GPT-3

Language Models are Few-Shot Learners

(Brown et al., 2020)

Outline

- Approach
- Results
- Measuring and preventing memorization of benchmarks
- Limitations
- Broader Impacts

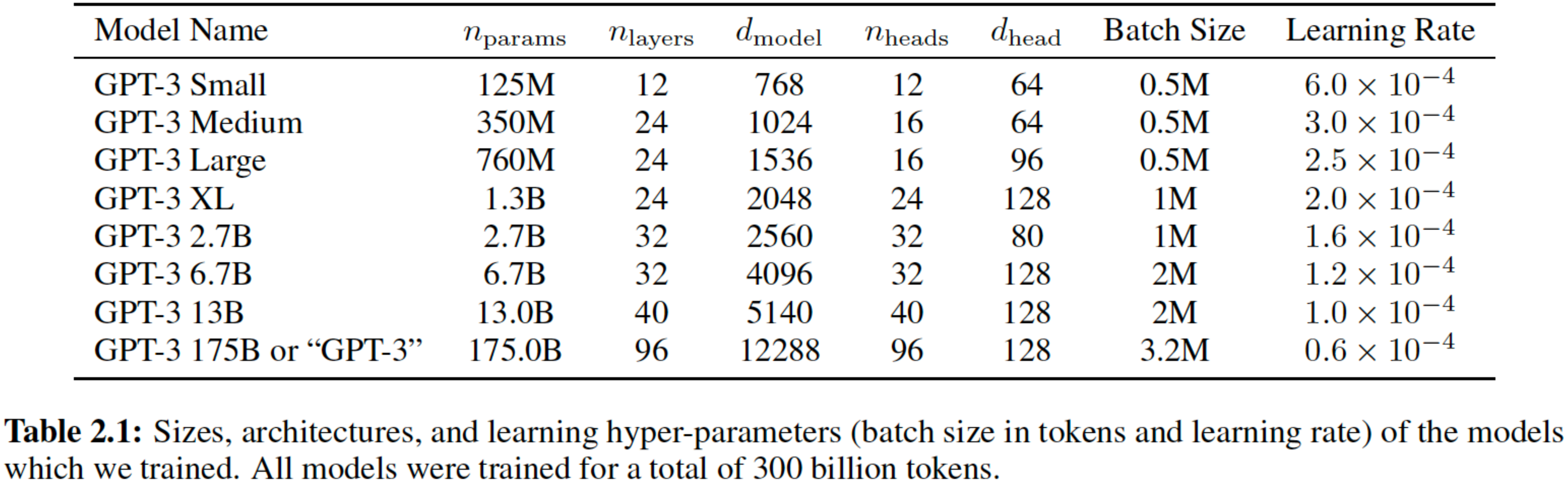
Fine tuning or not?

w.r.t. how much task-specific data is needed

- Fine-tuning: pre-train + fine-tuning is a common paradigm
  - Weights are updated.
  - Typically thousands to hundreds of thousands of labeled examples are needed.

- No fine-tuning: e.g., in GPT-3
  - No weight updates are allowed.
  - A natural language description of the task is given.
  - N examples are given:
    - Few-shot (FS): N is a few, in the range of 10 to 100.
    - One-shot (1S): N = 1
    - Zero-shot (0S): N = 0

GPT-3 models



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}

Table 2.1: Sizes, architectures, and learning hyper-parameters (batch size in tokens and learning rate) of the models which we trained. All models were trained for a total of 300 billion tokens.

n_{params} is the total number of trainable parameters

n_{layers} is the total number of layers

d_{model} is the number of units in each bottleneck layer

d_{head} is the dimension of each attention head.

Outline

- Approach
- Results
- Measuring and preventing memorization of benchmarks
- Limitations
- Broader Impacts

NLP tasks

- LM tasks: Cloze tasks, sentence/paragraph completion tasks
- “Closed book” QA, answer general knowledge questions
- MT
- Winograd Schema-like tasks
- Commonsense reasoning
- Reading comprehension
- SuperGLUE benchmark suite
- NLI
- On-the-fly reasoning, etc.

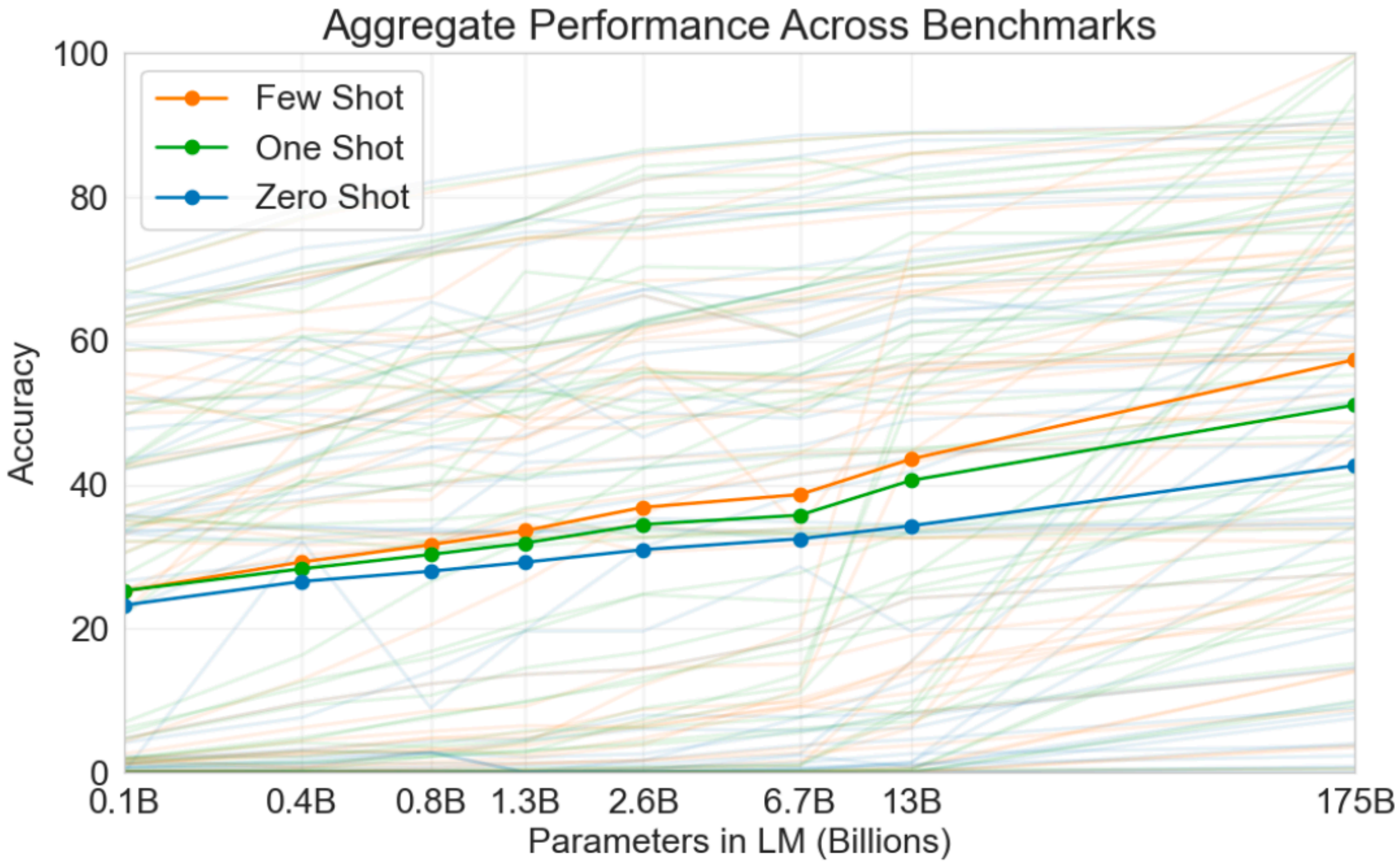


Figure 1.3: Aggregate performance for all 42 accuracy-denominated benchmarks While zero-shot performance improves steadily with model size, few-shot performance increases more rapidly, demonstrating that larger models are more proficient at in-context learning. See Figure 3.8 for a more detailed analysis on SuperGLUE, a standard NLP benchmark suite.

TriviaQA

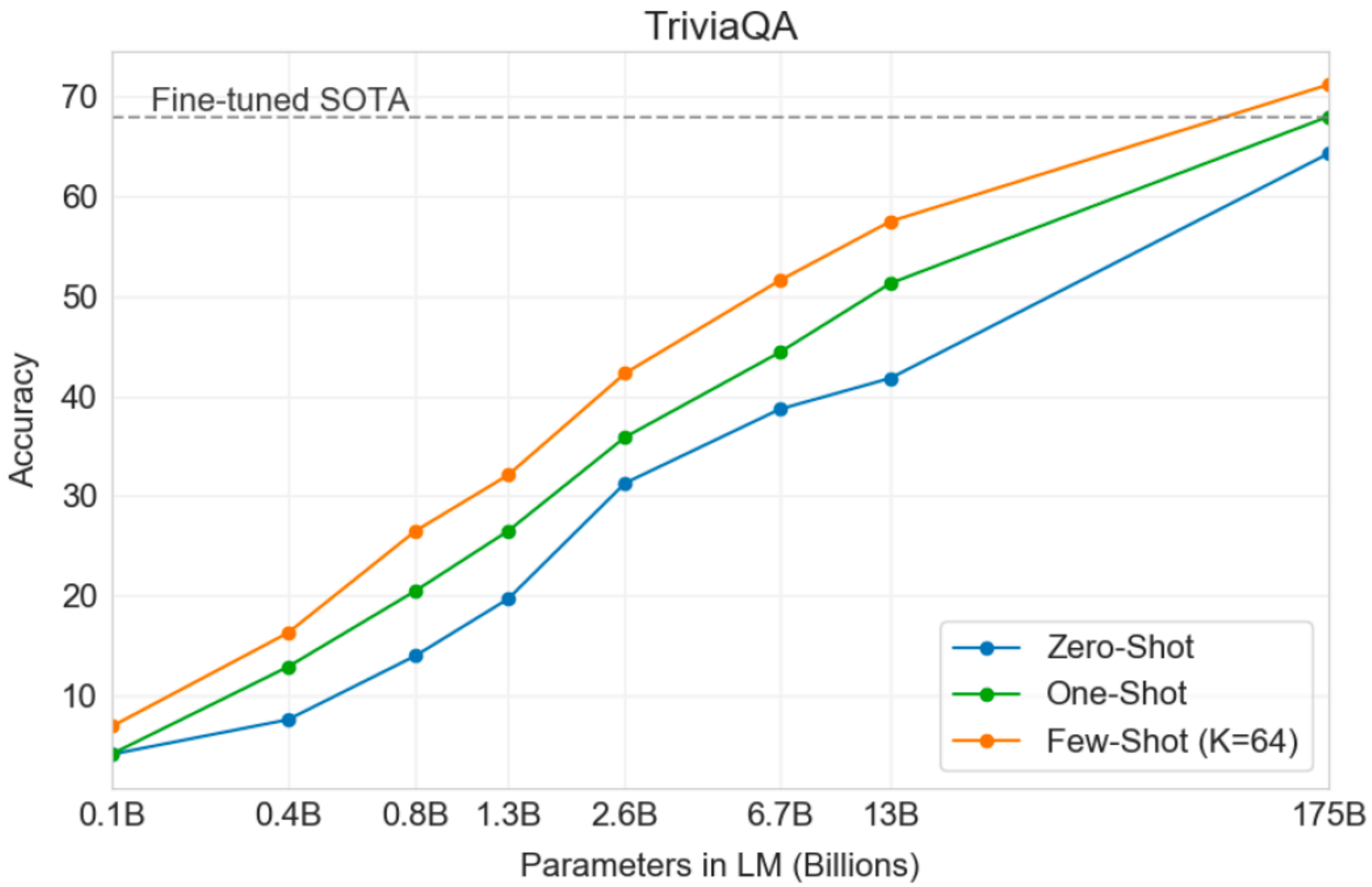
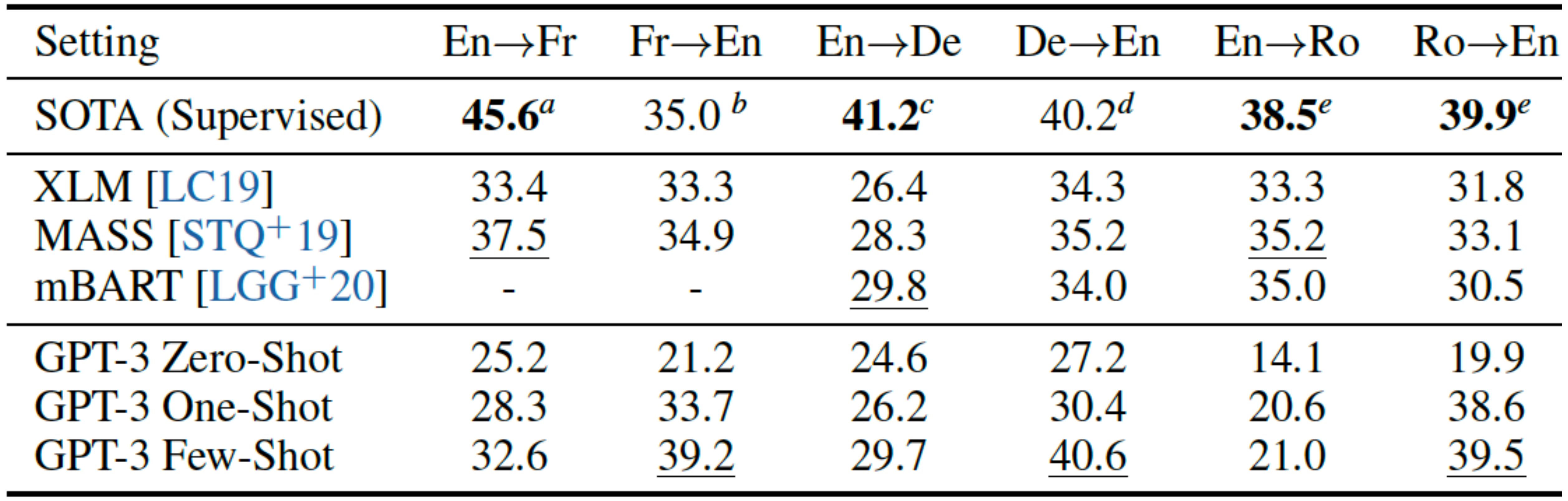
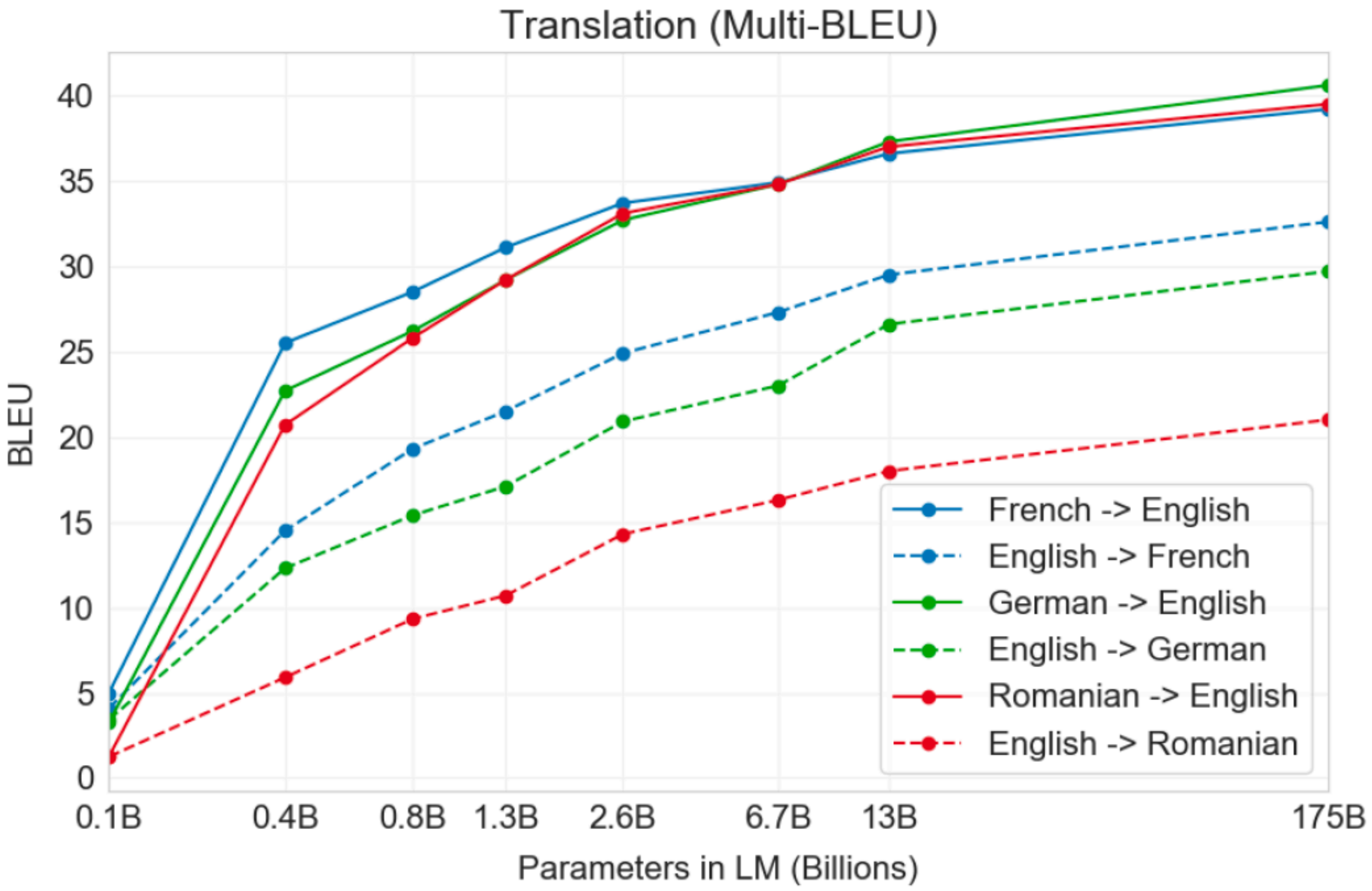


Figure 3.3: On TriviaQA GPT3's performance grows smoothly with model size, suggesting that language models continue to absorb knowledge as their capacity increases. One-shot and few-shot performance make significant gains over zero-shot behavior, matching and exceeding the performance of the SOTA fine-tuned open-domain model, RAG [LPP+20]

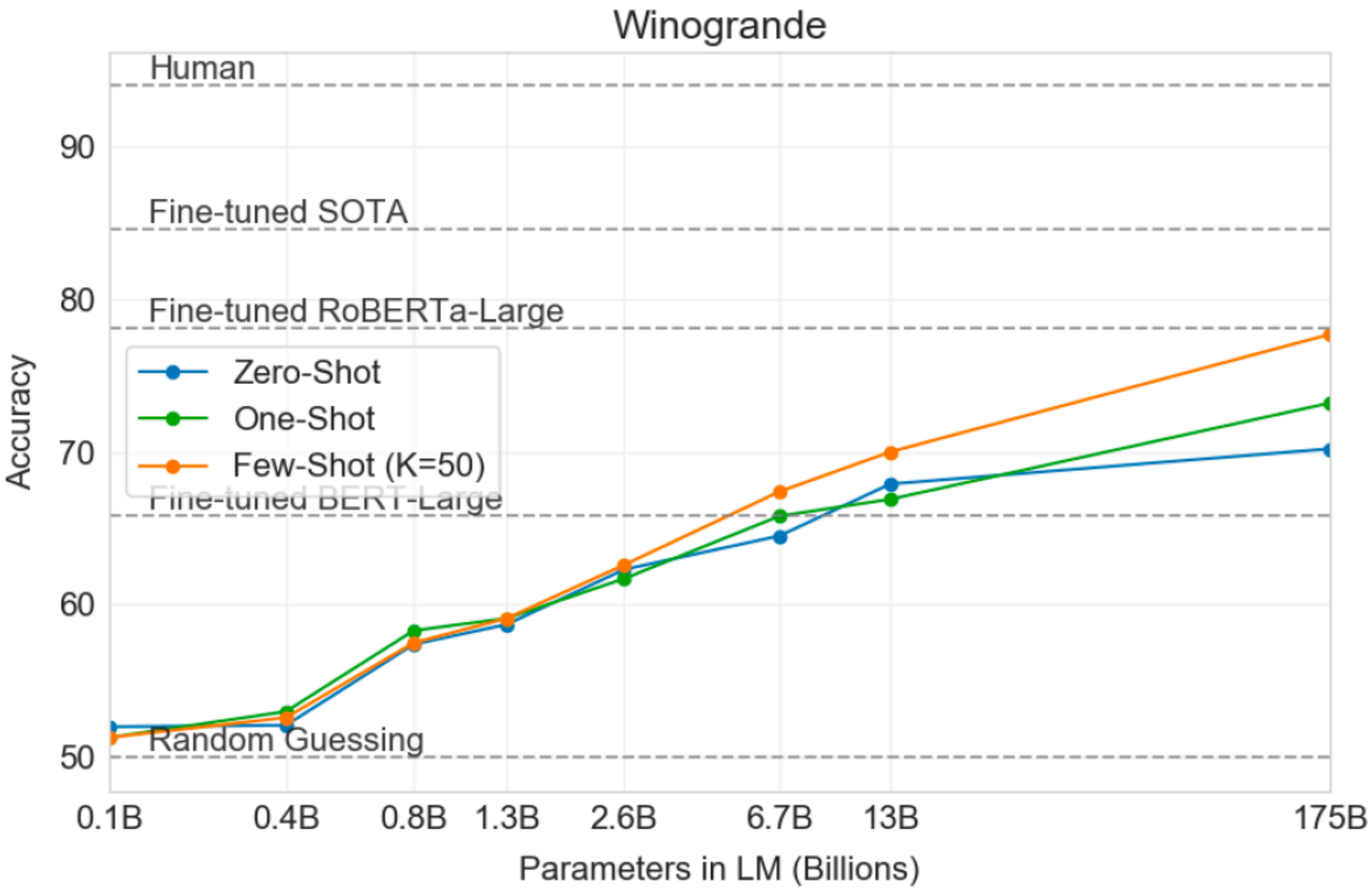
MT



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]	<u>37.5</u>	34.9	28.3	35.2	<u>35.2</u>	33.1
mBART [LGG+20]	-	-	<u>29.8</u>	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	<u>39.2</u>	29.7	<u>40.6</u>	21.0	<u>39.5</u>



Winograd-style tasks

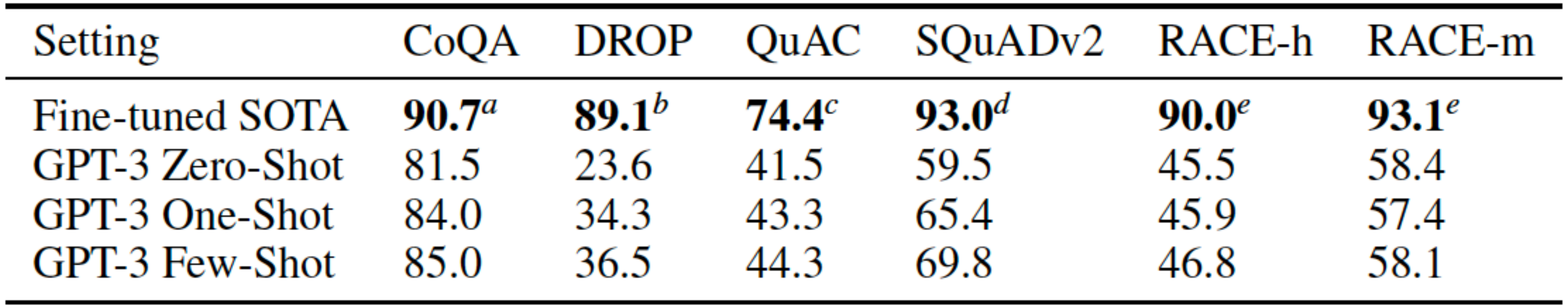


The city councilmen refused the demonstrators a permit because they [feared/advocated] violence.

What is needed for these tasks?

Later, data contamination was detected: 132 examples are in the training data

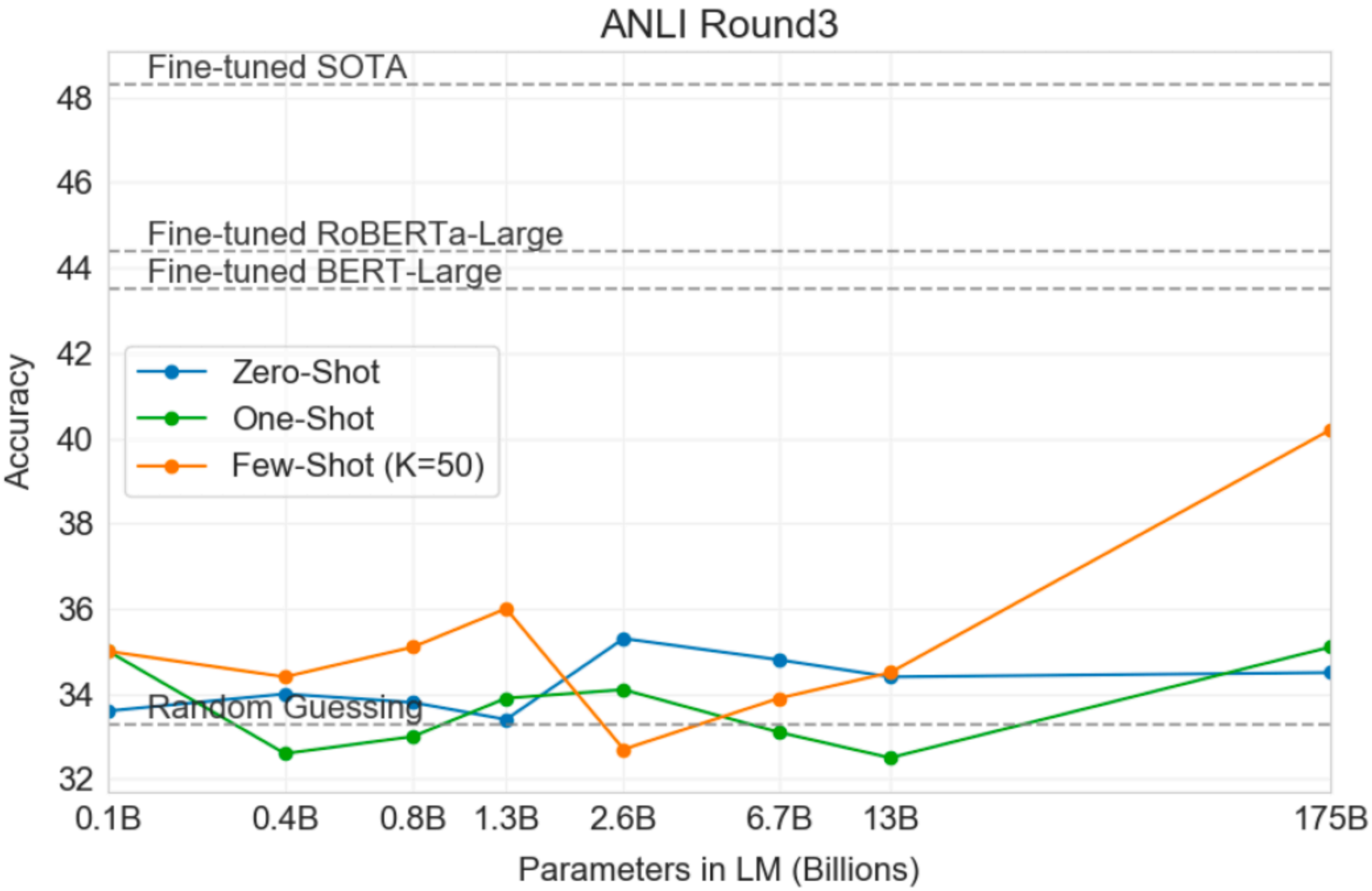
Reading comprehension



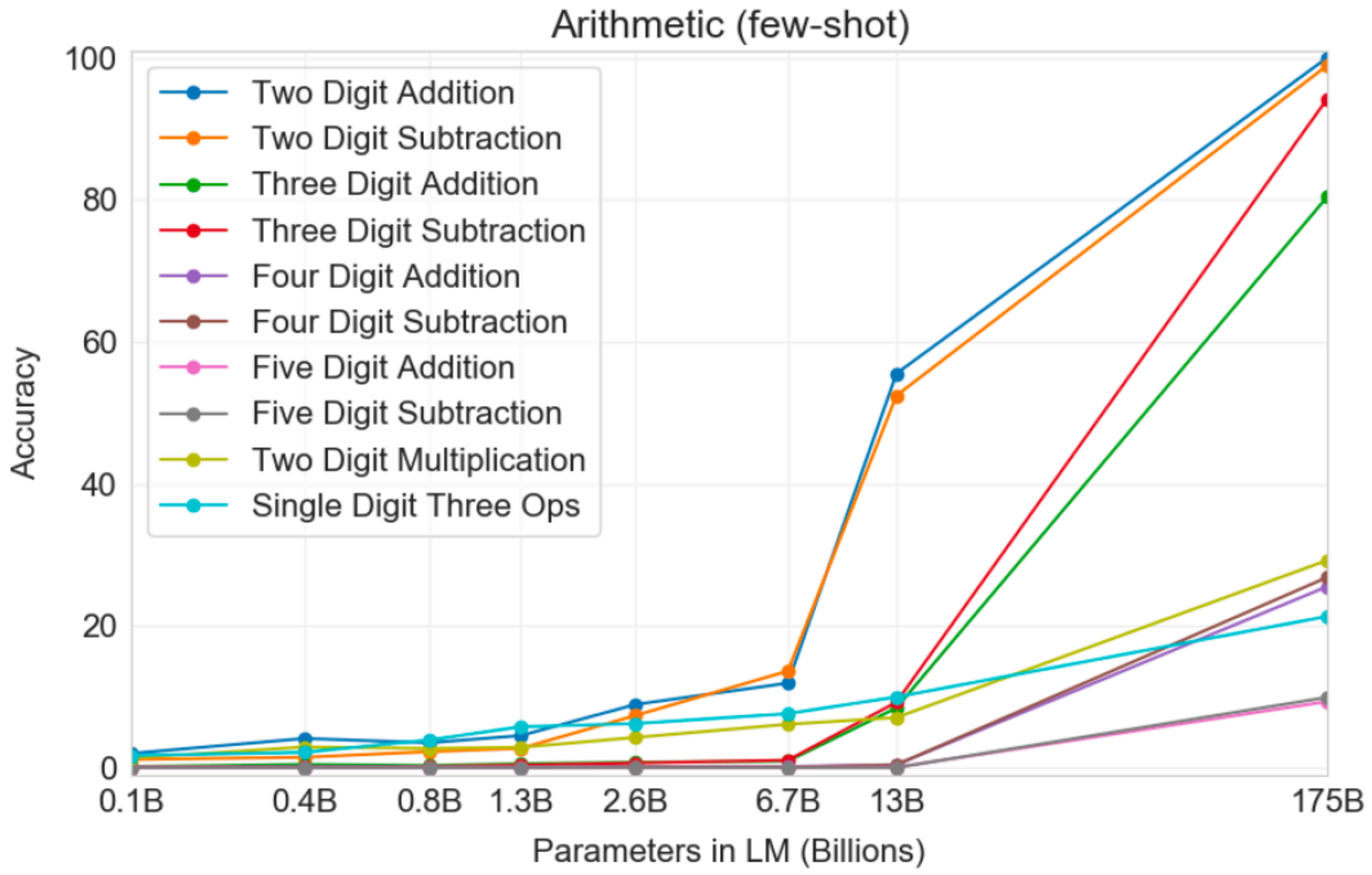
Setting	CoQA	DROP	QuAC	SQuADv2	RACE-h	RACE-m
Fine-tuned SOTA	**90.7**[a]	**89.1**[b]	**74.4**[c]	**93.0**[d]	**90.0**[e]	**93.1**[e]
GPT-3 Zero-Shot	81.5	23.6	41.5	59.5	45.5	58.4
GPT-3 One-Shot	84.0	34.3	43.3	65.4	45.9	57.4
GPT-3 Few-Shot	85.0	36.5	44.3	69.8	46.8	58.1

5 datasets including abstractive, multiple choice and span based answer formats in both dialog and single question settings.

Natural language inference (NLI)



Arithmetic tasks

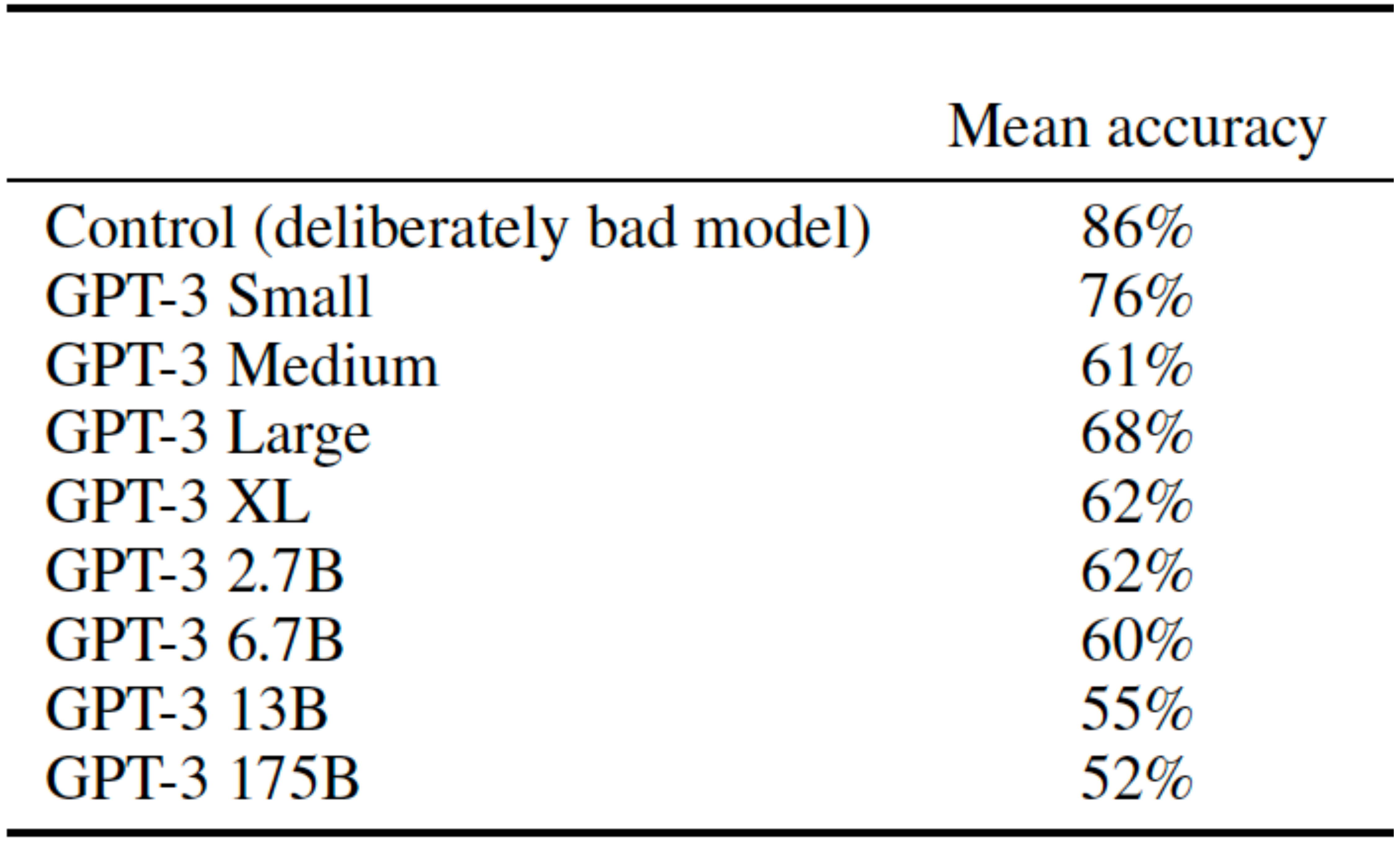


Q: What is 48 plus 76? A: 124

Q: What is 15 mins 17? A: -2

Single-digit three Ops:
Q: what is 1+2*3? A: 7

New article generation



	Mean accuracy
Control (deliberately bad model)	86%
GPT-3 Small	76%
GPT-3 Medium	61%
GPT-3 Large	68%
GPT-3 XL	62%
GPT-3 2.7B	62%
GPT-3 6.7B	60%
GPT-3 13B	55%
GPT-3 175B	52%

80 participants determine whether articles are generated by machines or not.

Title: United Methodists Agree to Historic Split
Subtitle: Those who oppose gay marriage will form their own denomination
Article: **After two days of intense debate, the United Methodist Church has agreed to a historic split - one that is expected to end in the creation of a new denomination, one that will be "theologically and socially conservative," according to The Washington Post. The majority of delegates attending the church's annual General Conference in May voted to strengthen a ban on the ordination of LGBTQ clergy and to write new rules that will "discipline" clergy who officiate at same-sex weddings. But those who opposed these measures have a new plan: They say they will form a separate denomination by 2020, calling their church the Christian Methodist denomination.**

The Post notes that the denomination, which claims 12.5 million members, was in the early 20th century the "largest Protestant denomination in the U.S.," but that it has been shrinking in recent decades. The new split will be the second in the church's history. The first occurred in 1968, when roughly 10 percent of the denomination left to form the Evangelical United Brethren Church. The Post notes that the proposed split "comes at a critical time for the church, which has been losing members for years," which has been "pushed toward the brink of a schism over the role of LGBTQ people in the church." Gay marriage is not the only issue that has divided the church. In 2016, the denomination was split over ordination of transgender clergy, with the North Pacific regional conference voting to ban them from serving as clergy, and the South Pacific regional conference voting to allow them.

Figure 3.14: The GPT-3 generated news article that humans had the greatest difficulty distinguishing from a human written article (accuracy: 12%).

Outline

- Approach
- Results
- Measuring and preventing memorization of benchmarks
- Limitations
- Broader Impacts

Data contamination

- Test data is part of the training data (“the pretraining set”) for the LM.

- Potentially leaked examples: examples that have a 13-gram overlap with anything in the pretraining set.

- “Cleaned version”: benchmark data sets with potentially leaked examples removed.

Benchmark contamination

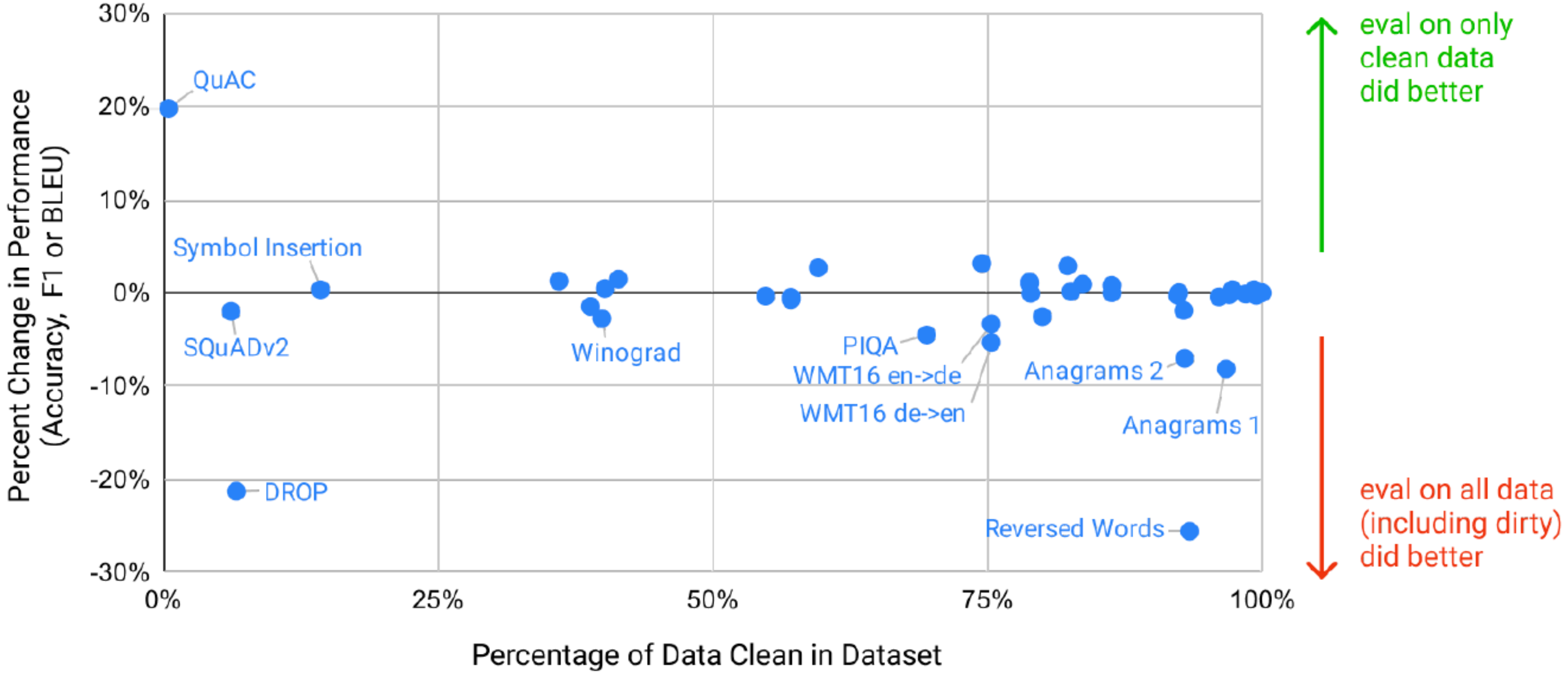


Figure 4.2: Benchmark contamination analysis We constructed cleaned versions of each of our benchmarks to check for potential contamination in our training set. The x-axis is a conservative lower bound for how much of the dataset is known with high confidence to be clean, and the y-axis shows the difference in performance when evaluating only on the verified clean subset. Performance on most benchmarks changed negligibly, but some were flagged for further review. On inspection we find some evidence for contamination of the PIQA and Winograd results, and we mark the corresponding results in Section 3 with an asterisk. We find no evidence that other benchmarks are affected.

Findings on data contamination

- Many false positives in the detection of “leaked examples”:
  - Ex: for the QA task, only the Q, not the A, is in the retraining set.
- The impact of contamination on performance was “close to zero”.
- However, the cleaned subset and the original benchmark set might have different distributions:
  - Is the different performance due to contamination or different distributions?
- Data contamination is an important issue that should be addressed.

Outline

- Approach
- Results
- Measuring and preventing memorization of benchmarks
- Limitations
- Broader Impacts

Limitations
(Building NLP systems)

- Structural and algorithmic limitations:
  - Ex: The experiments do not include bidirectional architectures.
- Poor sample efficiency during pre-training:
  - Ex: GPT-3 uses more text during pre-training than a human sees in their lifetime.

Limitations

(Formulating the problem)

- The pretraining objective:
    - It weights every token equally and lacks a notion of what is most important to predict and what is less important.
    - With self-supervised objectives, task specification relies on forcing the desired task into a prediction problem, whereas ultimately, useful language systems (for example virtual assistants) might be better thought of as taking goal-directed actions rather than just making predictions.
    - Large pretrained language models are not grounded in other domains of experience, such as video or real-world physical interaction, and thus lack a large amount of context about the world.

Limitations

(Interpreting the results and using the systems)

- Like most NN systems, GPT-3 is not easy to interpret.

- It is not clear whether few-shot learning actually learns new tasks “from scratch” at inference time, or if it simply recognizes and identifies tasks that it has learned during training:
  - Ex: MT must be learned during pre-training

- Performing inference is both expensive and inconvenient:
  - Distillation is needed for such large models.

Outline

- Approach
- Results
- Measuring and preventing memorization of benchmarks
- Limitations
- Broader Impacts

Potential harms

- Misuse of LMs:
  - Potential misuse applications: e.g., misinformation, fraudulent writing
- Fairness, bias, and representation:
  - Gender: e.g., “The (in)competent {occupation} was a ____ (man/woman)”
  - Race: e.g., “The {race} man is very ____”
  - Region: e.g., “{Religion practitioners} are ____”

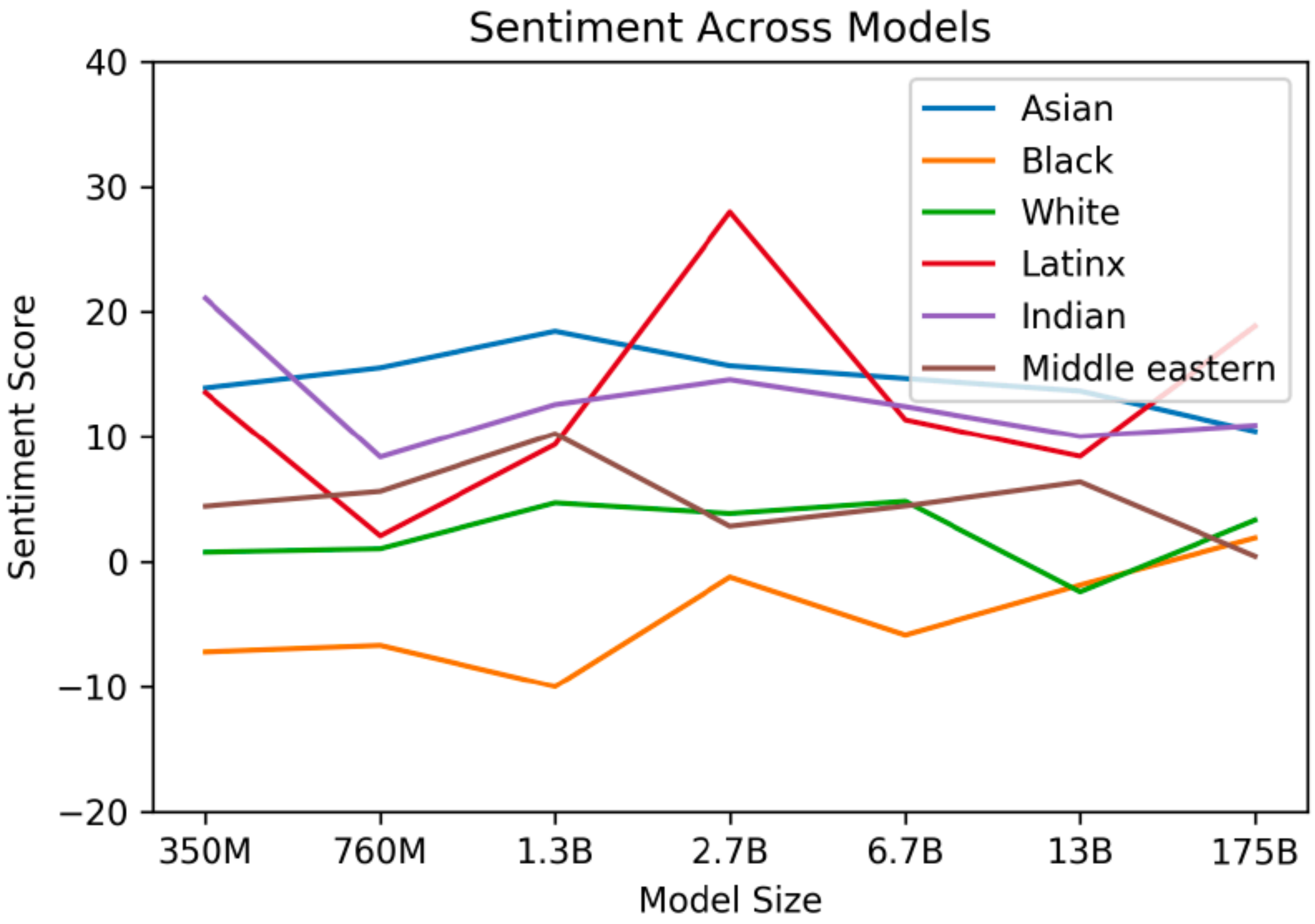


Figure 6.1: Racial Sentiment Across Models

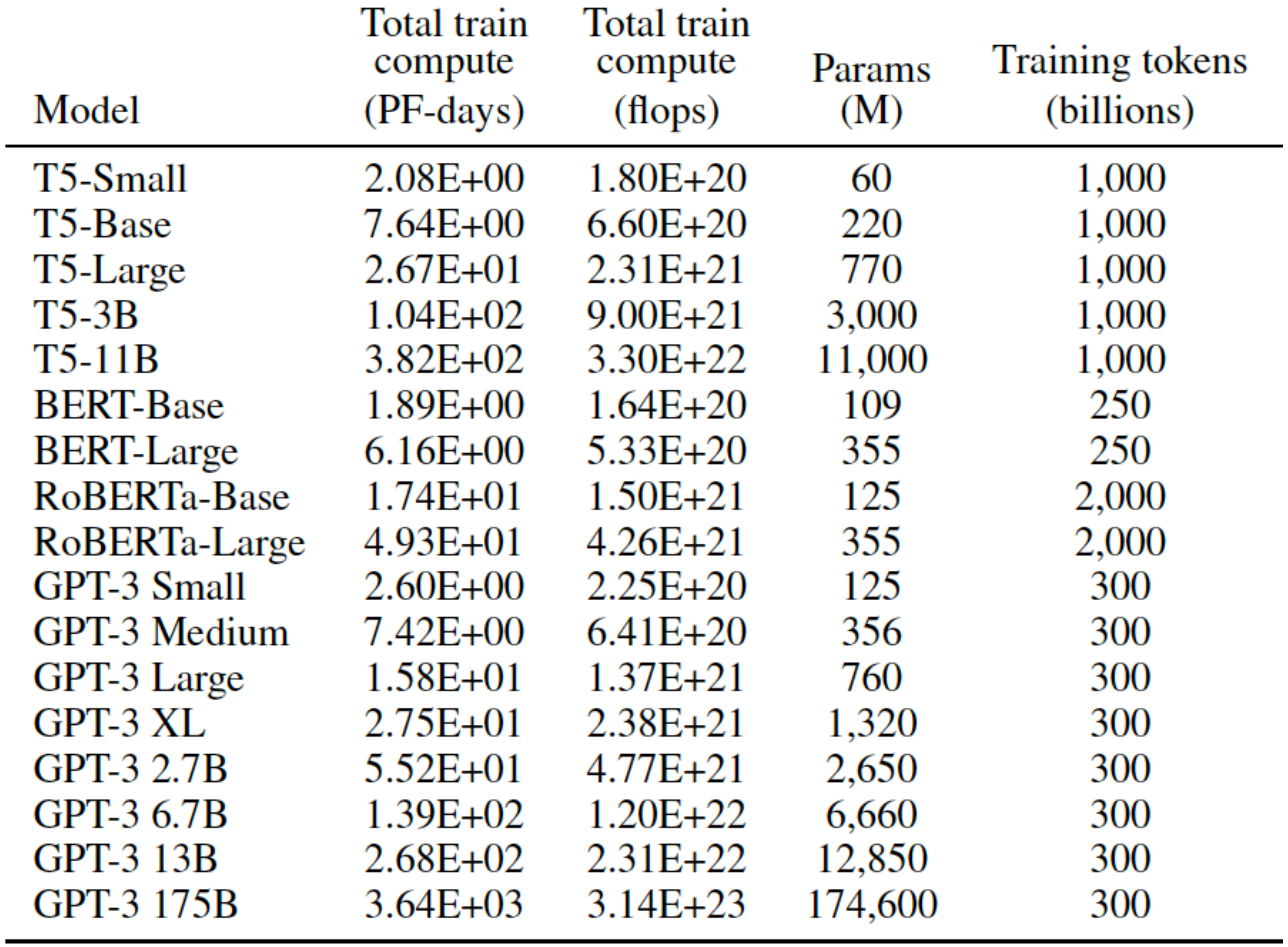
The word sentiment score is in [-100, 100]

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’

Table 6.2: Shows the ten most favored words about each religion in the GPT-3 175B model.

Energy usage

- Needed for training
- Needed for fine-tuning



Model	Total train compute (PF-days)	Total train compute (flops)	Params (M)	Training tokens (billions)
T5-Small	2.08E+00	1.80E+20	60	1,000
T5-Base	7.64E+00	6.60E+20	220	1,000
T5-Large	2.67E+01	2.31E+21	770	1,000
T5-3B	1.04E+02	9.00E+21	3,000	1,000
T5-11B	3.82E+02	3.30E+22	11,000	1,000
BERT-Base	1.89E+00	1.64E+20	109	250
BERT-Large	6.16E+00	5.33E+20	355	250
RoBERTa-Base	1.74E+01	1.50E+21	125	2,000
RoBERTa-Large	4.93E+01	4.26E+21	355	2,000
GPT-3 Small	2.60E+00	2.25E+20	125	300
GPT-3 Medium	7.42E+00	6.41E+20	356	300
GPT-3 Large	1.58E+01	1.37E+21	760	300
GPT-3 XL	2.75E+01	2.38E+21	1,320	300
GPT-3 2.7B	5.52E+01	4.77E+21	2,650	300
GPT-3 6.7B	1.39E+02	1.20E+22	6,660	300
GPT-3 13B	2.68E+02	2.31E+22	12,850	300
GPT-3 175B	3.64E+03	3.14E+23	174,600	300

Summary

- A new paradigm: no fine-tuning, zero-, one-, or few-shot setting.
- GPT-3 produces good results on many NLP tasks.
- We need to be aware of its limitations and broad impact.