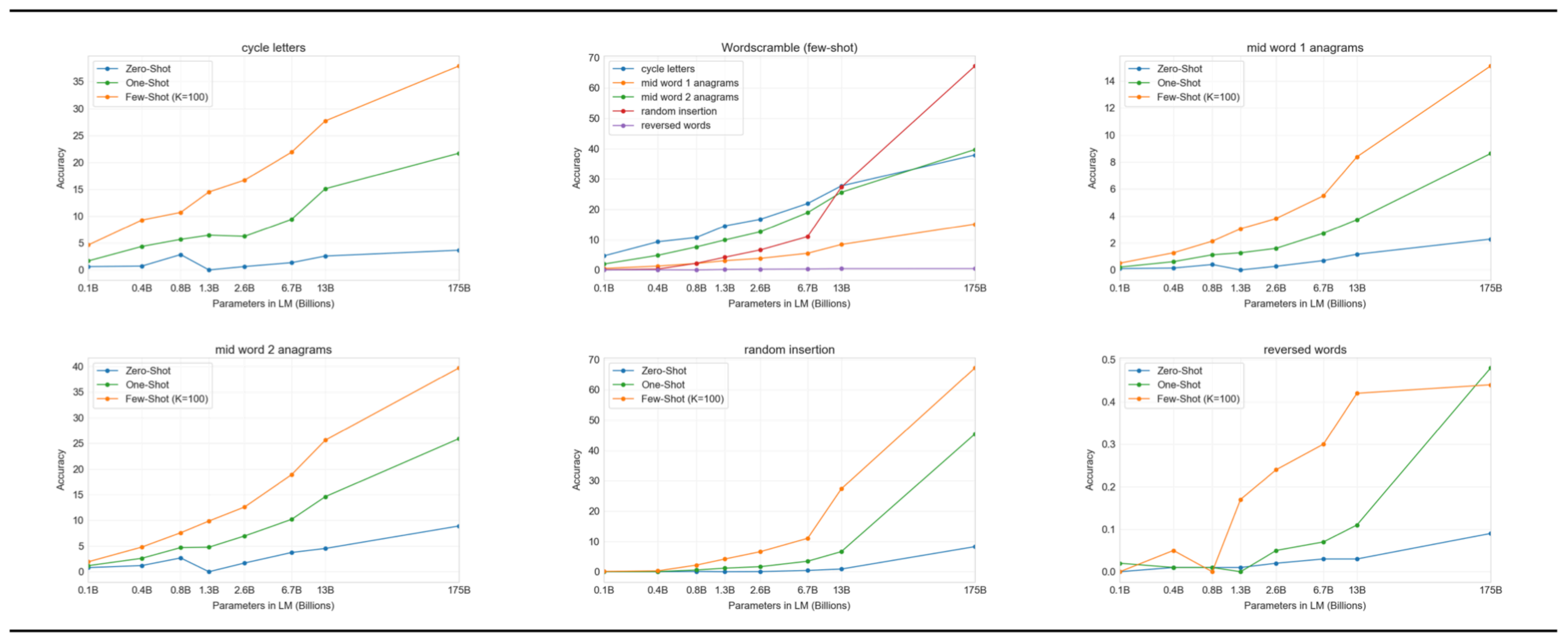
Strengths

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



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

Strengths

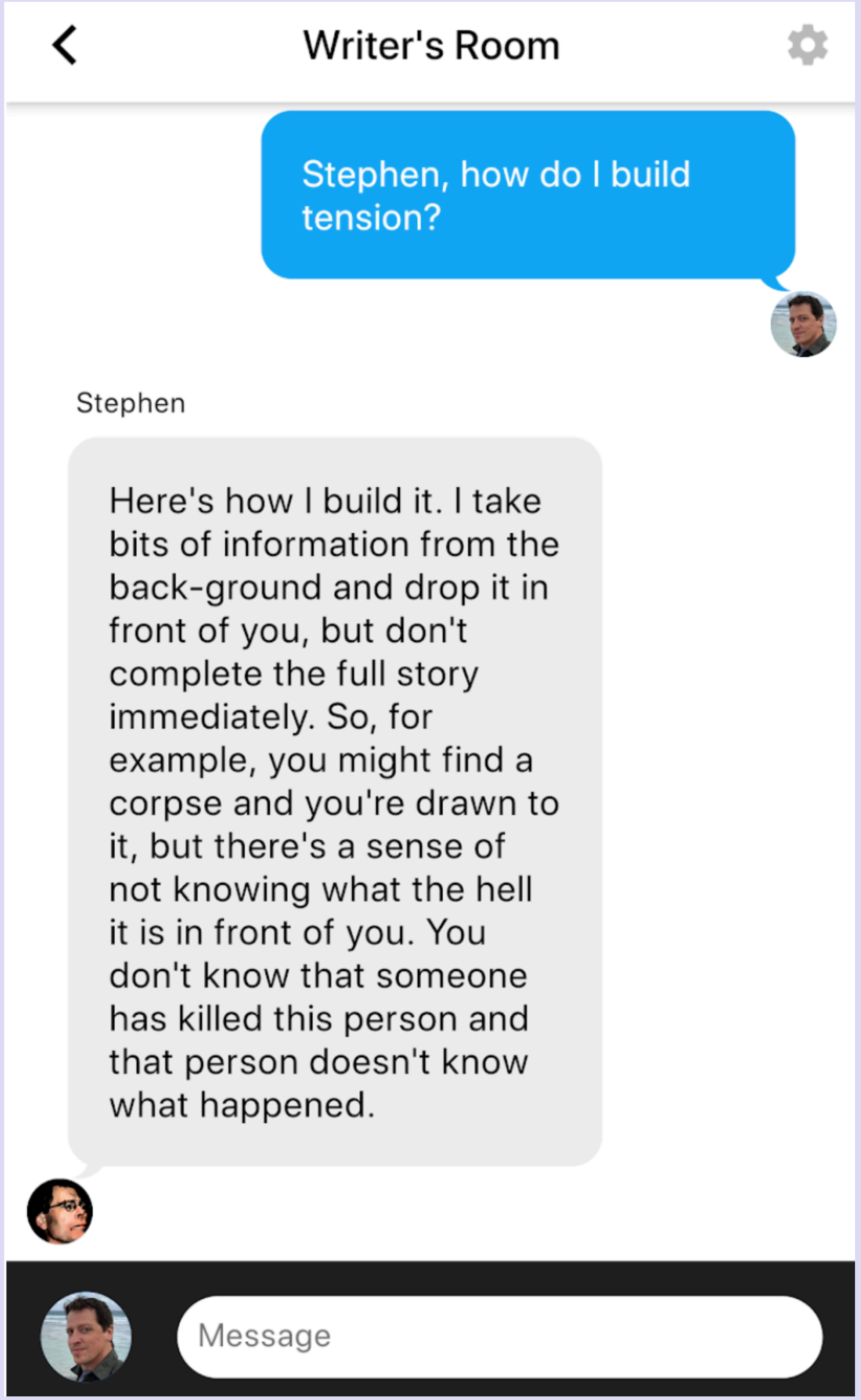


cycle letters
Zero-Shot
One-Shot
Few-Shot (K=100)
Accuracy
Parameters in LM (Billions)
0.1B 0.4B 0.8B 1.3B 2.6B 6.7B 13B 175B
Wordscramble (few-shot)
cycle letters
mid word 1 anagrams
mid word 2 anagrams
random insertion
reversed words
mid word 1 anagrams
mid word 2 anagrams
random insertion
reversed words

Real-World Applications - AI Channels

Andrew Mayne: How do I get great ideas for science fiction stories?

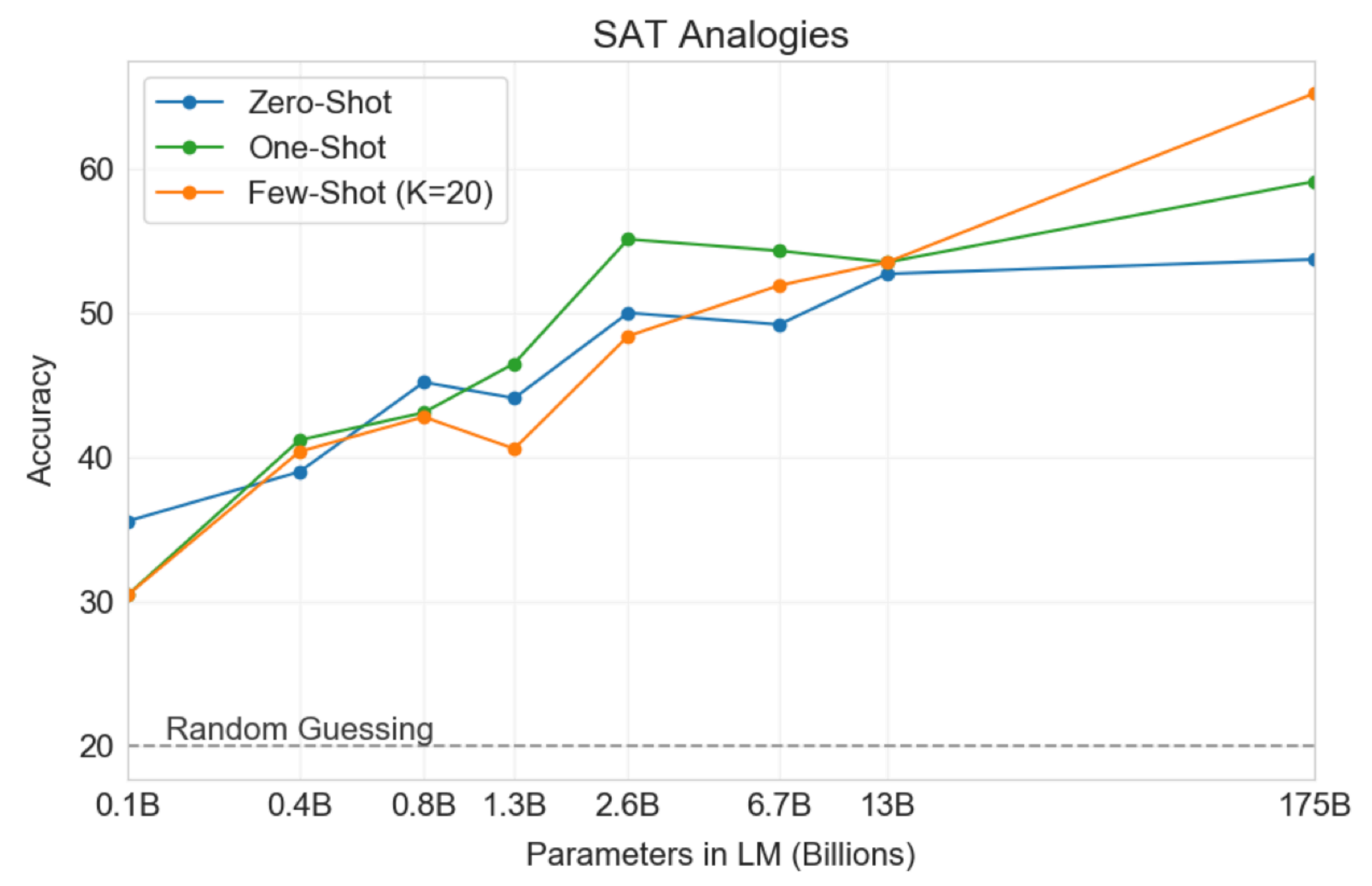
Neil Gaiman: I like to watch alien films, and I like to ask myself - what if? What if you went to Stockholm and you got mugged by Darth Vader?



https://andrewmayneblog.wordpress.com

Strengths

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



Turney, et al. Combining independent modules to solve multiple-choice synonym and analogy problems. 2003