Fall 2024 COS597R:
Deep Dive into Large Language Models

Danqi Chen, Sanjeev Arora

PLi

PRINCETON UNIVERSITY

Lecture 2: Pre-training I

https://princeton-cos597r.github.io/

Announcements

- TAs: Adithya Bhaskar (adithyab@), Tyler Zhu (tylerzhu@)

- Please make sure that you are on the Slack team (princeton-cos597r) asap!
  - We will use Slack for future announcements, and discussion of lectures (#lectures), research topics, and projects!

- All the panel and scribes assignments are on the website

- Office hours: Danqi (Tue 10-11; appointment-based), Sanjeev (TBA), Adithya (Wed 3-4), Tyler (Mon 4-5)

Focus: the GPT-3 paper

arXiv
https://arxiv.org › cs

[2005.14165] Language Models are Few-Shot Learners

by TB Brown · 2020 · Cited by 31178 — Specifically, we train **GPT-3**, an autoregressive language model with 175 billion parameters, 10x more than any previous non-sparse languag…

[Submitted on 28 May 2020 (v1), last revised 22 Jul 2020 (this version, v4)]

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

Recent work has demonstrated substantial gains on many NLP tasks and benchmarks by pre-training on a large corpus of text followed by fine-tuning on a specific task. While typically task-agnostic in architecture, this method still requires task-specific fine-tuning datasets of thousands or tens of thousands of examples. By contrast, humans can generally perform a new language task from only a few examples or from simple instructions - something which current NLP systems still largely struggle to do. Here we show that scaling up language models greatly improves task-agnostic, few-shot performance, sometimes even reaching competitiveness with prior state-of-the-art fine-tuning approaches. Specifically, we train GPT-3, an autoregressive language model with 175 billion parameters, 10x more than any previous non-sparse language model, and test its performance in the few-shot setting. For all tasks, GPT-3 is applied without any gradient updates or fine-tuning, with tasks and few-shot demonstrations specified purely via text interaction with the model. GPT-3 achieves strong performance on many NLP datasets, including translation, question-answering, and cloze tasks, as well as several tasks that require on-the-fly reasoning or domain adaptation, such as unscrambling words, using a novel word in a sentence, or performing 3-digit arithmetic. At the same time, we also identify some datasets where GPT-3's few-shot learning still struggles, as well as some datasets where GPT-3 faces methodological issues related to training on large web corpora. Finally, we find that GPT-3 can generate samples of news articles which human evaluators have difficulty distinguishing from articles written by humans. We discuss broader societal impacts of this finding and of GPT-3 in general.

"Language models form the backbone of modern techniques for solving a range of problems in natural language processing. The paper shows that when such language models are scaled up to an unprecedented number of parameters, **the language model itself can be used as a few-shot learner that achieves very competitive performance on many of these problems without any additional training**. This is a very surprising result that is expected to have substantial impact in the field, and that is likely to withstand the test of time. In addition to the scientific contribution of the work, the paper also **presents a very extensive and thoughtful exposition of the broader impact of the work**, which may serve as an example to the NeurIPS community on how to think about the real-world impact of the research performed by the community."

https://neuripsconf.medium.com/announcing-the-neurips-2020-award-recipients-73e4d3101537

Before I dive in

- I assume you have read the paper carefully
- My goal is to provide additional context + highlight important points of the paper
- Questions and discussion are welcome anytime

Q3 (optional): Do you have any questions from the reading that you would like to see addressed in class?

27 responses

I'm wondering how evaluation methods have changed since LLMs like ChatGPT went public and if increased usage has changed the reasoning presented in the paper or if the evaluation in this paper should be re-evaluated.

I'd like to discuss about how to correctly evaluate the MCQ (multiple-choice questions) tasks.

In the benchmarks, when it lists few-shot learning comparing with SOTA, does that mean using context that includes all the examples that few-shot learning SOTA is trained with?

Brief history and motivation

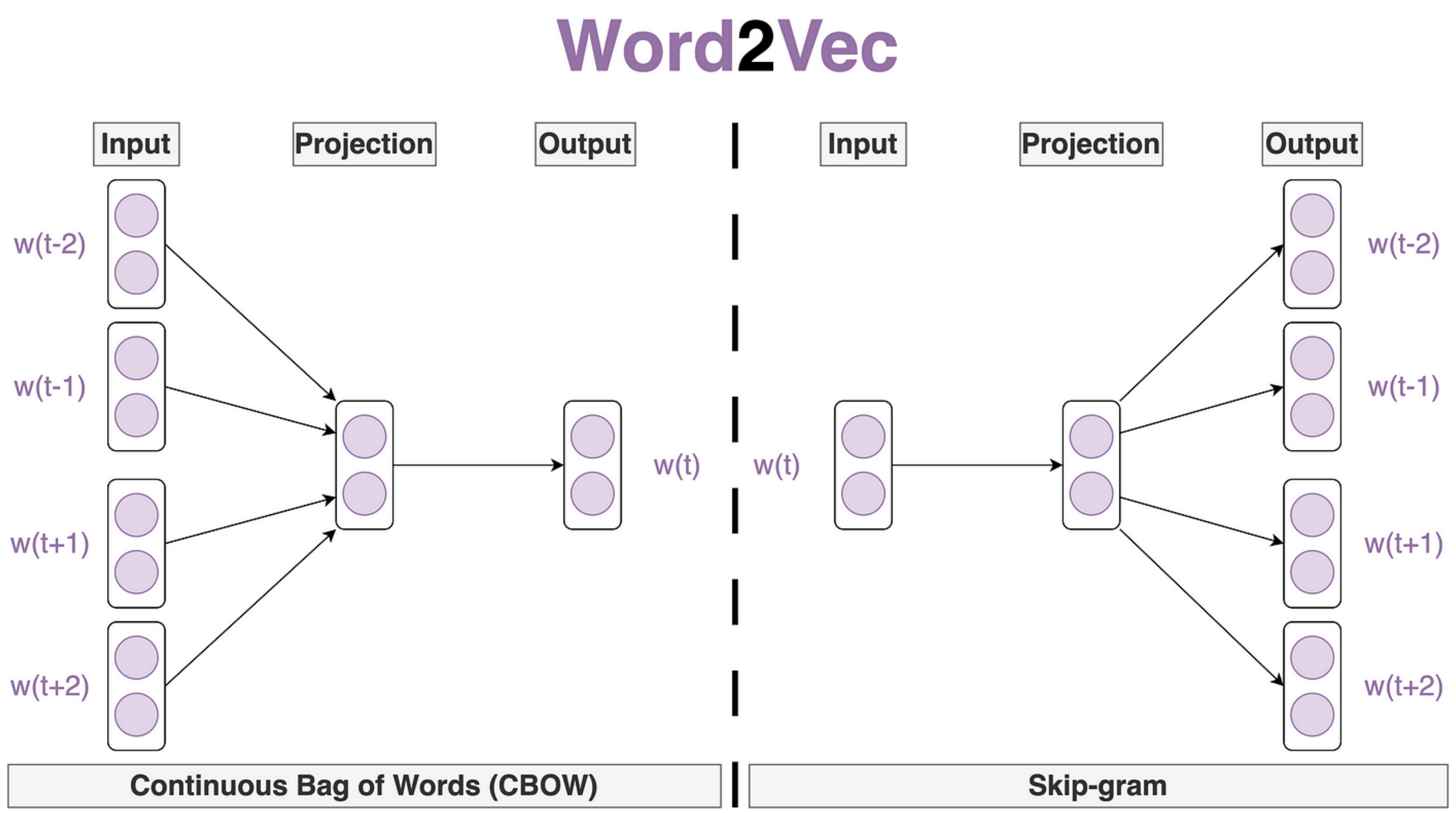
Word embeddings

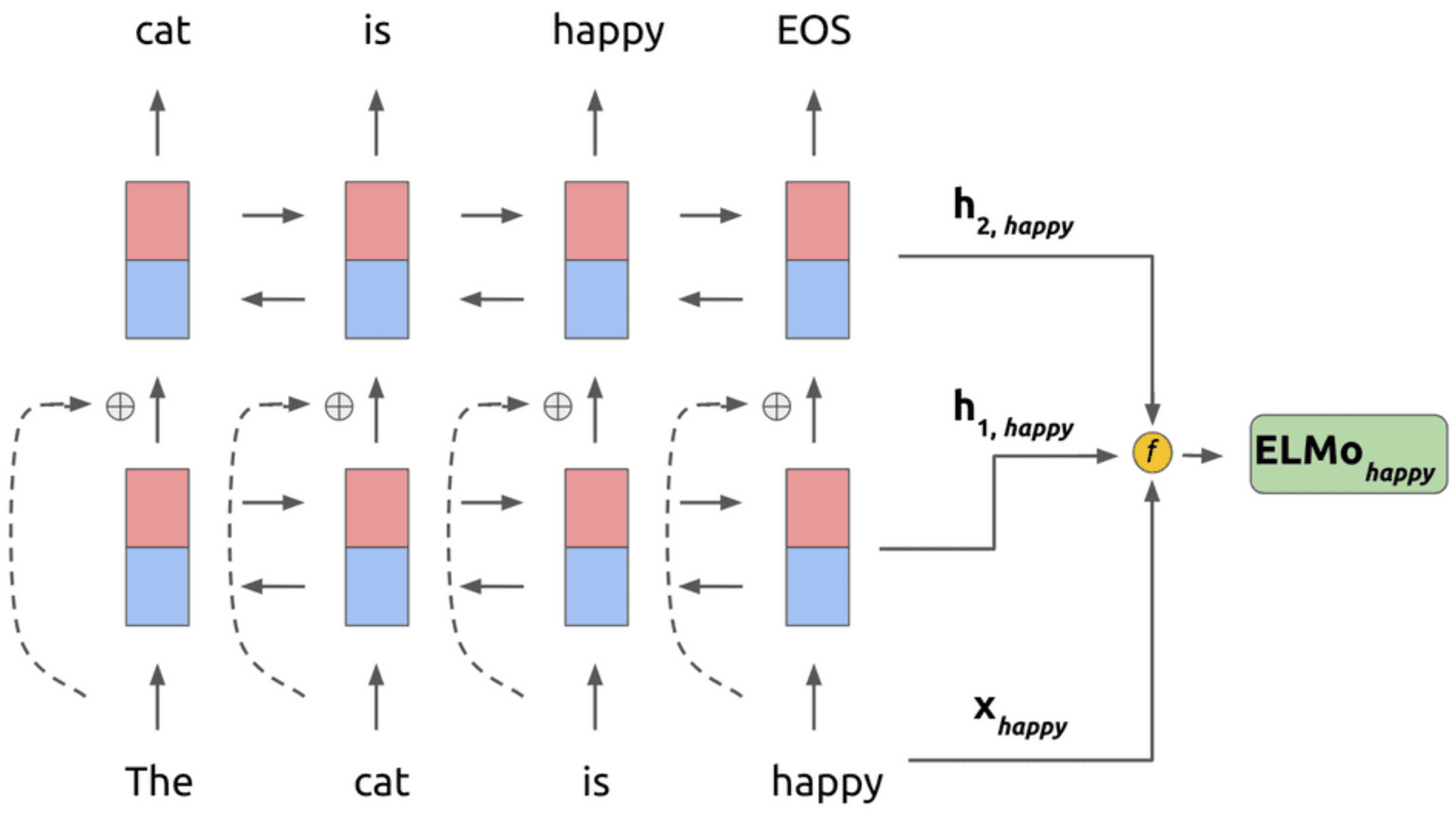
- **Word embeddings** e.g., word2vec (Mikolov et al., 2013), GloVe (Pennington et al., 2014)

  "single-layer representations were learned using word vectors"

- **Contextualized word embeddings** e.g., ELMo (Peters et al., 2018), CoVe (McCann et al., 2017)

  "RNNs with multiple layers of representations and contextual state were used to form stronger representations"



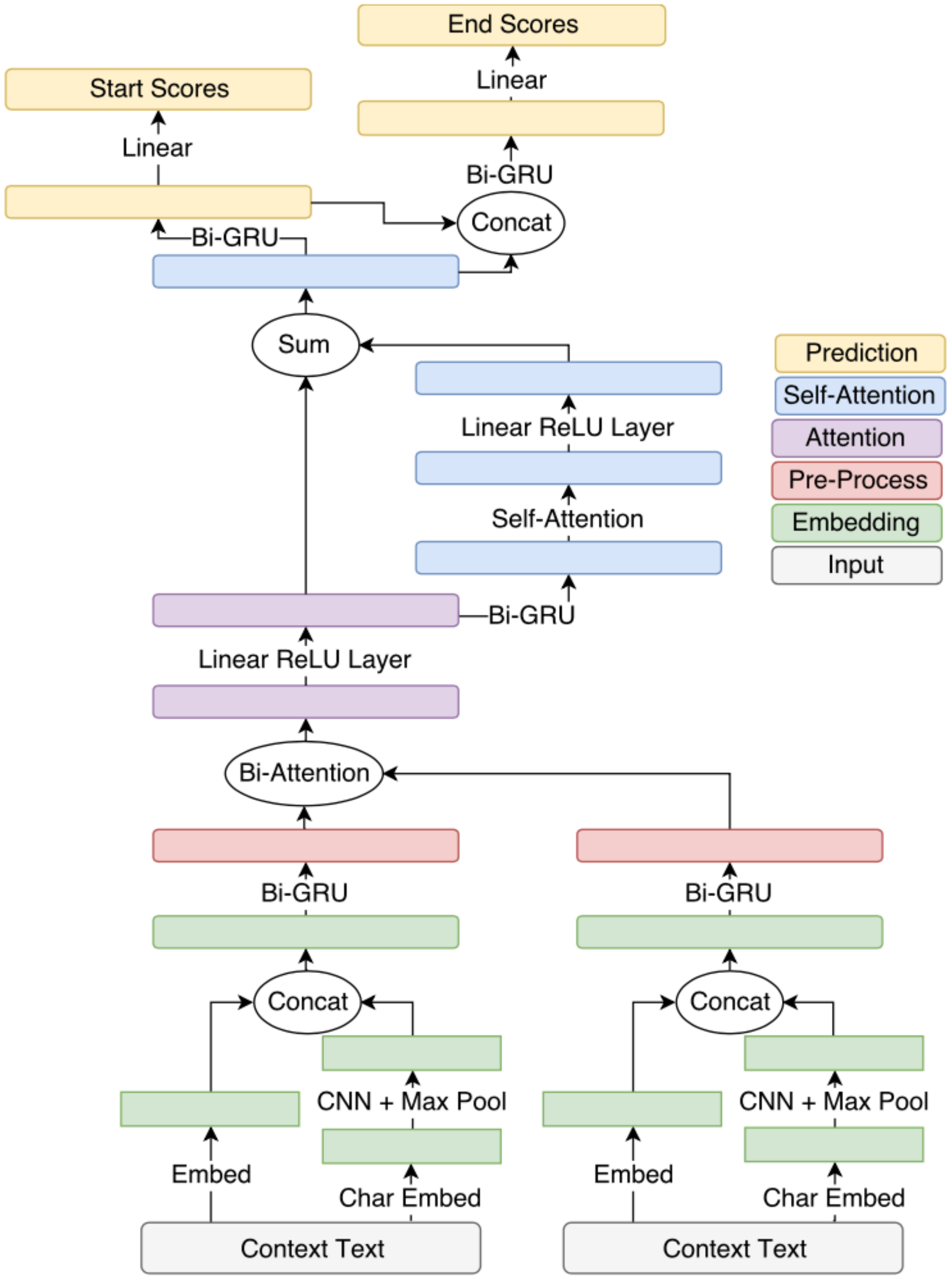


Used for task-specific neural architectures!

Word embeddings

- **Word embeddings**
- **Contextualized word embeddings**

Used for task-specific neural architectures!



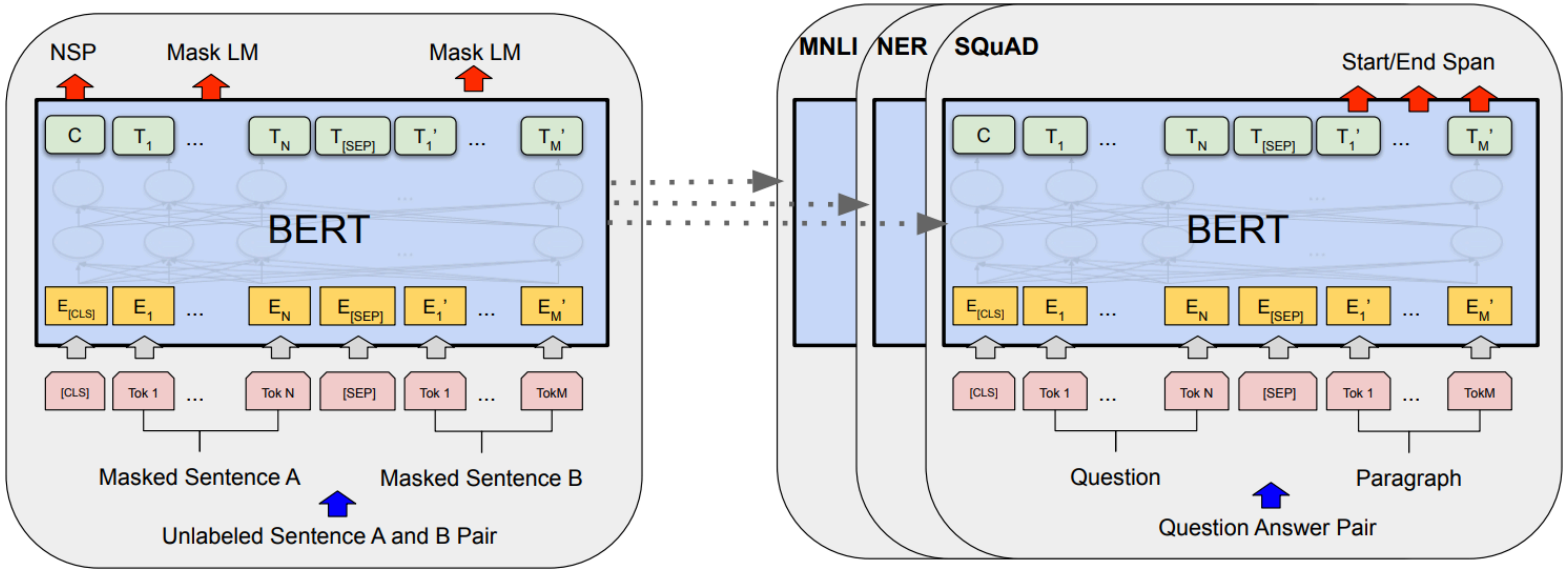
(Clark and Gardner, 2018)

One model for all tasks

- **One pre-trained model for all tasks**
  - BERT (Devlin et al., 2018), RoBERTa (Liu et al., 2019)
  - T5 (Raffel et al., 2019), BART (Lewis et al., 2019)
  - GPT-1 (Radford et al., 2018), GPT-2 (Radford et al., 2019)

minimal modifications to downstream tasks

still fine-tuning on $10^3 - 10^5$ downstream examples



Pre-training　　Fine-Tuning

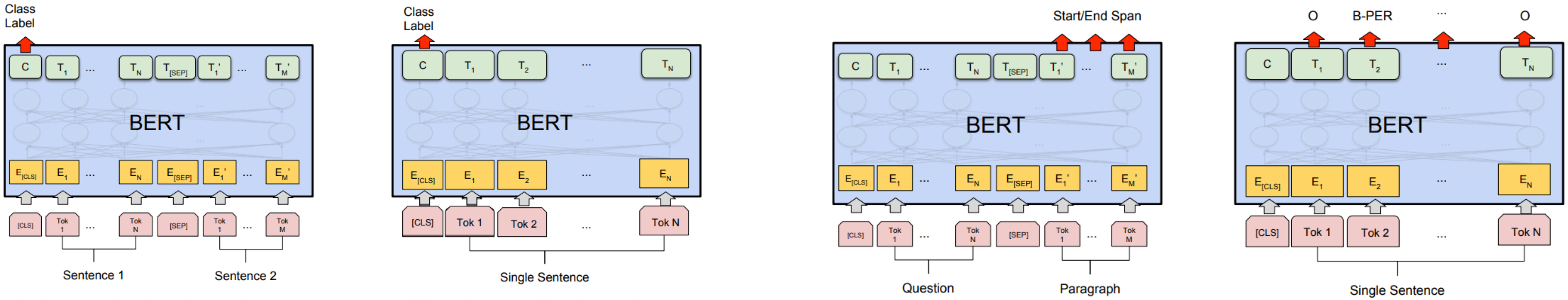
(Devlin et al. , 2018)

One model for all tasks

- **One pre-trained model for all tasks**

  minimal modifications to downstream tasks

  still fine-tuning on $10^3 - 10^5$ downstream examples



(a) Sentence Pair Classification Tasks: MNLI, QQP, QNLI, STS-B, MRPC, RTE, SWAG

(b) Single Sentence Classification Tasks: SST-2, CoLA

(c) Question Answering Tasks:

(d) Single Sentence Tagging Tasks:

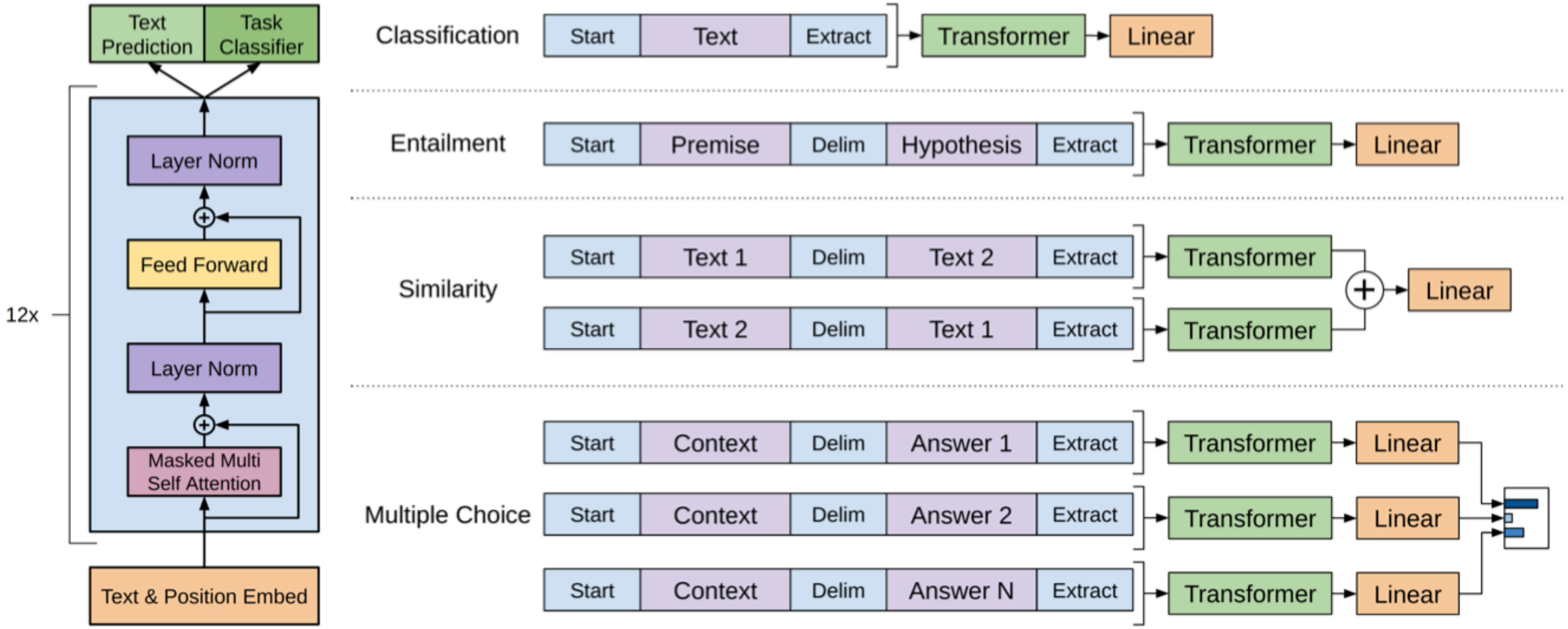
(Devlin et al. , 2018)

One model for all tasks

- **One pre-trained model for all tasks**

  minimal modifications to downstream tasks

  still fine-tuning on $10^3 - 10^5$ downstream examples



(Radford et al. , 2018)

One model for all tasks

- **One pre-trained model for all tasks**
  - BERT (Devlin et al., 2018), RoBERTa (Liu et al., 2019) — encoder models
  - T5 (Raffel et al., 2019), BART (Lewis et al., 2019) — encoder-decoder models
  - GPT-1 (Radford et al., 2018), GPT-2 (Radford et al., 2019) — decoder models

- All based on **Transformers**
- They mainly differ in the pre-training objectives (slight difference in fine-tuning)
- Model sizes and pre-training data are also different!

The Annotated Transformer

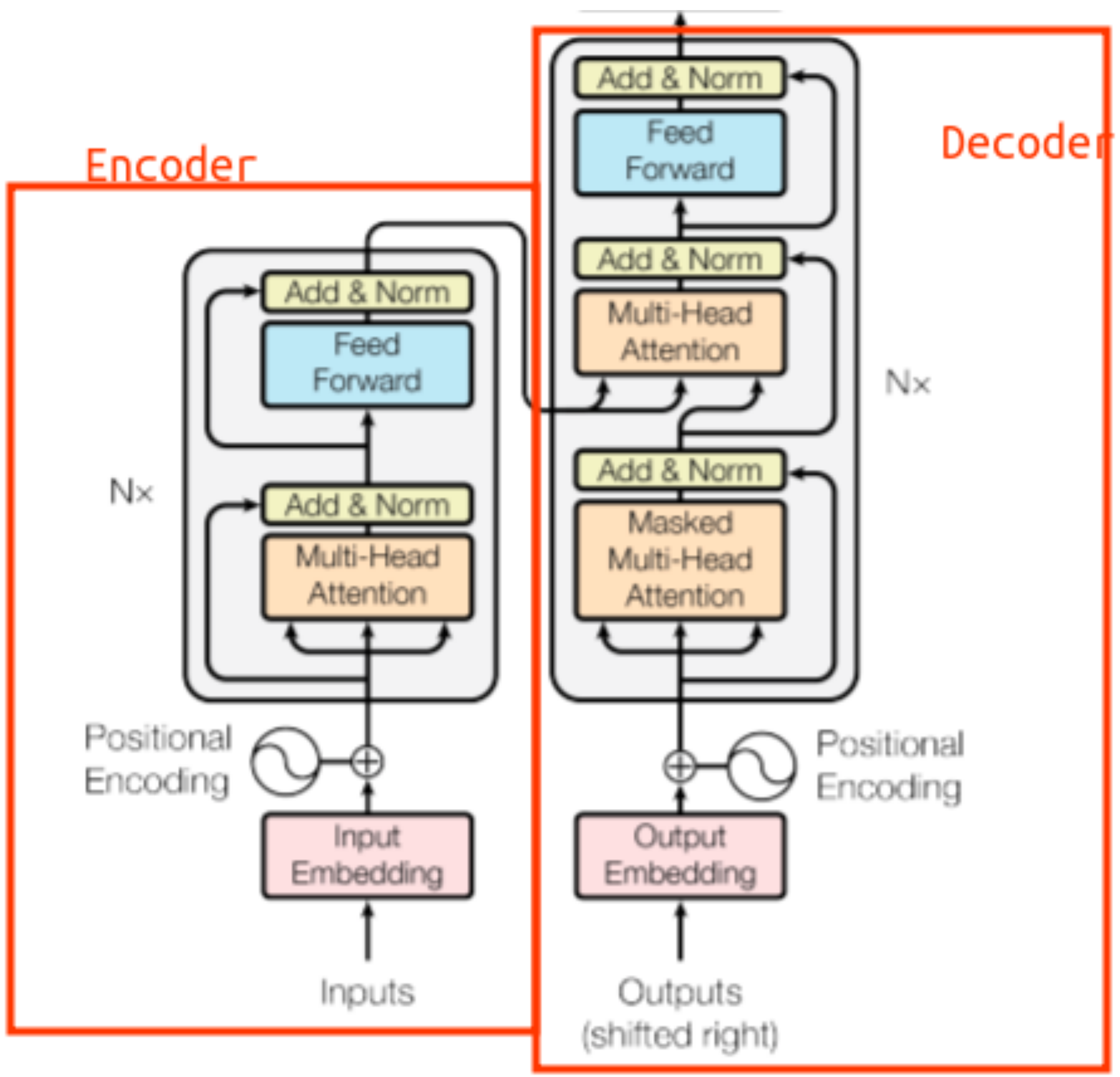
Attention is All You Need

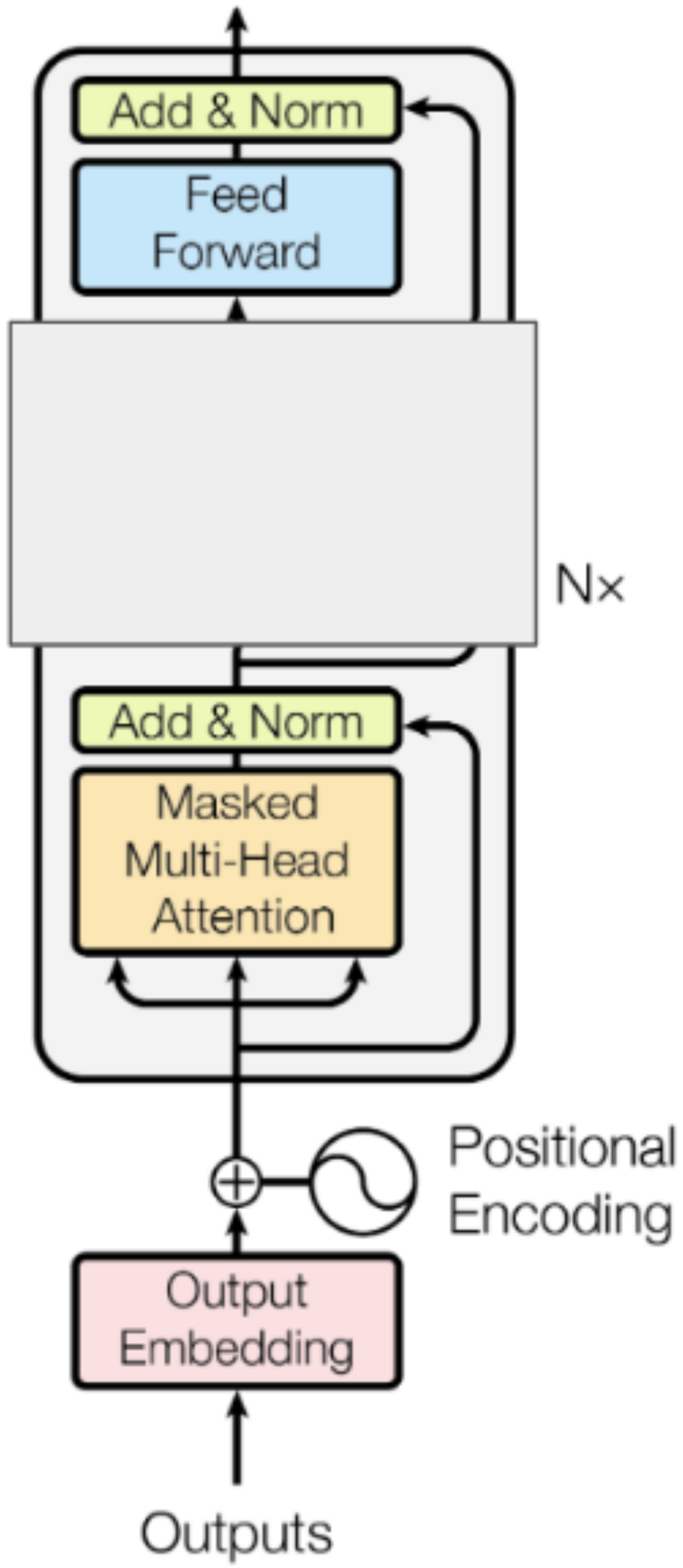
Ashish Vaswani* Google Brain avaswani@google.com
Noam Shazeer* Google Brain noam@google.com
Niki Parmar* Google Research nikip@google.com
Jakob Uszkoreit* Google Research usz@google.com
Llion Jones* Google Research llion@google.com
Aidan N. Gomez* † University of Toronto aidan@cs.toronto.edu
Łukasz Kaiser* Google Brain lukaszkaiser@google.com
Illia Polosukhin* ‡ illia.polosukhin@gmail.com

- *v2022: Austin Huang, Suraj Subramanian, Jonathan Sum, Khalid Almubarak, and Stella Biderman.*
- *Original: Sasha Rush.*

(If you are not familiar with Transformers)

One model for all tasks





Encoder vs decoder models

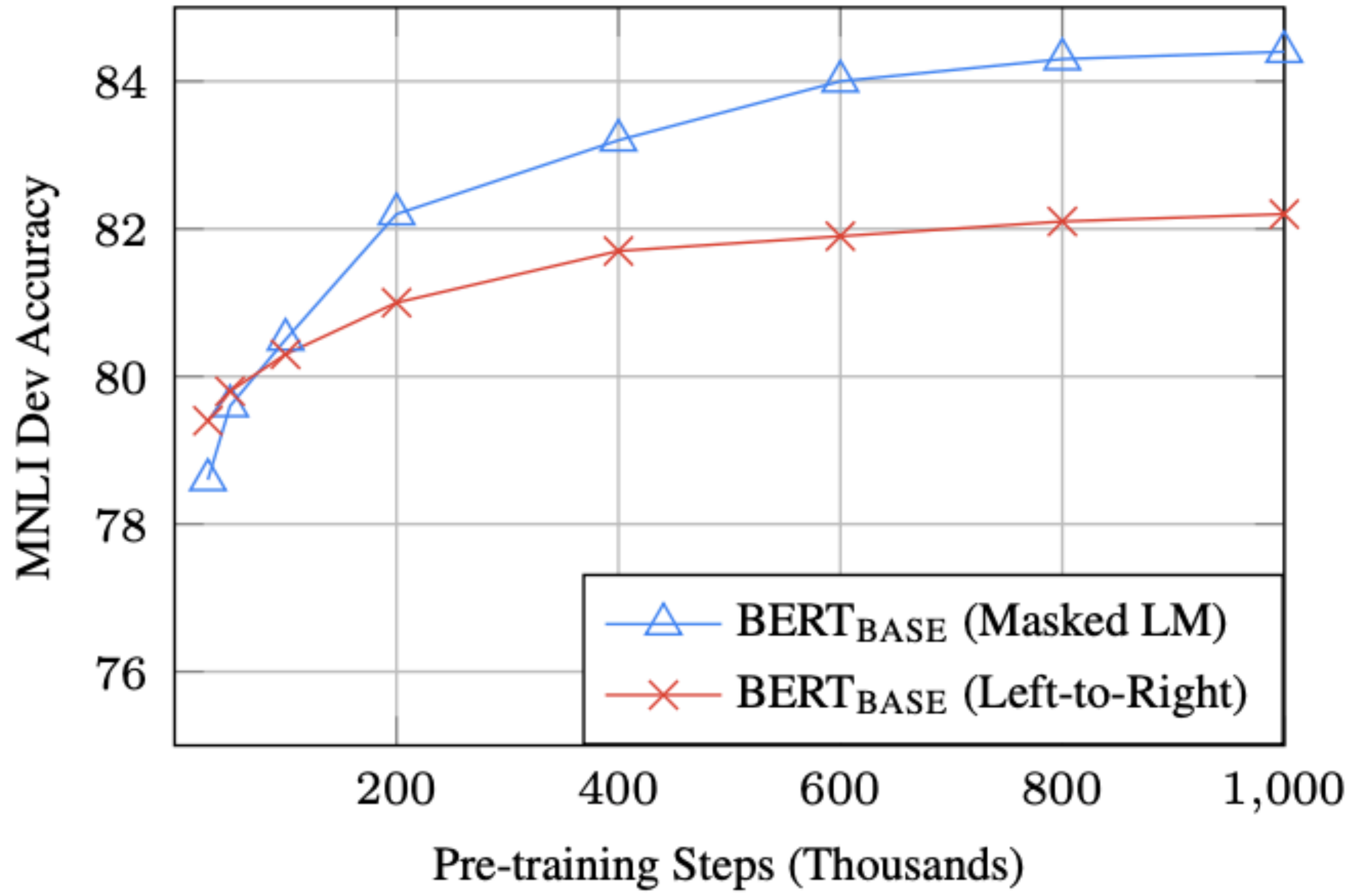


Figure 5: Ablation over number of training steps. This shows the MNLI accuracy after fine-tuning, starting from model parameters that have been pre-trained for k steps. The x-axis is the value of k.

(Devlin et al. , 2018)

- BERT/RoBERTa: 110M/330M parameters
- T5: up to 11B parameters

Yi Tay

2024 JUL 16 - WRITTEN BY YI TAY

What happened to BERT & T5? On Transformer Encoders, PrefixLM and Denoising Objectives

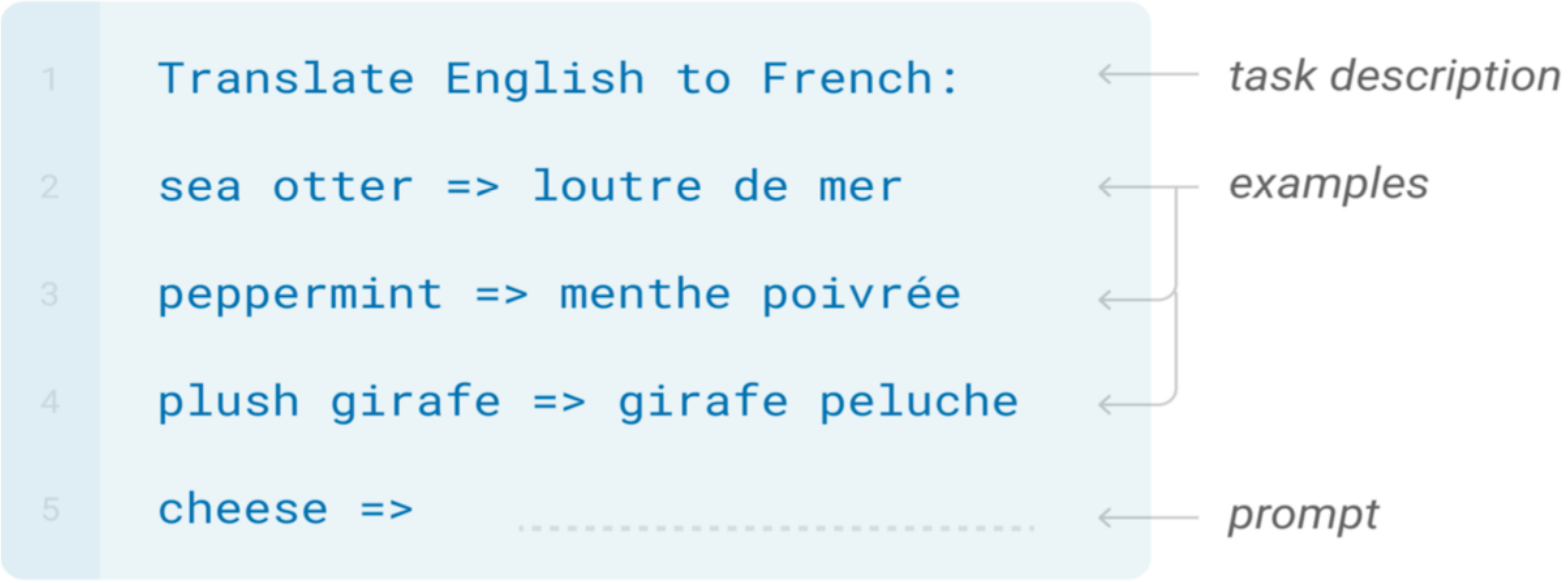
https://www.yitay.net/blog/model-architecture-blogpost-encoders-prefixlm-denoising

- Encoder-only models can’t generate text (easily); harder to scale up
- Bidirectional attention is only important at smaller scale?
- “Masking objectives” can be still combined with autoregressive LMs

GPT-3: main contributions

- An autoregressive language model of 175B parameters, 10x larger than any previous LMs
- Introduced the concept of "in-context learning", and showed competitive performance

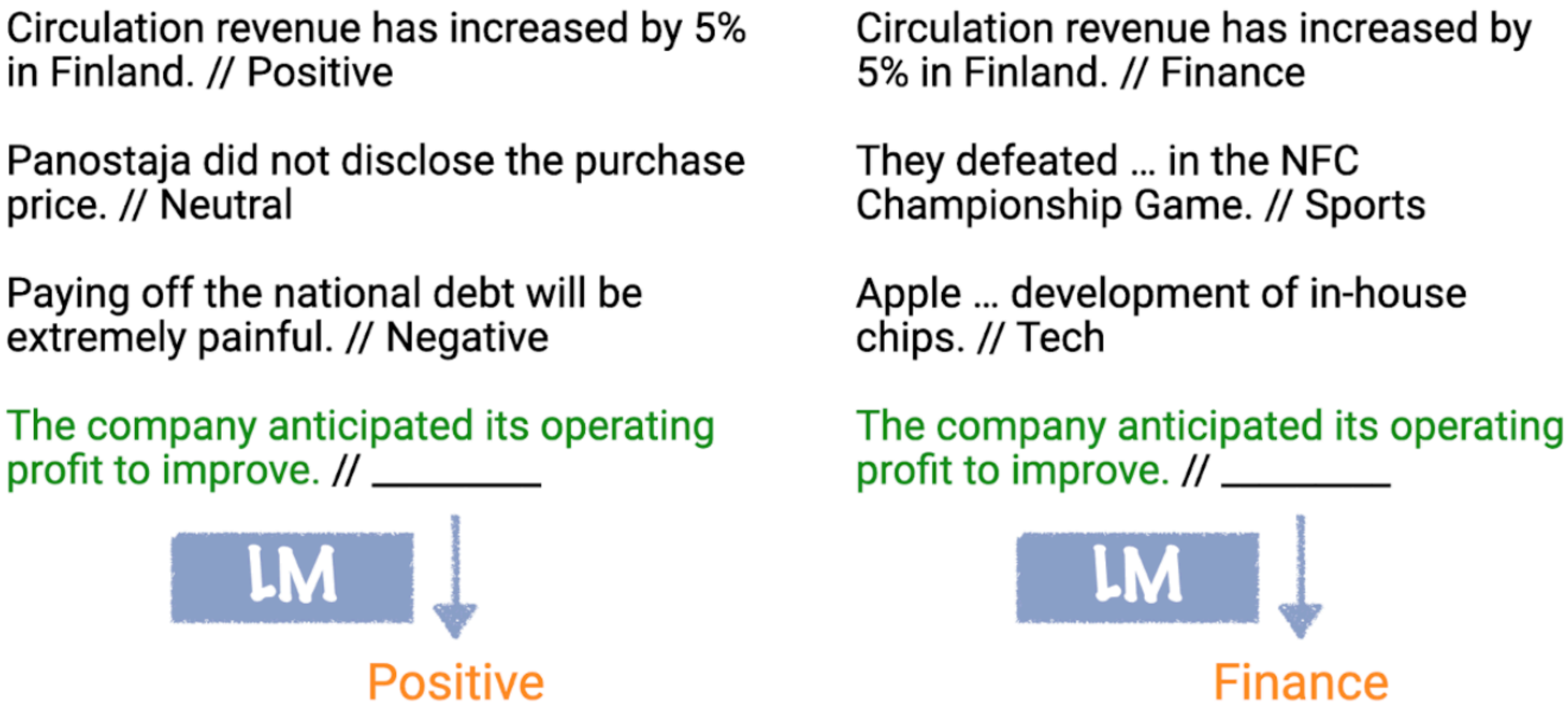
In-context learning: you can perform a task from only **a few examples** or **simple instructions** without any gradient updates or fine-tuning!



https://ai.stanford.edu/blog/understanding-incontext/

GPT-3: main contributions

In-context learning: you can perform a task from only **a few examples** or **simple instructions** without any gradient updates or fine-tuning!



https://ai.stanford.edu/blog/understanding-incontext/

GPT-3: main contributions

In-context learning: you can perform a task from only **a few examples** or **simple instructions** without any gradient updates or fine-tuning!

- **Interesting note**: the idea of in-context learning starts from GPT-2, "though with much more limited results and no systematic study."

3.7. Translation

We test whether GPT-2 has begun to learn how to translate from one language to another. In order to help it infer that this is the desired task, we condition the language model on a context of example pairs of the format `english sentence = french sentence` and then after a final prompt of `english sentence =` we sample from the model with greedy decoding and use the first generated sentence as the translation. On the WMT-14 English-French

3.8. Question Answering

tively. Similar to translation, the context of the language model is seeded with example question answer pairs which helps the model infer the short answer style of the dataset. GPT-2 answers 4.1% of questions correctly when evaluated by the exact match metric commonly used on reading

(Radford et al. , 2019)

Why few-shot learning?

- Collecting large supervised training sets is expensive

Corpus	|Train|	|Test|	Task	Metrics	Domain
			Single-Sentence Tasks		
CoLA	8.5k	**1k**	acceptability	Matthews corr.	misc.
SST-2	67k	1.8k	sentiment	acc.	movie reviews
			Similarity and Paraphrase Tasks		
MRPC	3.7k	1.7k	paraphrase	acc./F1	news
STS-B	7k	1.4k	sentence similarity	Pearson/Spearman corr.	misc.
QQP	364k	**391k**	paraphrase	acc./F1	social QA questions
			Inference Tasks		
MNLI	393k	**20k**	NLI	matched acc./mismatched acc.	misc.
QNLI	105k	5.4k	QA/NLI	acc.	Wikipedia
RTE	2.5k	3k	NLI	acc.	news, Wikipedia
WNLI	634	**146**	coreference/NLI	acc.	fiction books

GLUE (Devlin et al. , 2018)

Why few-shot learning?

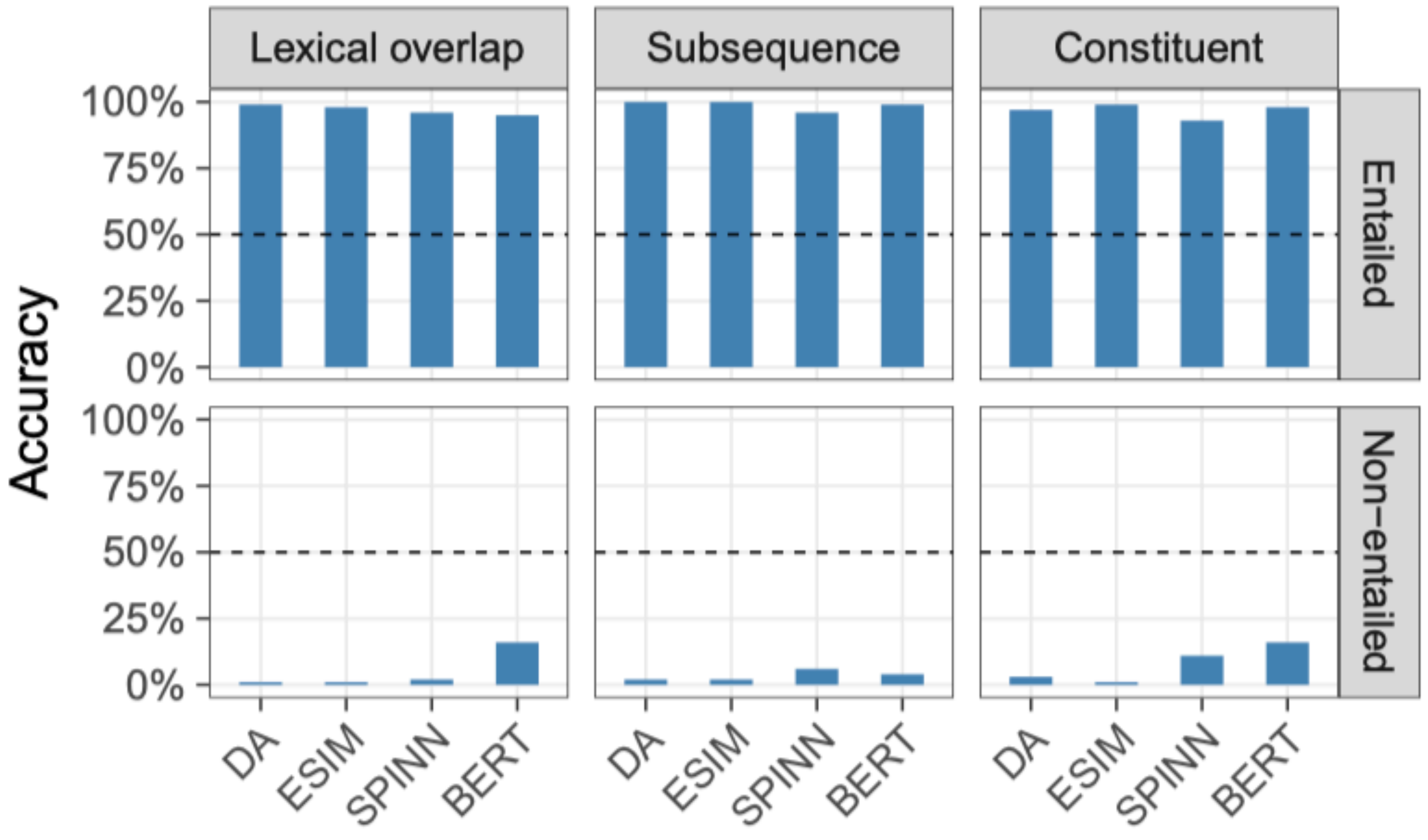
- Fine-tuning can exploit **spurious correlation** and do not generalize well out-of-distribution

NATURAL LANGUAGE INFERENCE E.G., MNLI

- **Premise**: The banker near the judge saw the actor.
- **Hypothesis**: The banker saw the actor.
- **Label**: Entailment

<u>Lexical overlap heuristic: a premise entails all hypotheses constructed from words in the premise</u>

- **Premise**: The doctors visited the lawyer.
- **Hypothesis**: The lawyer visited the doctors.
- **Label**: Not Entailment ❌



(McCoy et al. , 2019)

Why few-shot learning?

- Humans do not require large supervised datasets to learn most language tasks
- It allows humans to seamlessly **mix together** or **switch** between many tasks and tasks when interacting with NLP systems
    - Fluidity
    - Generality

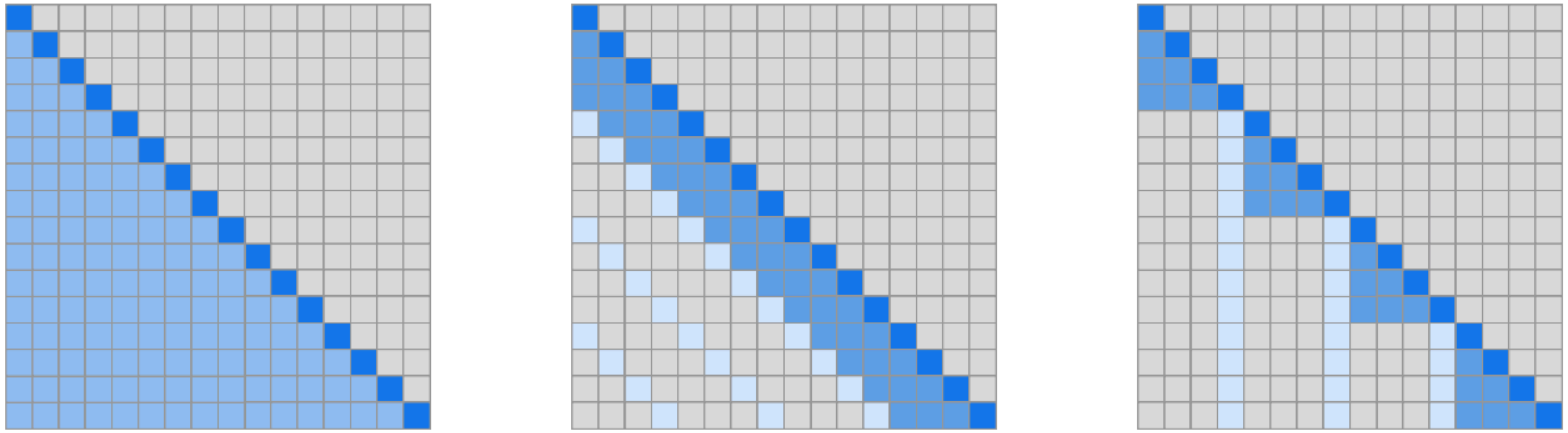
GPT-3: details

Overview of GPT-3

- GPT-3 is a Transformer decoder only trained on large amounts of unlabeled text
- Training objective: next-token prediction

$$L_1(\mathcal{U}) = \sum_i \log P(u_i | u_{i-k}, \ldots, u_{i-1}; \Theta)$$

- Model architecture the same as GPT-2, including modified initialization, pre-normalization
    - Except that "we use alternating dense and locally banded sparse attention patterns in the layers of the Transformer"



(Child et al. , 2019)

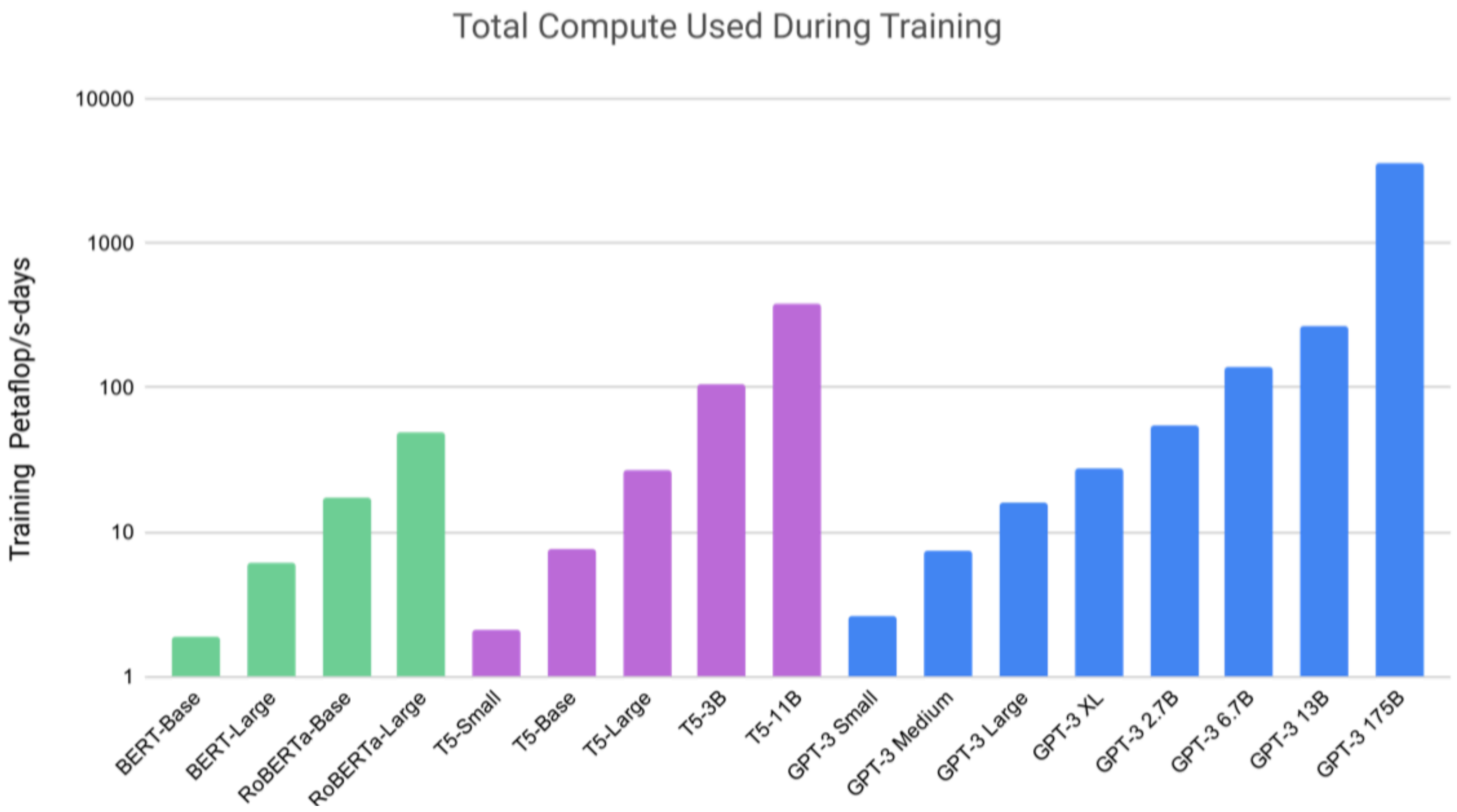
Overview of GPT-3

- GPT-3 is a Transformer decoder only trained on large amounts of unlabeled text
- All models were trained on 300B tokens

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}

- Scaling laws (next week): "scaling of validation loss should be approximately a smooth power law as a function of size"
- Larger models typically use a larger batch size but require a smaller learning rate
- **Context window size = 2048**
- Use a lot of "model parallelism" during training
- Use Adam optimizer $\beta_1 = 0.9, \beta_2 = 0.95, \text{and } \epsilon = 10^{-8}$

GPT-3: training compute



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

“We train much larger models on many fewer tokens”

GPT-3: training data

- Common Crawl (CC) + a set of high-quality, curated data
  - Common Crawl is a nonprofit organization that crawls the web and freely provides its archives and datasets to the public.
  - Lots of low-quality and duplicated content - requires heavy filtering
  - We will see lots of efforts later, e.g., RefineWeb, FineWeb-edu
  - Data in the mix: WebText, Books1, Books2, English Wikipedia
- Filtering CC:
  - Filtering based on similarity to a range of high-quality reference corpora
  - Fuzzy deduplication at the document level
- Data sampling: sample from high-quality data more frequently!

GPT-3: training data

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

Approach

- **Few-shot:** a few demonstrations are prepended in the context (no weights updated allowed)
  - The demonstrations are randomly sampled from training set
  - K: typically 10-100, depending on how many examples can fit in context (2048)
  - Not always “the larger K, the better” => use a development set to decide K
  - Optionally add a natural language prompt
- **One-shot:** special case when K = 1.

  “it most closely matches the way in which some tasks are communicated to humans”

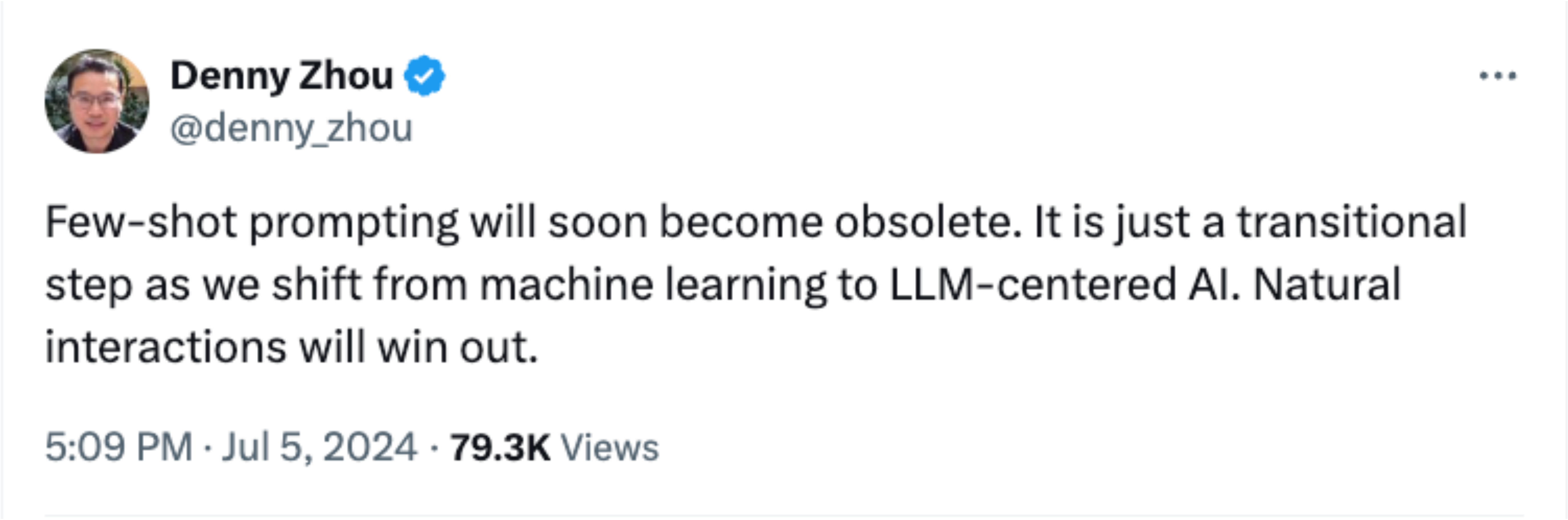
  “it is sometimes difficult to communicate the content or format of a task if no examples are given”
- **Zero-shot:** avoidance of spurious correlation, “unfairly hard”

  “at least some settings zero-shot is closest to how humans perform tasks”

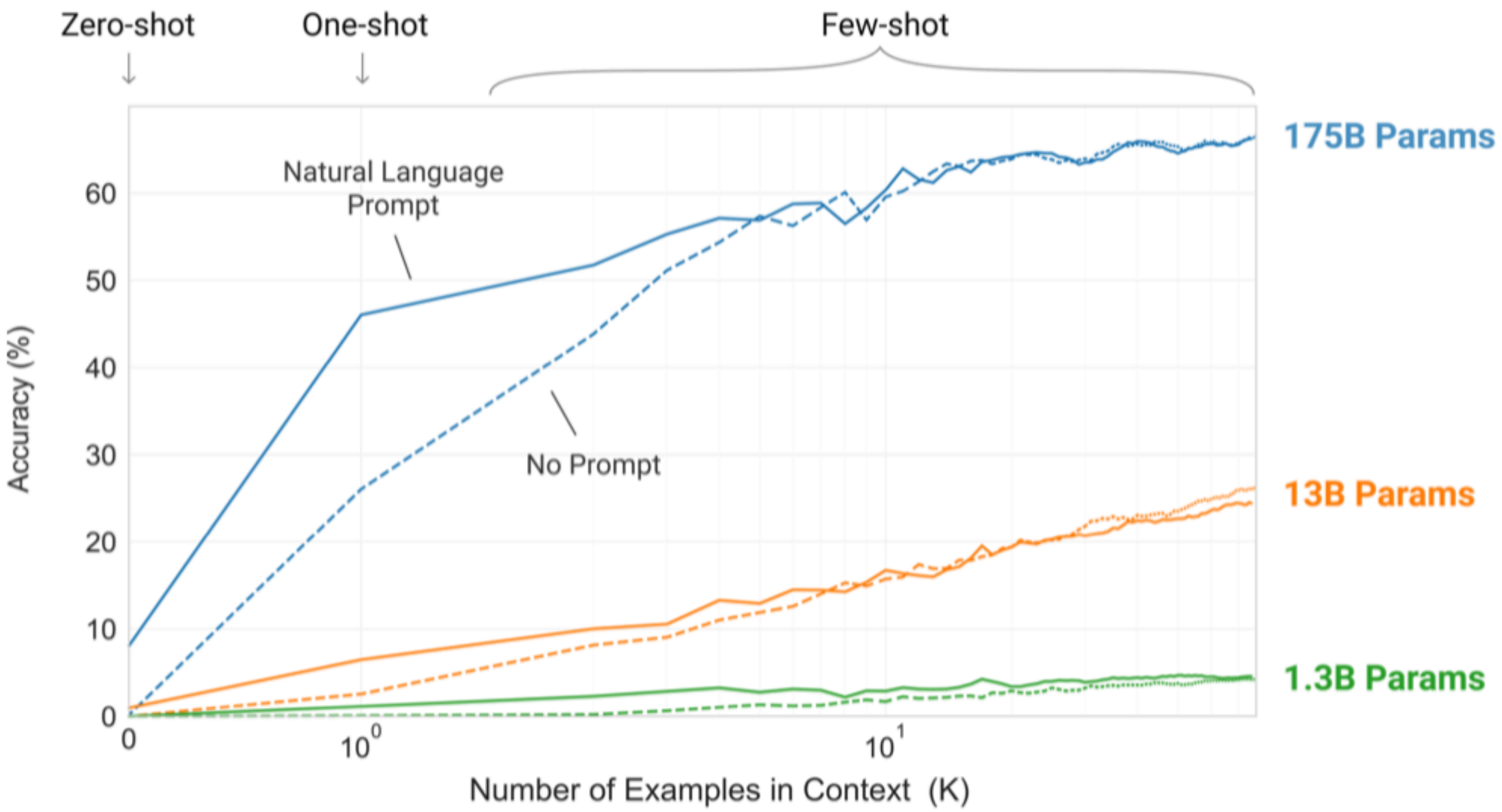
Approach

- **Few-shot:** stronger performance, only slightly behind state-of-the-art fine-tuned models

“however, one-shot, or even sometimes zero-shot, seem like the fairest comparisons to human performance, and **are important targets for future work**."



A summary of results



Evaluation

Evaluation tasks

- Tasks similar to language modeling
- Closed-book question answering
- Machine translation
- Winograd schema and commonsense reasoning
- Reading comprehension
- SuperGLUE
- NLI
- Novel tasks: on-the-fly reasoning, adaptation, open-ended text synthesis

Evaluation protocol

- Open-ended generation: beam search (size = 4), length penalty ($\alpha = 0.6$)
- Multiple choices questions (MCQ):
  - K In-context examples (context + correct completion) + query context
  - Feed each answer choice separately and compare per-token likelihood
  - Additional benefits: $\frac{P(\text{completion}|\text{context})}{P(\text{completion}|\text{answer\_context})}$

Published as a conference paper at ICLR 2024

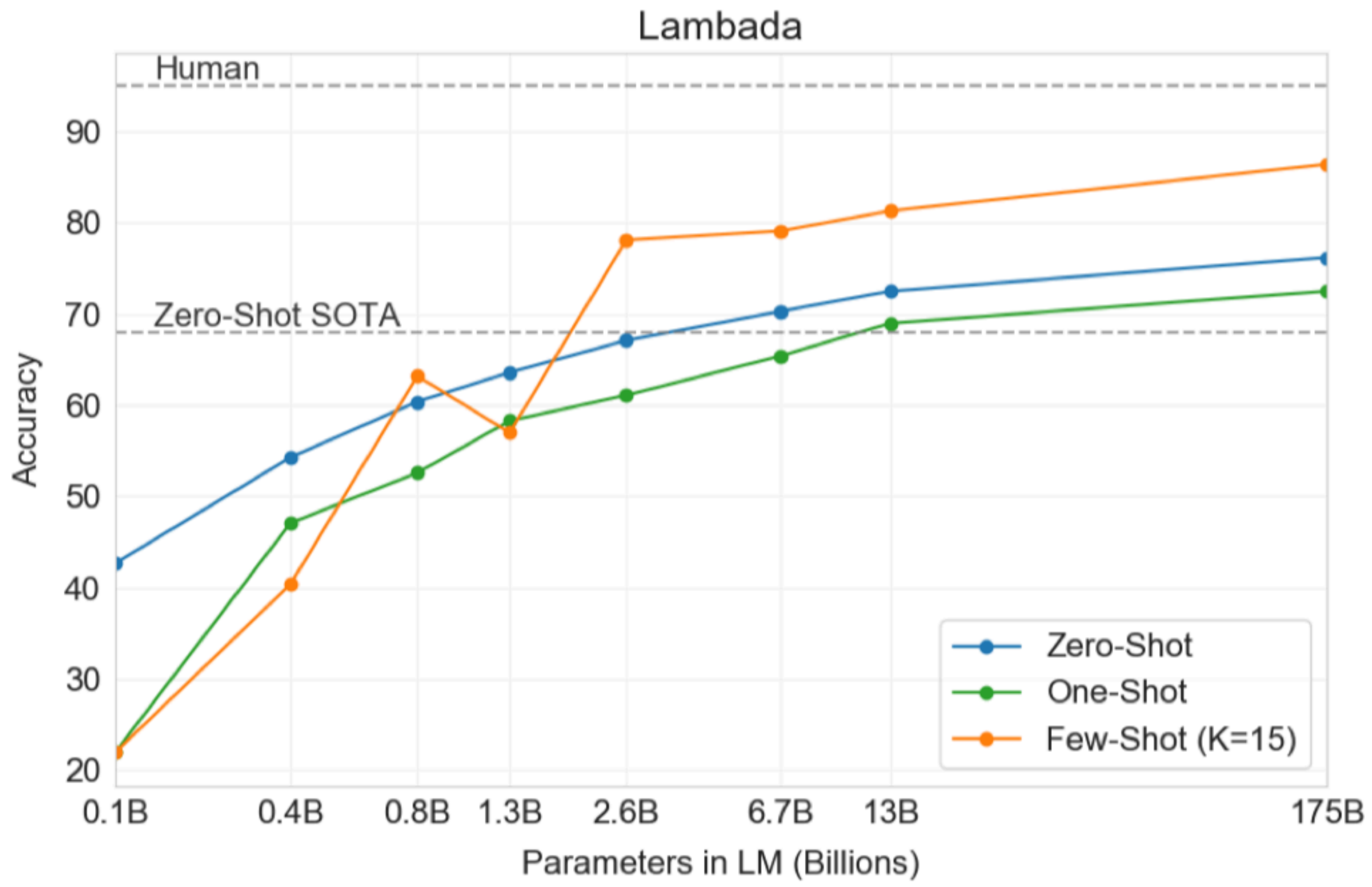
LARGE LANGUAGE MODELS ARE NOT ROBUST MULTIPLE CHOICE SELECTORS

Chujie Zheng[†] **Hao Zhou**[‡] **Fandong Meng**[‡] **Jie Zhou**[‡] **Minlie Huang**[†*]
[†]The CoAI Group, DCST, BNRist, Tsinghua University, Beijing 100084, China
[‡]Pattern Recognition Center, WeChat AI, Tencent Inc., China
chujiezhengchn@gmail.com aihuang@tsinghua.edu.cn

- Yes/no questions: use True/False instead of 0/1

Language modeling

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



LAMBADA

Context: He shook his head, took a step back and held his hands up as he tried to smile without losing a cigarette. “Yes you can,” Julia said in a reassuring voice. “I ’ve already focused on my friend. You just have to click the shutter, on top, here.”
Target sentence: He nodded sheepishly, through his cigarette away and took the _____.
Target word: camera

(Paperno et al. , 2016)

Language modeling

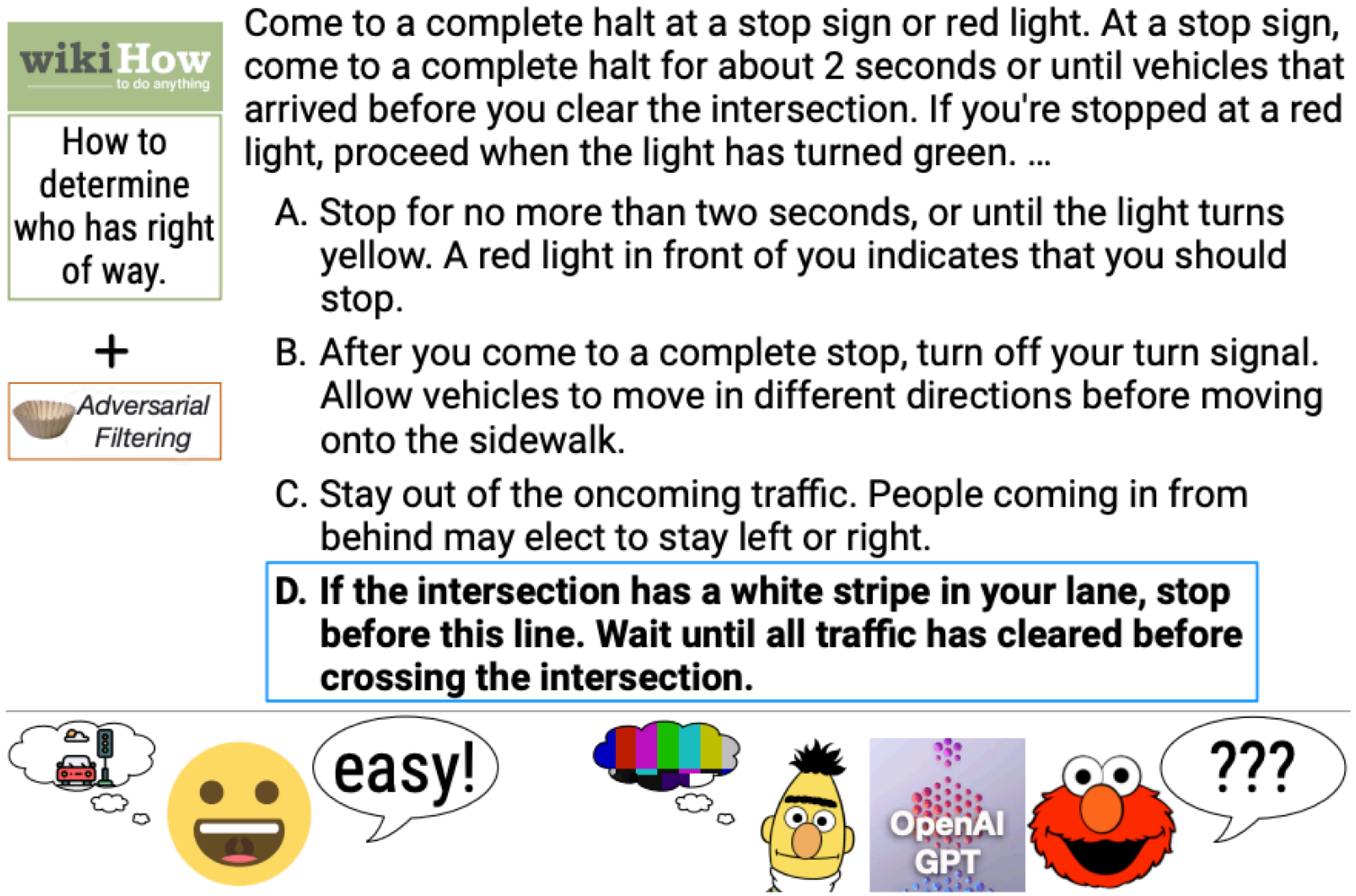
STORYCLOZE

Context	Right Ending	Wrong Ending
Karen was assigned a roommate her first year of college. Her roommate asked her to go to a nearby city for a concert. Karen agreed happily. The show was absolutely exhilarating.	Karen became good friends with her roommate.	Karen hated her roommate.
Jim got his first credit card in college. He didn't have a job so he bought everything on his card. After he graduated he amounted a $10,000 debt. Jim realized that he was foolish to spend so much money.	Jim decided to devise a plan for repayment.	Jim decided to open another credit card.
Gina misplaced her phone at her grandparents. It wasn't anywhere in the living room. She realized she was in the car before. She grabbed her dad's keys and ran outside.	She found her phone in the car.	She didn't want her phone anymore.

(Mostafazadeh et al., 2016)

Language modeling

HELLASWAG



(Zellers et al., 2019)

Open-domain question answering

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**

- Open-book vs closed-book QA

Question	Generated Answer	Correct	Probability
Who wrote the book the origin of species?	Charles Darwin	✓	83.4%
Who is the founder of the ubuntu project?	Mark Shuttleworth	✓	82.0%
Who is the quarterback for the green bay packers?	Aaron Rodgers	✓	81.1%
Panda is a national animal of which country?	China	✓	76.8%
Who came up with the theory of relativity?	Albert Einstein	✓	76.4%
When was the first star wars film released?	1977	✓	71.4%
What is the most common blood type in sweden?	A	✗	70.6%
Who is regarded as the founder of psychoanalysis?	Sigmund Freud	✓	69.3%
Who took the first steps on the moon in 1969?	Neil Armstrong	✓	66.8%
Who is the largest supermarket chain in the uk?	Tesco	✓	65.3%
What is the meaning of shalom in english?	peace	✓	64.0%
Who was the author of the art of war?	Sun Tzu	✓	59.6%
Largest state in the us by land mass?	California	✗	59.2%
Green algae is an example of which type of reproduction?	parthenogenesis	✗	56.5%

Machine translation

- GPT-3's training data: 93% English (by word count)

unsupervised NMT

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

Winograd-style and commonsense reasoning

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

Example: Grace was happy to trade me her sweater for my jacket. She thinks the [sweater | jacket] looks dowdy to her

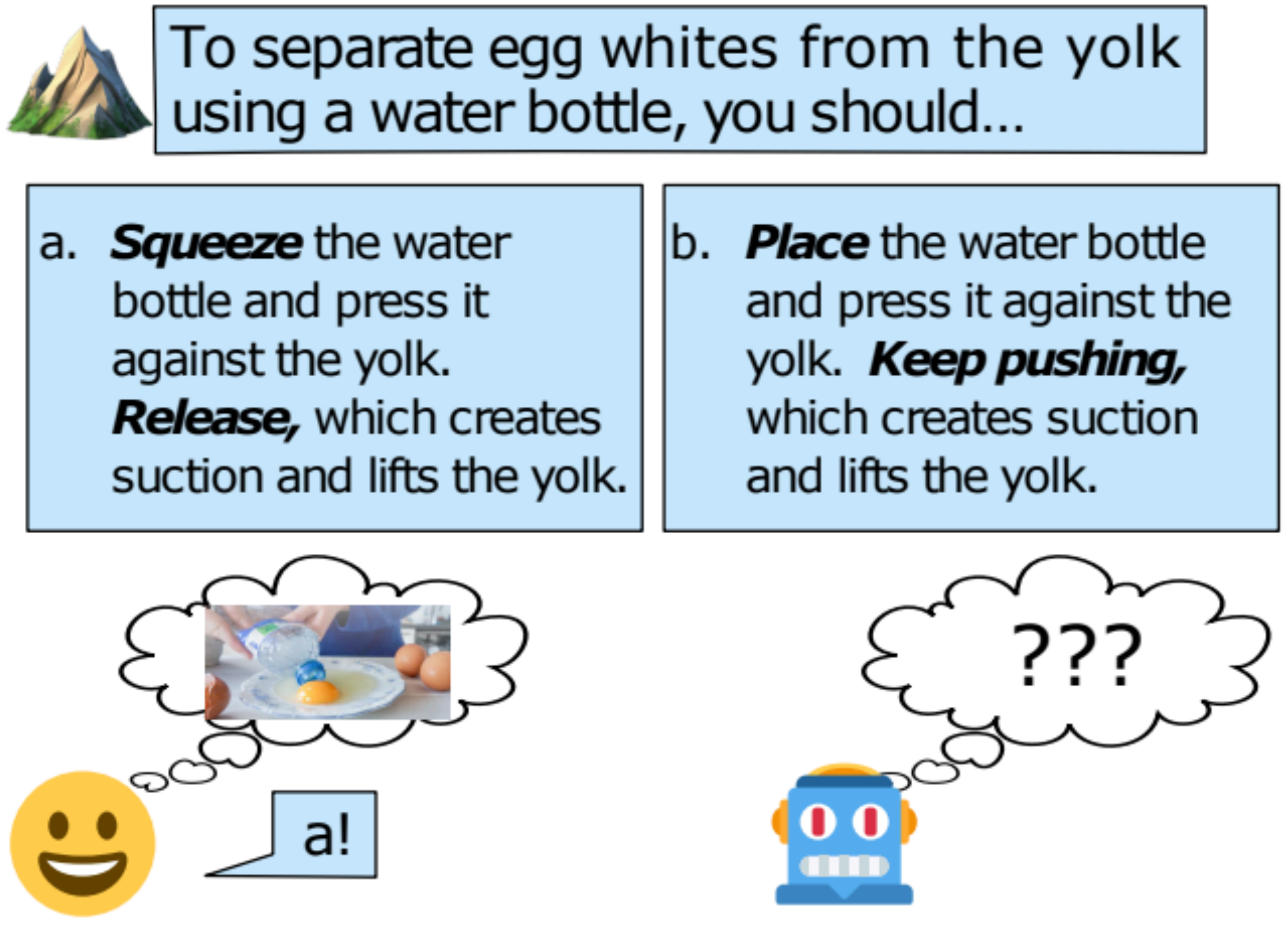
`Correct Context →`	`Grace was happy to trade me her sweater for my jacket. She thinks the sweater`
`Incorrect Context →`	`Grace was happy to trade me her sweater for my jacket. She thinks the jacket`
`Target Completion →`	`looks dowdy on her.`

Figure G.13: Formatted dataset example for Winograd. The 'partial' evaluation method we use compares the probability of the completion given a correct and incorrect context.

Winograd-style and commonsense reasoning

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

PIQA (PHYSICAL QA)



(Bisk et al., 2019)

Winograd-style and commonsense reasoning

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

- ARC: 3rd to 9th grade science exams

Knowledge Type	**Example**
Definition	What is a worldwide increase in temperature called? (A) greenhouse effect (B) global warming (C) ozone depletion (D) solar heating
Basic Facts & Properties	Which element makes up most of the air we breathe? (A) carbon (B) nitrogen (C) oxygen (D) argon
Structure	The crust, the mantle, and the core are structures of Earth. Which description is a feature of Earth's mantle? (A) contains fossil remains (B) consists of tectonic plates (C) is located at the center of Earth (D) has properties of both liquids and solids
Processes & Causal	What is the first step of the process in the formation of sedimentary rocks? (A) erosion (B) deposition (C) compaction (D) cementation
Teleology / Purpose	What is the main function of the circulatory system? (1) secrete enzymes (2) digest proteins (3) produce hormones (4) transport materials
Algebraic	If a red flowered plant (RR) is crossed with a white flowered plant (rr), what color will the offspring be? (A) 100% pink (B) 100% red (C) 50% white, 50% red (D) 100% white
Experiments	Scientists perform experiments to test hypotheses. How do scientists try to remain objective during experiments? (A) Scientists analyze all results. (B) Scientists use safety precautions. (C) Scientists conduct experiments once. (D) Scientists change at least two variables.
Spatial / Kinematic	In studying layers of rock sediment, a geologist found an area where older rock was layered on top of younger rock. Which best explains how this occurred? (A) Earthquake activity folded the rock layers...

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

Subtraction (28.8%)	That year, his **Untitled (1981)**, a painting of a haloed, black-headed man with a bright red skeletal body, depicted amid the artists signature scrawls, was **sold by Robert Lehrman for $16.3 million, well above its $12 million high estimate**.	How many more dollars was the Untitled (1981) painting sold for than the 12 million dollar estimation?	4300000
Comparison (18.2%)	In **1517, the seventeen-year-old King sailed to Castile**. There, his Flemish court .... **In May 1518, Charles traveled to Barcelona in Aragon**.	Where did Charles travel to first, Castile or Barcelona?	Castile
Selection (19.4%)	In 1970, to commemorate the 100th anniversary of the founding of Baldwin City, **Baker University professor and playwright Don Mueller and Phyllis E. Braun, Business Manager, produced a musical play entitled The Ballad Of Black Jack** to tell the story of the events that led up to the battle.	Who was the University professor that helped produce The Ballad Of Black Jack, Ivan Boyd or Don Mueller?	Don Mueller

DROP (Dua et al. , 2019)

What did the General Conference on Weights and Measures name after Tesla in 1960?

Ground Truth Answers: SI unit of magnetic flux density

Tesla was renowned for his achievements and showmanship, eventually earning him a reputation in popular culture as an archetypal "mad scientist". His patents earned him a considerable amount of money, much of which was used to finance his own projects with varying degrees of success.:121,154 He lived most of his life in a series of New York hotels, through his retirement. Tesla died on 7 January 1943. His work fell into relative obscurity after his death, but in 1960 the General Conference on Weights and Measures named the SI unit of magnetic flux density the tesla in his honor. There has been a resurgence in popular interest in Tesla since the 1990s.

SQuAD (Rajpurkar et al. , 2017)

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

Passage:
In a small village in England about 150 years ago, a mail coach was standing on the street. It didn't come to that village often. People had to pay a lot to get a letter. The person who sent the letter didn't have to pay the postage, while the receiver had to.
"Here's a letter for Miss Alice Brown," said the mailman.
" I'm Alice Brown," a girl of about 18 said in a low voice.
Alice looked at the envelope for a minute, and then handed it back to the mailman.
"I'm sorry I can't take it, I don't have enough money to pay it", she said.
A gentleman standing around were very sorry for her. Then he came up and paid the postage for her.
When the gentleman gave the letter to her, she said with a smile, " Thank you very much, This letter is from Tom. I'm going to marry him. He went to London to look for work. I've waited a long time for this letter, but now I don't need it, there is nothing in it."
"Really? How do you know that?" the gentleman said in surprise.
"He told me that he would put some signs on the envelope. Look, sir, this cross in the corner means that he is well and this circle means he has found work. That's good news."
The gentleman was Sir Rowland Hill. He didn't forgot Alice and her letter.
"The postage to be paid by the receiver has to be changed," he said to himself and had a good plan.
"The postage has to be much lower, what about a penny? And the person who sends the letter pays the postage. He has to buy a stamp and put it on the envelope." he said . The government accepted his plan. Then the first stamp was put out in 1840. It was called the "Penny Black". It had a picture of the Queen on it.

Questions:

1): The first postage stamp was made _.
A. in England B. in America C. by Alice D. in 1910

2): The girl handed the letter back to the mailman because _ .
A. she didn't know whose letter it was
B. she had no money to pay the postage
C. she received the letter but she didn't want to open it
D. she had already known what was written in the letter

3): We can know from Alice's words that _ .
A. Tom had told her what the signs meant before leaving
B. Alice was clever and could guess the meaning of the signs
C. Alice had put the signs on the envelope herself
D. Tom had put the signs as Alice had told him to

4): The idea of using stamps was thought of by _ .
A. the government
B. Sir Rowland Hill
C. Alice Brown
D. Tom

5): From the passage we know the high postage made _ .
A. people never send each other letters
B. lovers almost lose every touch with each other
C. people try their best to avoid paying it
D. receivers refuse to pay the coming letters

Answer: ADABC

RACE (Lai et al. , 2017)

- Reading comprehension tests for middle and high school Chinese students (age between 12 and 18)

Reading comprehension

Context →	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. The Helsinki metropolitan area includes the urban core of Helsinki, Espoo, Vantaa, Kauniainen, and surrounding commuter towns. It is the world's northernmost metro area of over one million people, and the city is the northernmost capital of an EU member state. The Helsinki metropolitan area is the third largest metropolitan area in the Nordic countries after Stockholm and Copenhagen, and the City of Helsinki is the third largest after Stockholm and Oslo. Helsinki is Finland's major political, educational, financial, cultural, and research center as well as one of northern Europe's major cities. Approximately 75% of foreign companies that operate in Finland have settled in the Helsinki region. The nearby municipality of Vantaa is the location of Helsinki Airport, with frequent service to various destinations in Europe and Asia. 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:
Target Completion →	Helsinki, Espoo, Vantaa, Kauniainen, and surrounding commuter towns

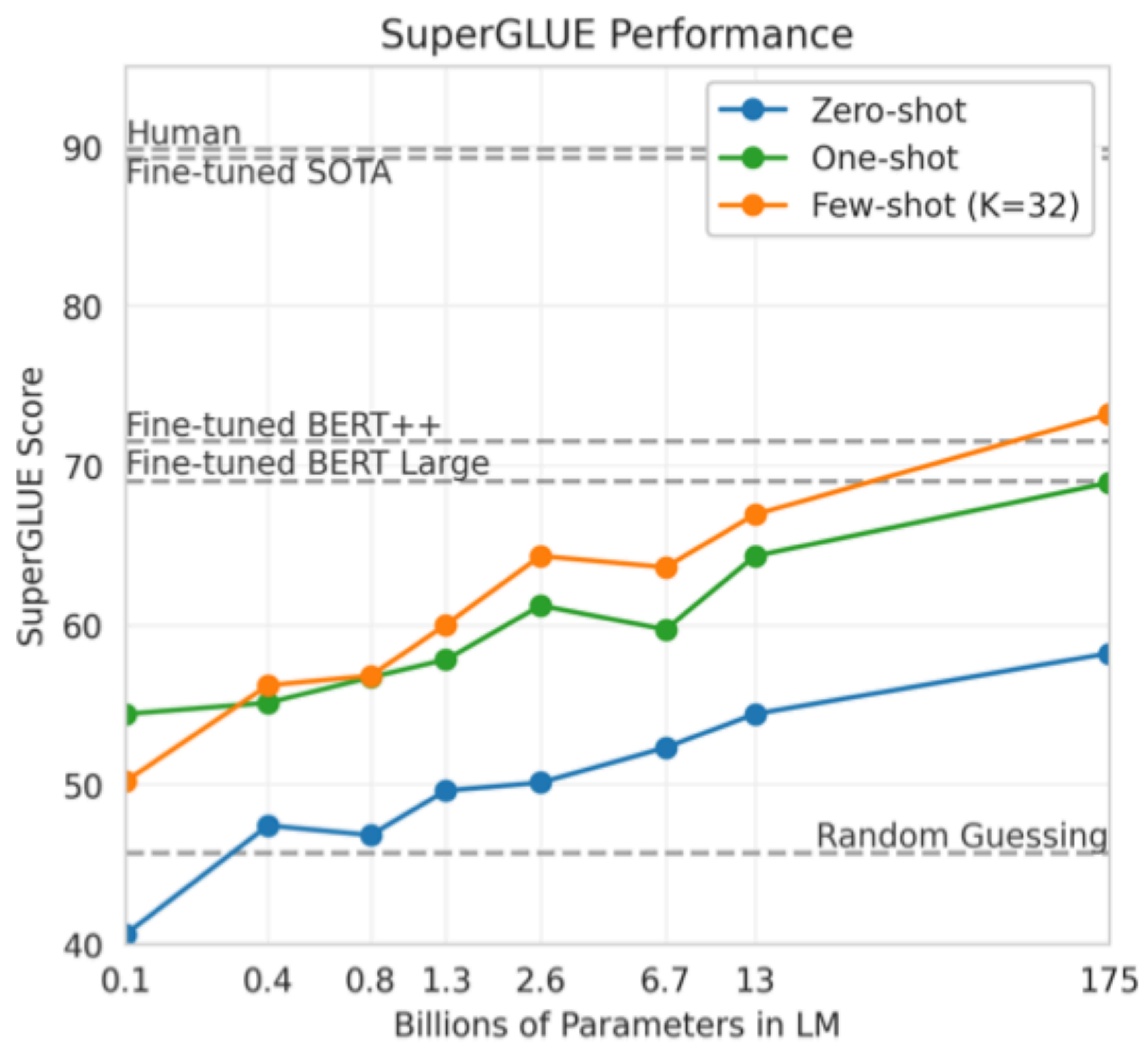
Figure G.18: Formatted dataset example for CoQA

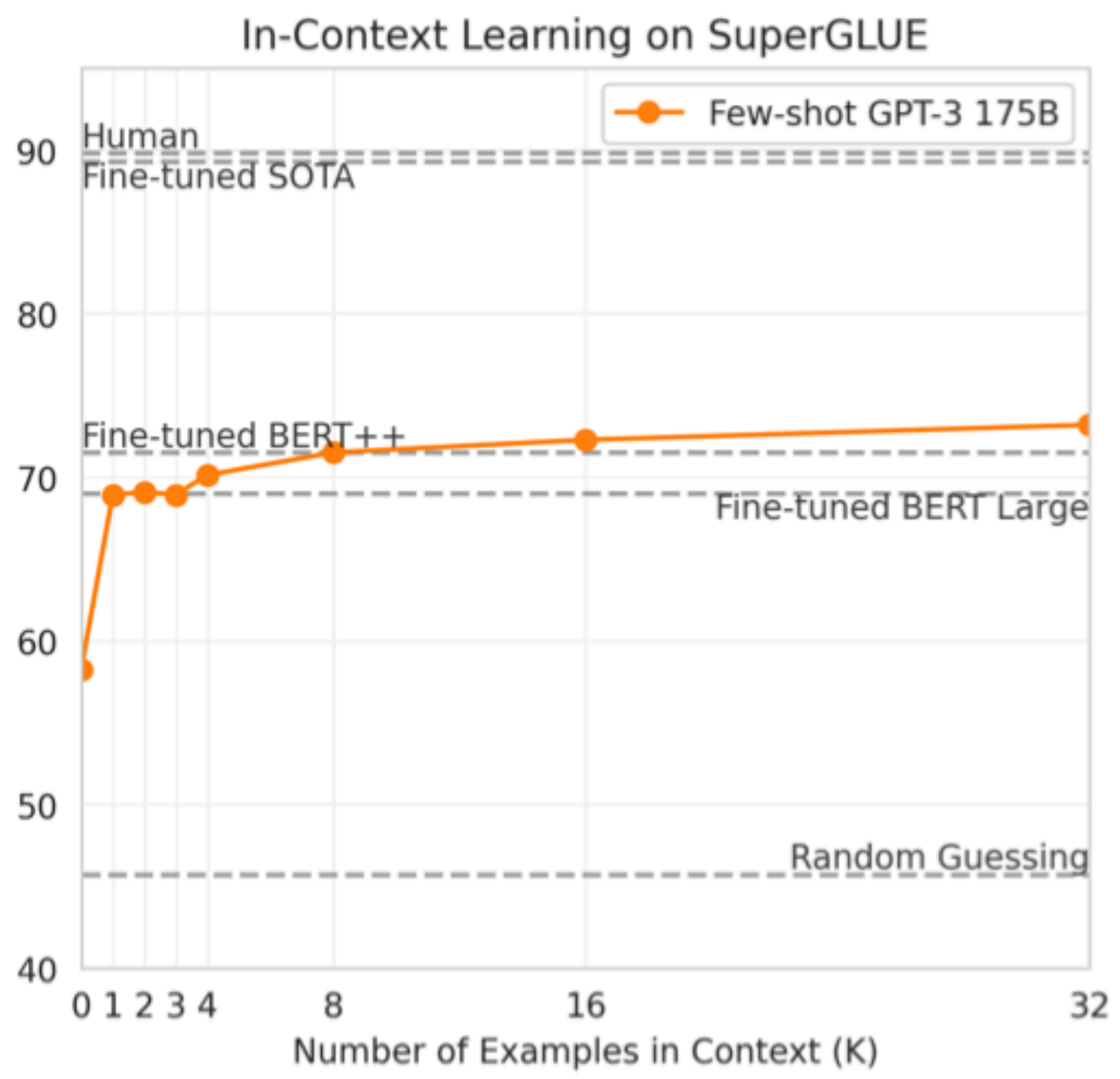
CoQA (Reddy et al., 2019)

SuperGLUE

	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

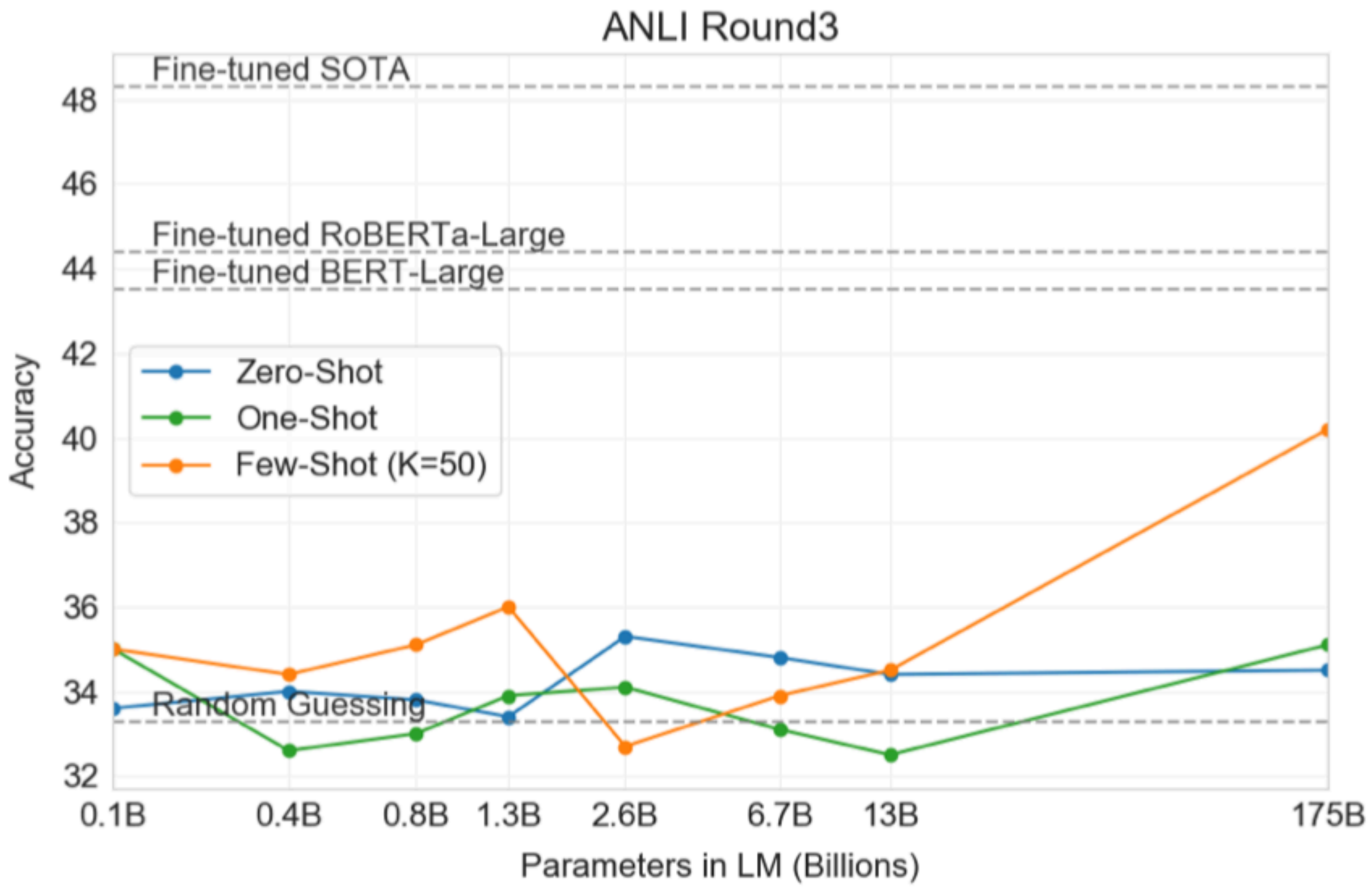
	WiC Accuracy	WSC Accuracy	MultiRC Accuracy	MultiRC F1a	ReCoRD Accuracy	ReCoRD F1
Fine-tuned SOTA	**76.1**	**93.8**	**62.3**	**88.2**	**92.5**	**93.3**
Fine-tuned BERT-Large	69.6	64.6	24.1	70.0	71.3	72.0
GPT-3 Few-Shot	49.4	80.1	30.5	75.4	90.2	91.1





SuperGLUE (Wang et al. , 2019)

Natural language inference (NLI)



ANLI (Nie et al., 2019)

Novel tasks

- Arithmetic
- Word scrambling and manipulation
- SAT analogies
- News article generation
- Learning and using novel words

Why synthetic tasks?

- Easier to control, scale and manipulate
- Less data contamination
- Sometimes provides very clear insights of what is going on

Novel tasks

Context →	Please unscramble the letters into a word, and write that word: asinoc =
Target Completion →	casino

Figure G.19: Formatted dataset example for Cycled Letters

Context →	Please unscramble the letters into a word, and write that word: r e!c.i p r o.c a/l =
Target Completion →	reciprocal

Figure G.26: Formatted dataset example for Symbol Insertion

Context →	Please unscramble the letters into a word, and write that word: taefed =
Target Completion →	defeat

Figure G.27: Formatted dataset example for Reversed Words

Novel tasks

Context →	Q: What is 98 plus 45? A:
Target Completion →	143

Figure G.44: Formatted dataset example for Arithmetic 2D+

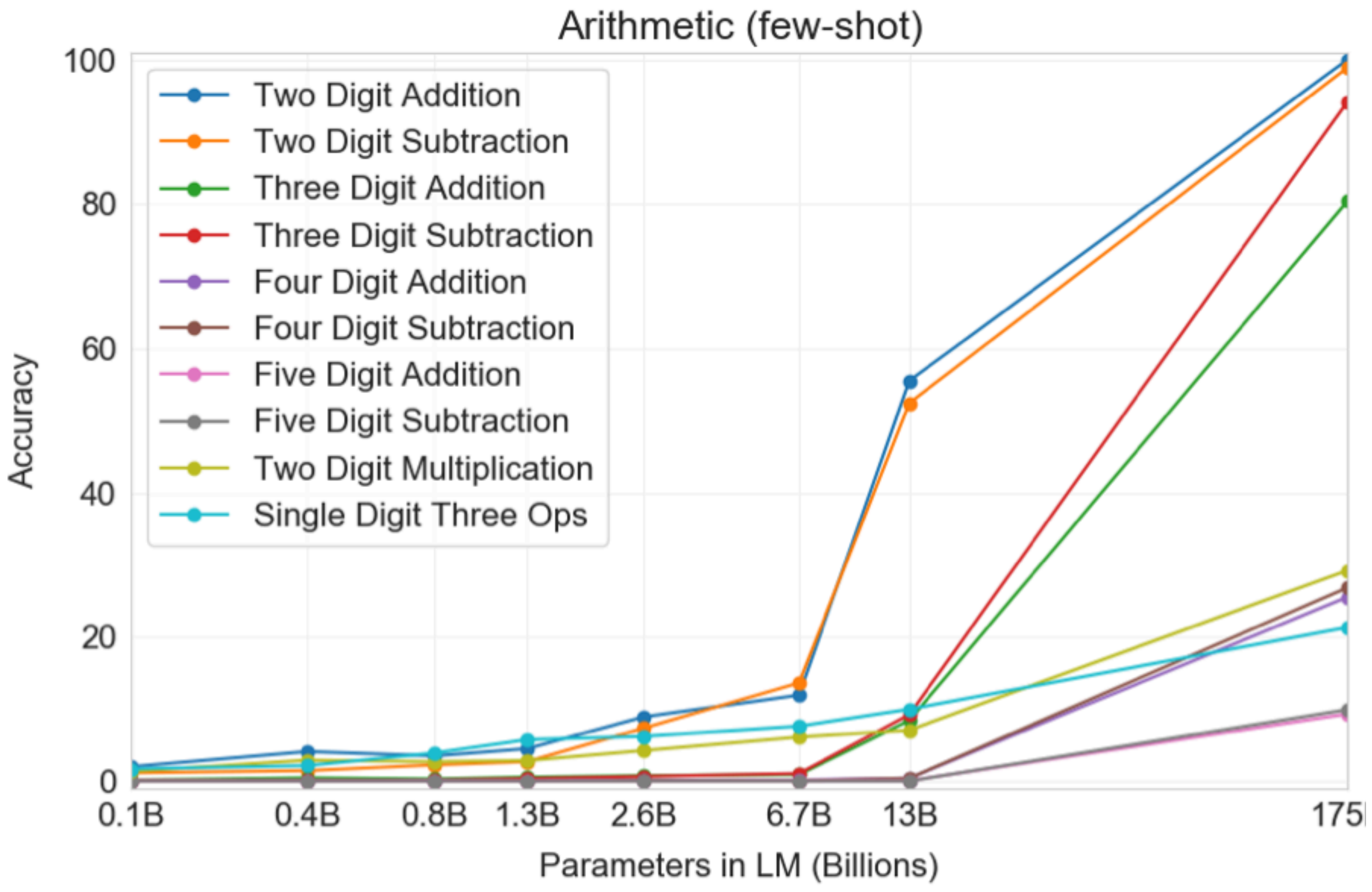
Context →	Q: What is 6209 minus 3365? A:
Target Completion →	2844

Figure G.48: Formatted dataset example for Arithmetic 4D-

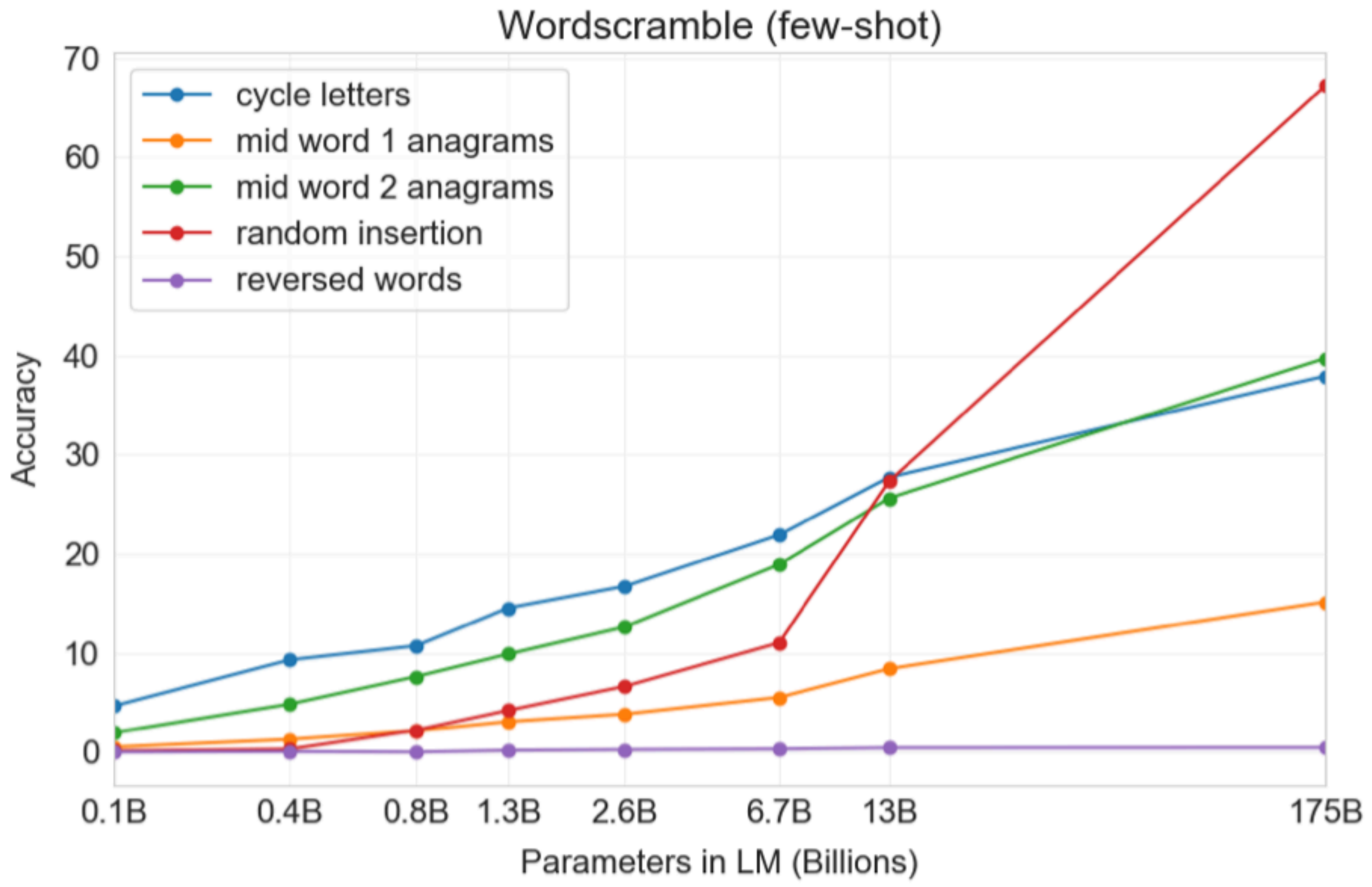
Context →	lull is to trust as
Correct Answer →	cajole is to compliance
Incorrect Answer →	balk is to fortitude
Incorrect Answer →	betray is to loyalty
Incorrect Answer →	hinder is to destination
Incorrect Answer →	soothe is to passion

Figure G.12: Formatted dataset example for SAT Analogies

Novel tasks



Arithmetic (few-shot)
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
Accuracy
0
20
40
60
80
100
0.1B
0.4B
0.8B
1.3B
2.6B
6.7B
13B
175
Parameters in LM (Billions)



Wordscramble (few-shot)
cycle letters
mid word 1 anagrams
mid word 2 anagrams
random insertion
reversed words
Accuracy
0
10
20
30
40
50
60
70
0.1B
0.4B
0.8B
1.3B
2.6B
6.7B
13B
175B
Parameters in LM (Billions)

Next lecture

- Limitations (e.g., contamination) and broader impact
- (Brief) understanding in-context learning
- GPT-3 $\Longrightarrow$ LlaMA 3.1
    - **Required reading**: Sections 1, 2, 3.1, 3.2, 3.4, 5.1
    - What are the major changes in terms of data, architecture, training and evaluation?