Fall 2024 COS597R: Deep Dive into Large Language Models

Danqi Chen, Sanjeev Arora

PLi PRINCETON UNIVERSITY

Lecture 3: Pre-training II

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

This lecture

- GPT-3 (cont'd)
- Understanding in-context learning (brief)
- GPT-3 vs Llama 3

Required reading: LLaMA 3

[Submitted on 31 Jul 2024 (v1), last revised 15 Aug 2024 (this version, v2)]

The Llama 3 Herd of Models

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone et al. (434 additional authors not shown)

Modern artificial intelligence (AI) systems are powered by foundation models. This paper presents a new set of foundation models, called Llama 3. It is a herd of language models that natively support multilinguality, coding, reasoning, and tool usage. Our largest model is a dense Transformer with 405B parameters and a context window of up to 128K tokens. This paper presents an extensive empirical evaluation of Llama 3. We find that Llama 3 delivers comparable quality to leading language models such as GPT-4 on a plethora of tasks. We publicly release Llama 3, including pre-trained and post-trained versions of the 405B parameter language model and our Llama Guard 3 model for input and output safety. The paper also presents the results of experiments in which we integrate image, video, and speech capabilities into Llama 3 via a compositional approach. We observe this approach performs competitively with the state-of-the-art on image, video, and speech recognition tasks. The resulting models are not yet being broadly released as they are still under development.

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

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

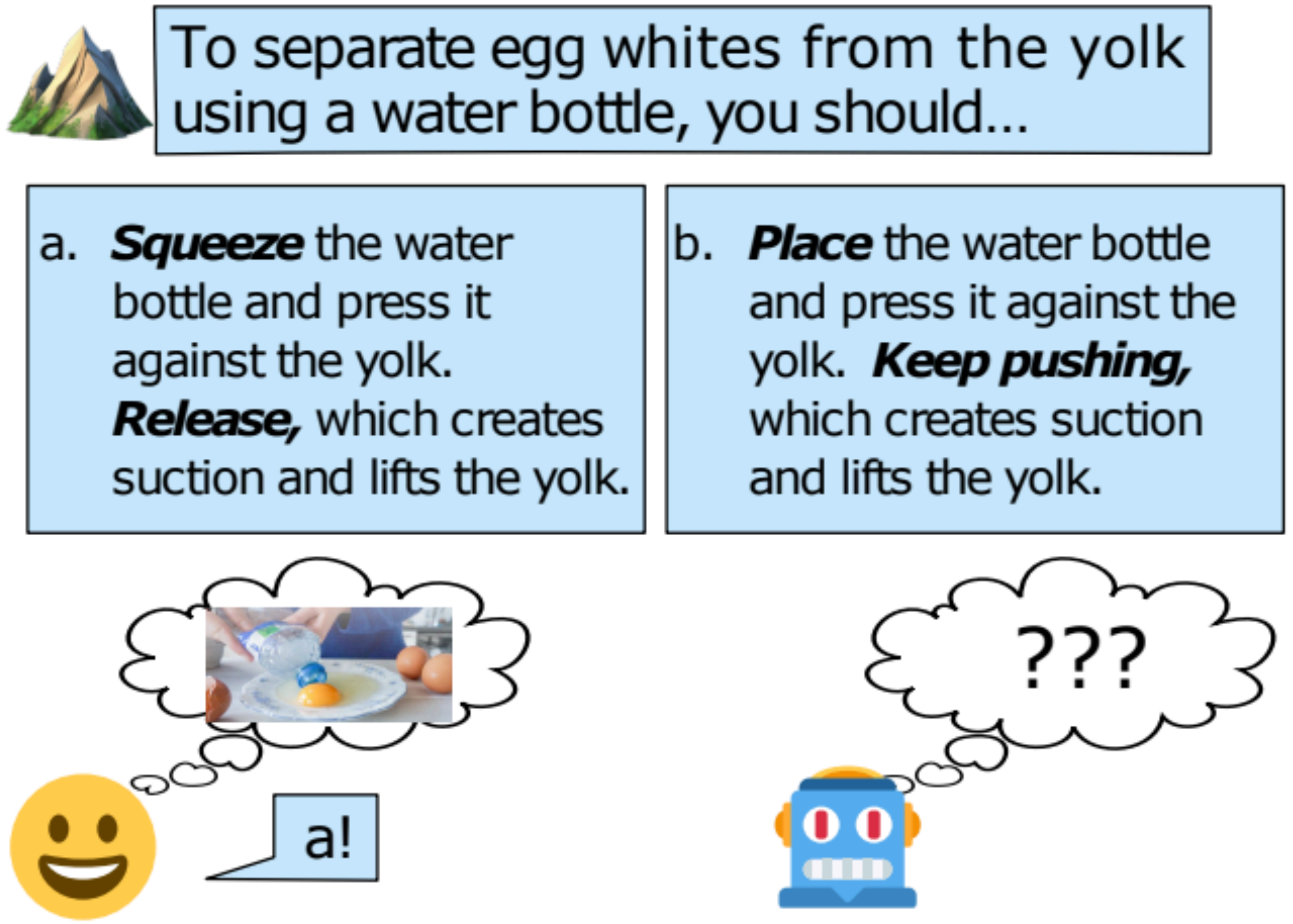
`Correct Context →`	`Grace was happy to trade me her sweater for my jacket. She thinks the sweater`
`Incorrect Context →`	`Grace was happy to trade me her sweater for my jacket. She thinks the jacket`
`Target Completion →`	`looks dowdy on her.`

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

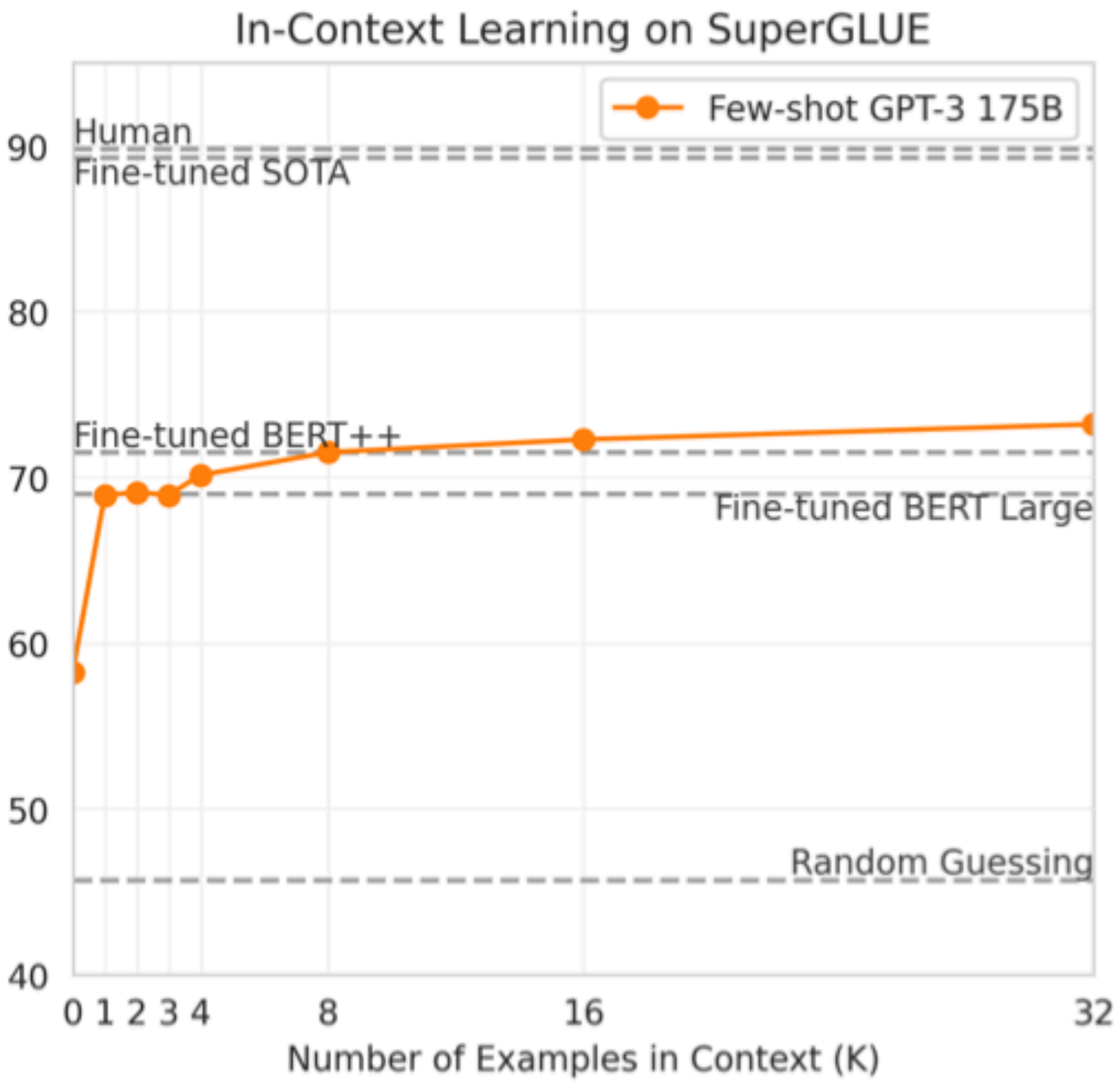
Figure G.18: Formatted dataset example for CoQA

CoQA (Reddy et al., 2019)

SuperGLUE

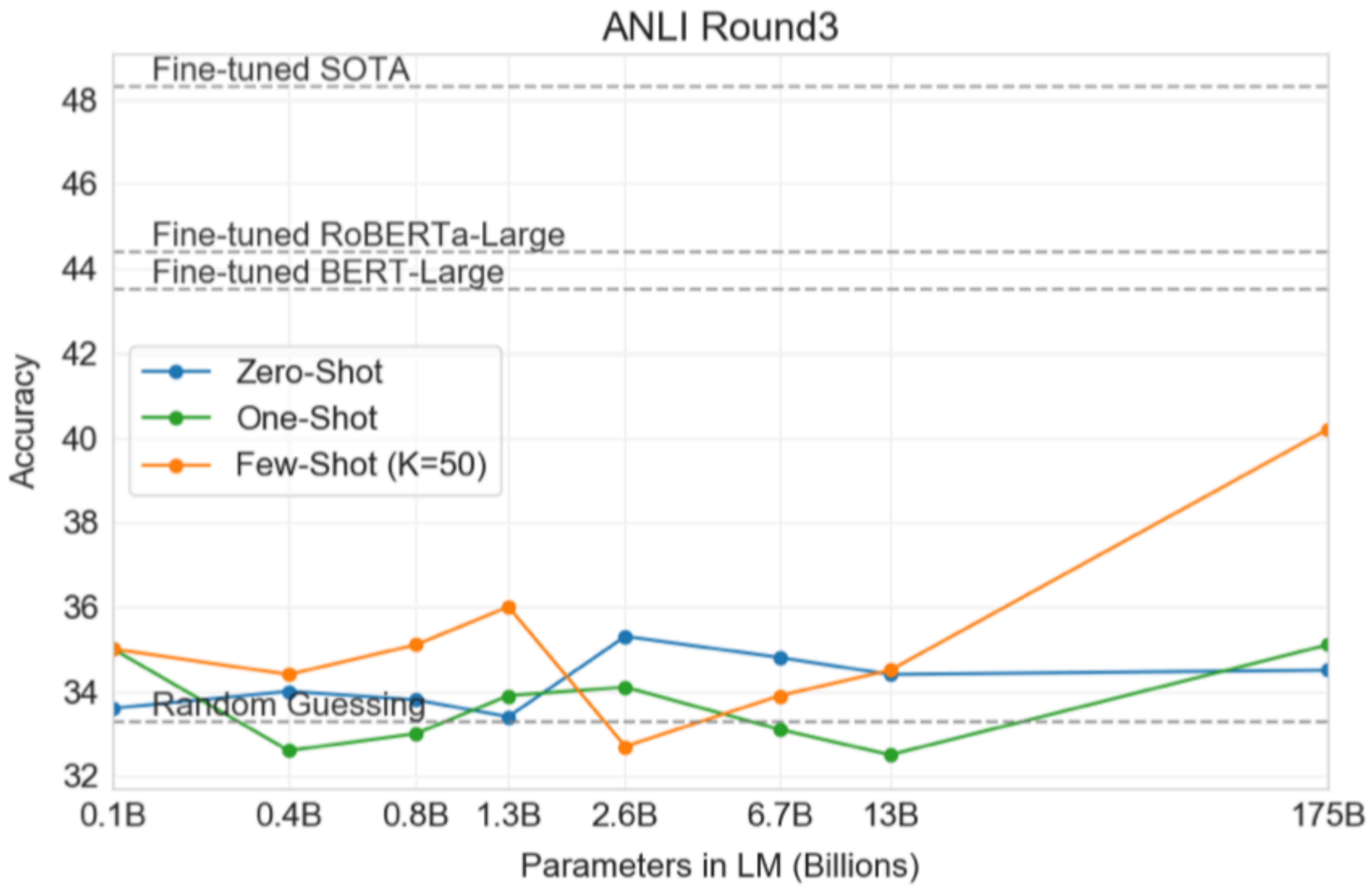
	SuperGLUE Average	BoolQ Accuracy	CB Accuracy	CB F1	COPA Accuracy	RTE Accuracy
Fine-tuned SOTA	**89.0**	**91.0**	**96.9**	**93.9**	**94.8**	**92.5**
Fine-tuned BERT-Large	69.0	77.4	83.6	75.7	70.6	71.7
GPT-3 Few-Shot	71.8	76.4	75.6	52.0	92.0	69.0

	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+

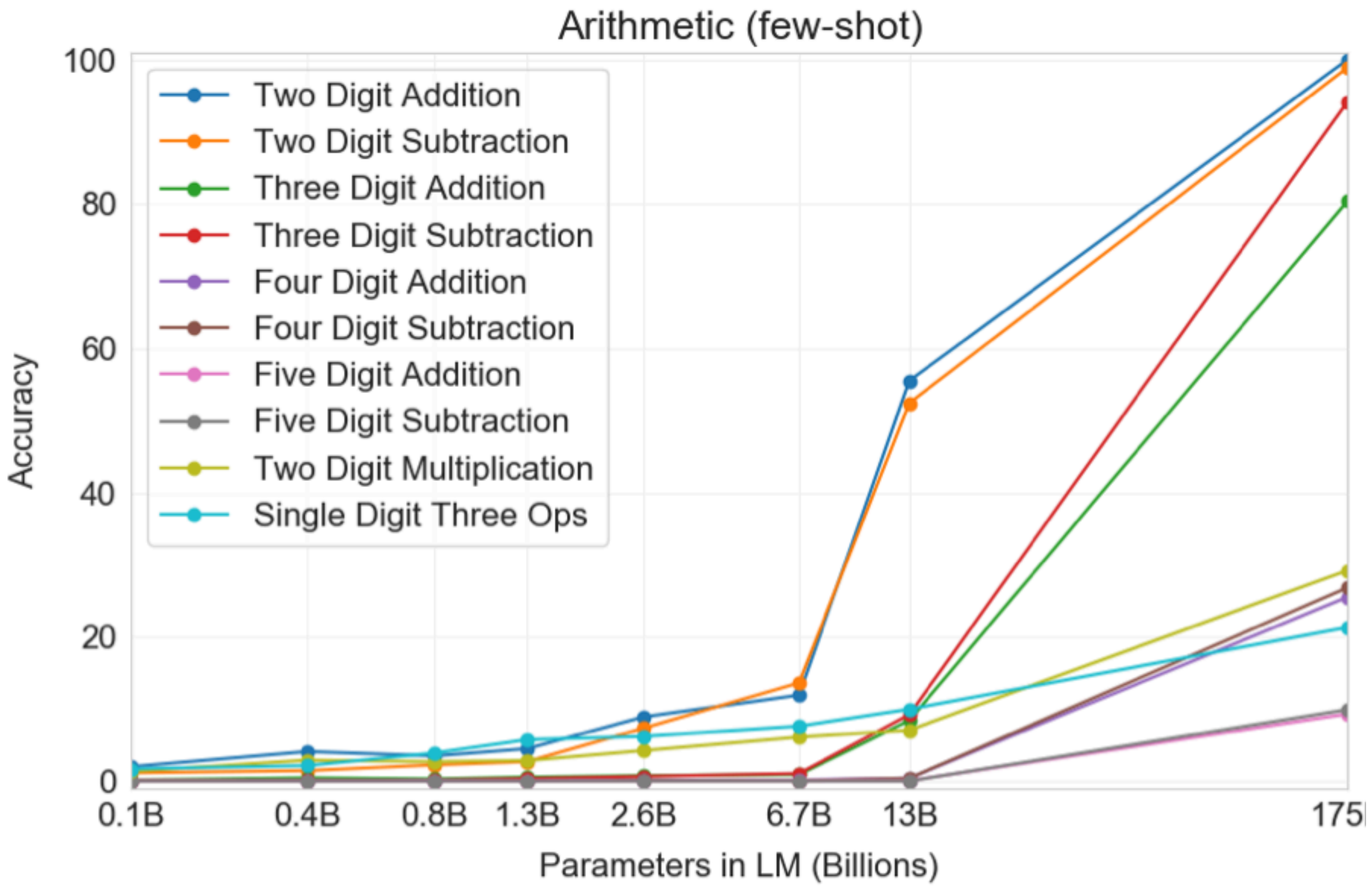
Context →	Q: What is 6209 minus 3365? A:
Target Completion →	2844

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

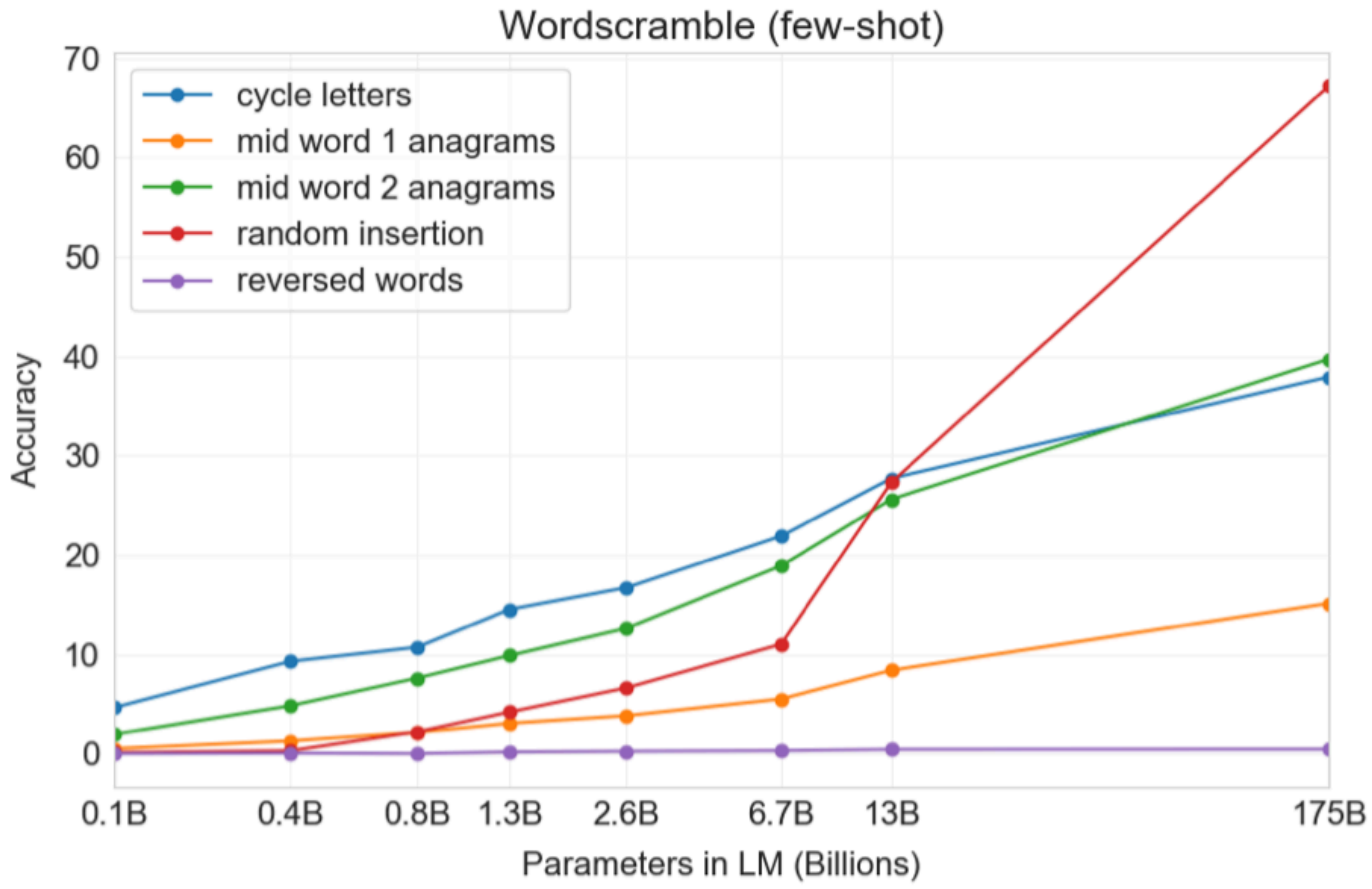
Context →	lull is to trust as
Correct Answer →	cajole is to compliance
Incorrect Answer →	balk is to fortitude
Incorrect Answer →	betray is to loyalty
Incorrect Answer →	hinder is to destination
Incorrect Answer →	soothe is to passion

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)

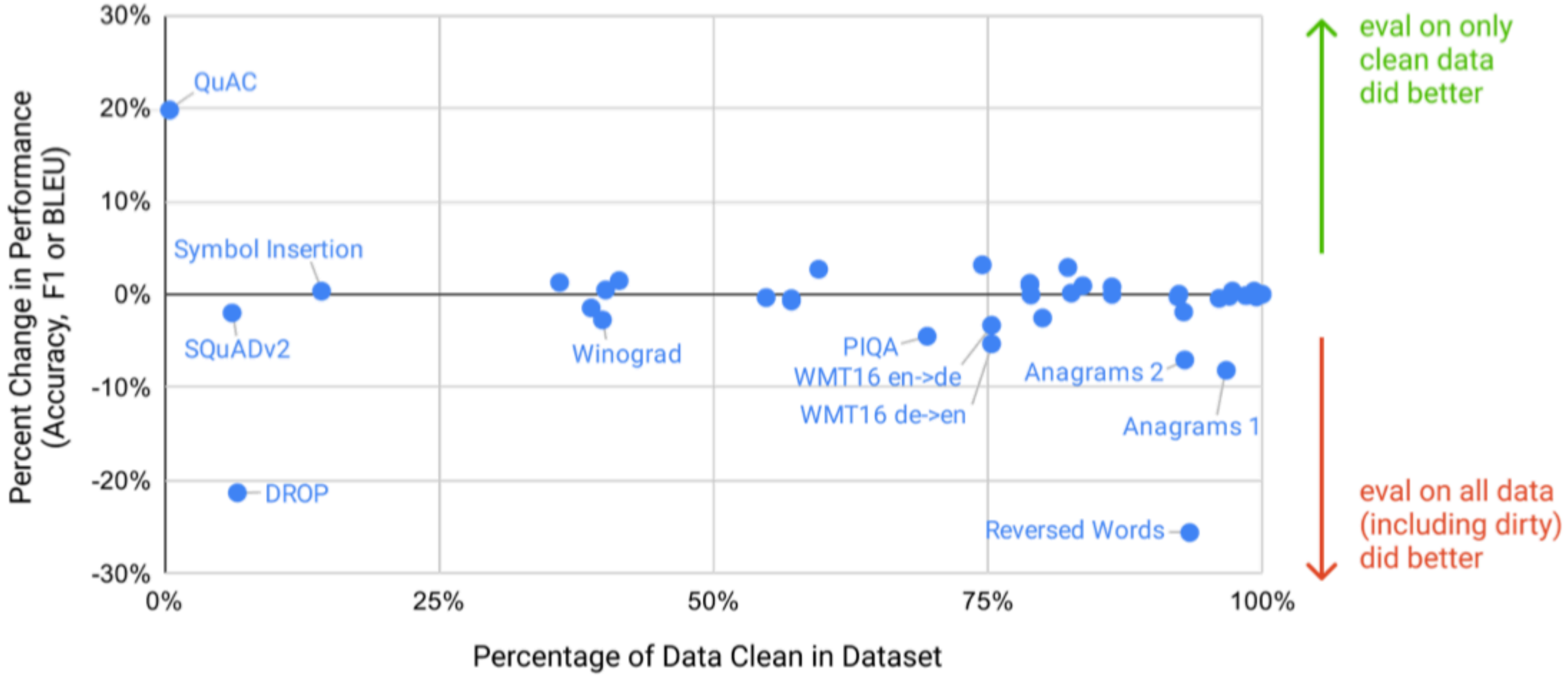


Wordscramble (few-shot)
cycle letters
mid word 1 anagrams
mid word 2 anagrams
random insertion
reversed words
Accuracy
0
10
20
30
40
50
60
70
0.1B
0.4B
0.8B
1.3B
2.6B
6.7B
13B
175B
Parameters in LM (Billions)

Contamination analysis

- How to decide which examples are contaminated?
  - "defined roughly as examples that have a 13-gram overlap with anything in the pretraining set"
- How to decide estimated performance gains from contamination?
  - Compare the performance on the “clean” subset vs entire dataset

A major methodological concern with language models pretrained on a broad swath of internet data, particularly large models with the capacity to memorize vast amounts of content, is potential contamination of downstream tasks by having their test or development sets inadvertently seen during pre-training. To reduce such contamination, we searched for and attempted to remove any overlaps with the development and test sets of all benchmarks studied in this paper. Unfortunately, a bug in the filtering caused us to ignore some overlaps, and due to the cost of training it was not feasible to retrain the model. In Section 4 we characterize the impact of the remaining overlaps, and in future work we will more aggressively remove data contamination.



Understanding in-context learning

Extrapolating to Unnatural Language Processing with GPT-3's In-context Learning: The Good, the Bad, and the Mysterious

Frieda Rong

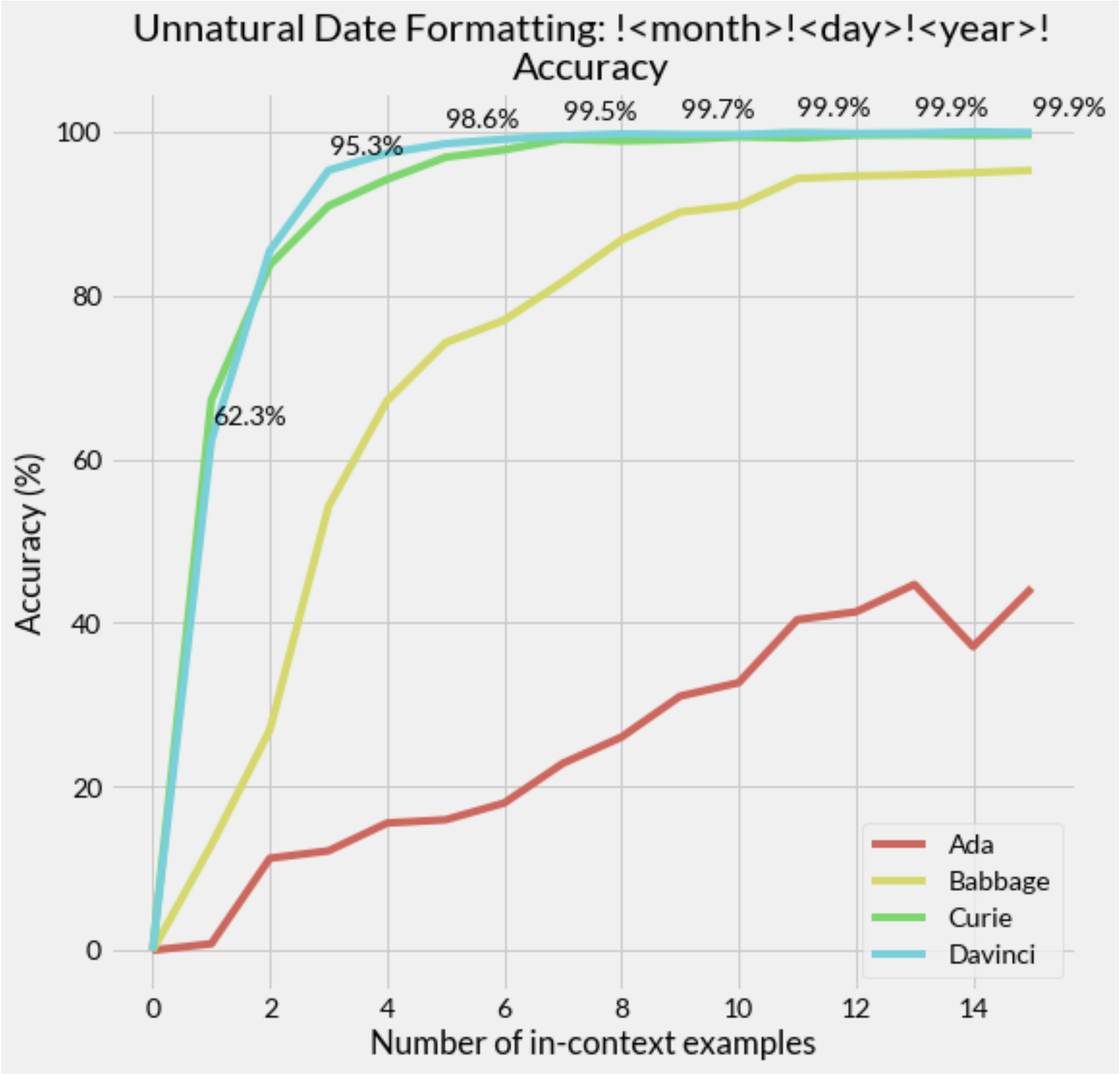
May 28, 2021

```
Input: 2014-06-01
Output: !06!01!2014!
Input: 2007-12-13
Output: !12!13!2007!
Input: 2010-09-23
Output: !09!23!2010!
Input: 2005-07-23
Output: !07!23!2005!
```

in-context examples

test example

model completion



http://ai.stanford.edu/blog/in-context-learning/

Understanding in-context learning

- **Hypothesis #1**: Transformers perform implicit gradient descent to update an “inner model”

Transformers Learn In-Context by Gradient Descent

Johannes von Oswald [1 2] Eyvind Niklasson [2] Ettore Randazzo [2] João Sacramento [1]
Alexander Mordvintsev [2] Andrey Zhmoginov [2] Max Vladymyrov [2]

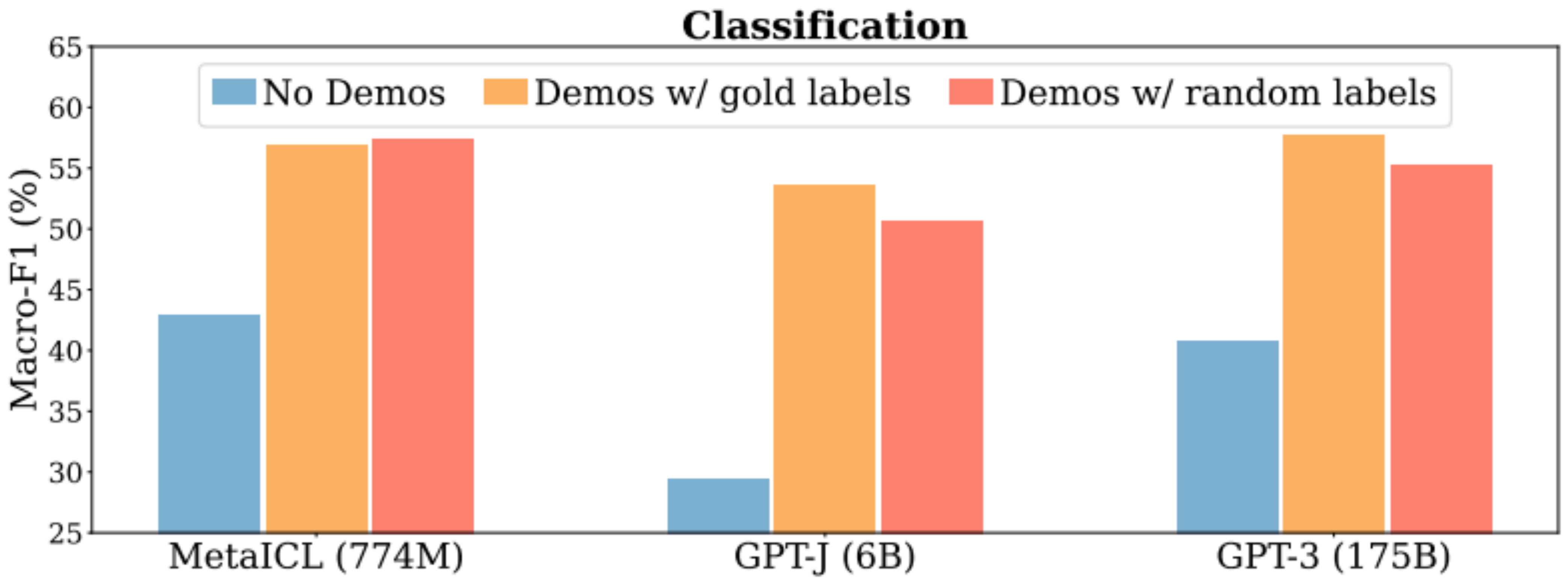
Why Can GPT Learn In-Context?
Language Models Implicitly Perform Gradient Descent as Meta-Optimizers

Damai Dai[†*], Yutao Sun[‖*], Li Dong[‡], Yaru Hao[‡], Shuming Ma[‡], Zhifang Sui[†], Furu Wei[‡]

- **Hypothesis #2**: Transformers learn tasks required for downstream applications during pre-training, and in-context demonstrations are only used to recognize which task is required

Rethinking the Role of Demonstrations:
What Makes In-Context Learning Work?

Sewon Min[1,2] Xinxi Lyu[1] Ari Holtzman[1] Mikel Artetxe[2]
Mike Lewis[2] Hannaneh Hajishirzi[1,3] Luke Zettlemoyer[1,2]
[1]University of Washington [2]Meta AI [3]Allen Institute for AI
{sewon,alrope,ahai,hannaneh,lsz}@cs.washington.ed
{artetxe,mikelewis}@meta.com



Ground-truth labels don’t matter!

Understanding in-context learning

We disentangle In-context learning into two roles - **task recognition (TR)** vs **task learning (TL)**

- TR: recognizes the task from demonstrations and applies LLMs' pre-trained priors
- TL: learns a new input-label mapping from demonstrations
- ICL performs both TR and TL, but TL emerges with **larger models** and **more demonstrations**

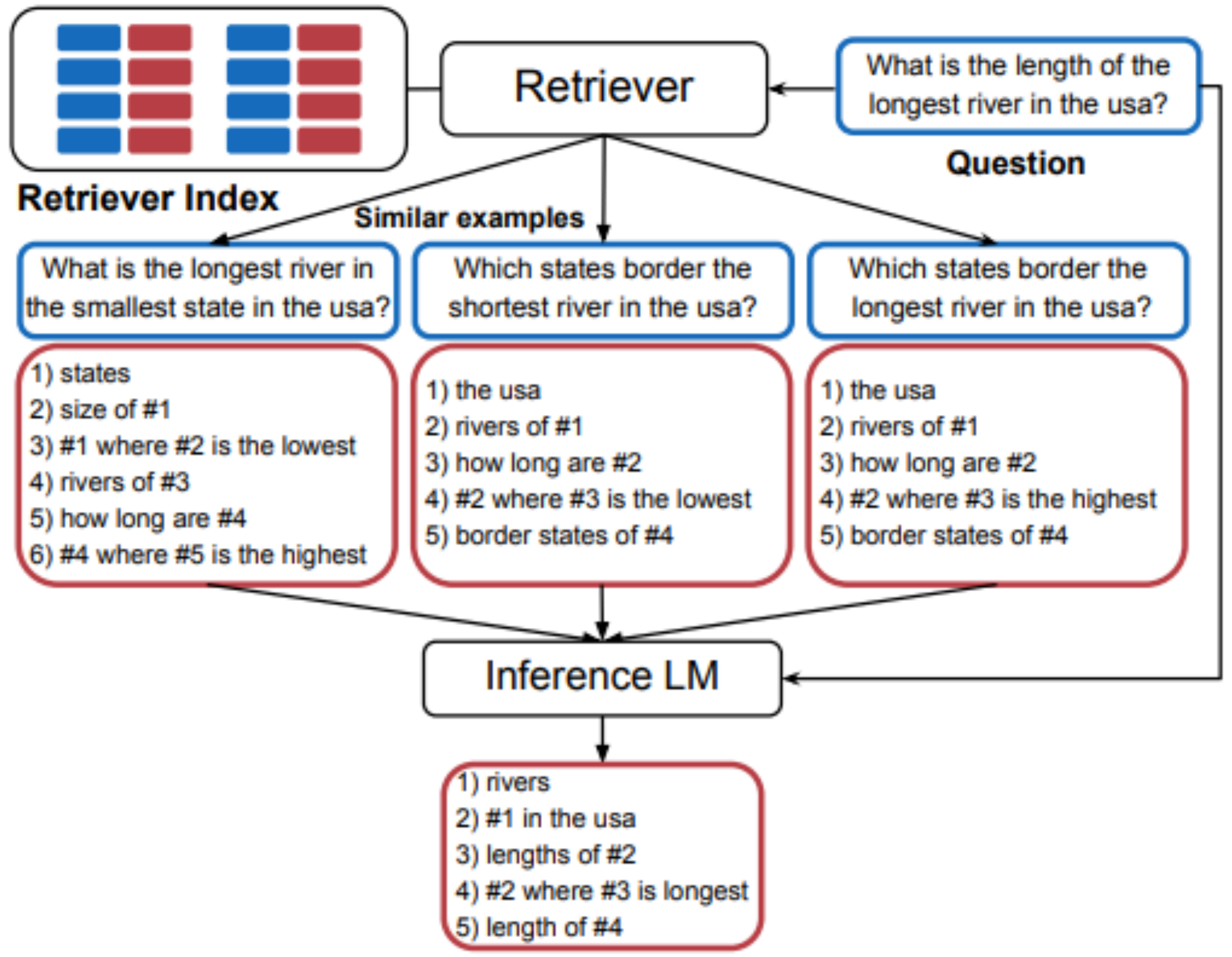
What In-Context Learning "Learns" In-Context: Disentangling Task Recognition and Task Learning

Improving in-context learning performance

- Instead of randomly sampling K in-context examples, you should use "high-quality" and similar ones!

Learning To Retrieve Prompts for In-Context Learning

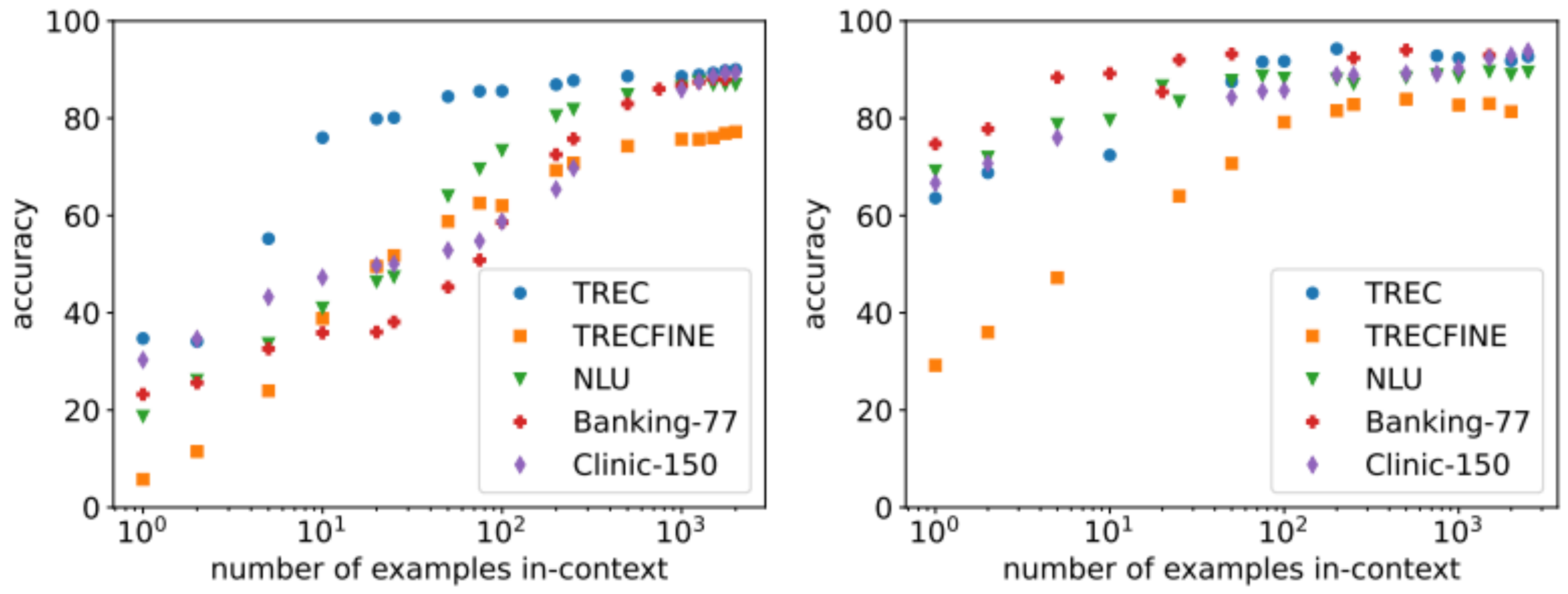
Ohad Rubin **Jonathan Herzig** **Jonathan Berant**
The Blavatnik School of Computer Science, Tel Aviv University
{ohad.rubin,jonathan.herzig,joberant}@cs.tau.ac.il



- Pack more examples in long-context models!

In-Context Learning with Long-Context Models: An In-Depth Exploration

Amanda Bertsch[γ] abertsch@cs.cmu.edu
Maor Ivgi[τ] maor.ivgi@cs.tau.ac.il
Uri Alon[γ*] urialon@cs.cmu.edu
Jonathan Berant[τ] joberant@cs.tau.ac.il
Matthew R. Gormley[γ] mgormley@cs.cmu.edu
Graham Neubig[γ] gneubig@cs.cmu.edu



(a) Using randomly selected examples. (b) Using retrieved examples.

Llama 3

From GPT-3 to Llama 3

- GPT-1, GPT-2, GPT-3, GPT-3.5/ChatGPT, GPT-4, GPT-4-turbo, GPT-4o
- Llama 1, Llama 2, Llama 3
- Mistral, Mixtral
- Claude 1, Claude 2, Claude 3, Claude 3.5 (Haiku, Sonnet, Opus)
- Qwen 1, Qwen 2
- Bard, Gemini, Gemini Pro, Gemma 1, Gemma 2
- …
- Truly open LMs: OLMo, Pythia, BLOOM

Llama 3.1: overview

- **Dense Transformers** - 8B, 70B, 405B
  - Dense vs mixture-of-experts
  - Smaller models are getting more attention
- **Long-context:** 128K tokens (remember, GPT-3 had only 2048 tokens)
- **Pre-trained** on 15T multilingual tokens (remember, GPT-3 was trained on 300B tokens)
- Pre-training vs **post-training:**
  - SFT, rejection sampling, direct preference optimization
  - multilinguality, coding, reasoning, tool use
  - Safety mitigations: helpfulness vs harmlessness
- **Multi-modal** training and adaptation

Pre-training data

- "To train the best language model, the curation of a large, high-quality training dataset is paramount."
- PII and safety filtering
- Text extraction and cleaning from raw HTML pages
- De-duplication: URL, document, line-level, …
- **Heuristic filtering**:
  - Remove lines that consist of repeated content (e.g., n-gram coverage ratio)
  - Dirty word counting
  - KL divergence of token-distribution compared "high-quality corpus"
- **Model-based quality classifier:** important and new trend!
- **Code, reasoning, and multilingual** data

Heuristic filtering

C4 rules (Raffel et al., 2020)

- We only retained lines that ended in a terminal punctuation mark (i.e. a period, exclamation mark, question mark, or end quotation mark).
- We discarded any page with fewer than 5 sentences and only retained lines that contained at least 3 words.
- We removed any page that contained any word on the "List of Dirty, Naughty, Obscene or Otherwise Bad Words".[6]
- Many of the scraped pages contained warnings stating that Javascript should be enabled so we removed any line with the word Javascript.
- Some pages had placeholder "lorem ipsum" text; we removed any page where the phrase "lorem ipsum" appeared.
- Some pages inadvertently contained code. Since the curly bracket "{" appears in many programming languages (such as Javascript, widely used on the web) but not in natural text, we removed any pages that contained a curly bracket.

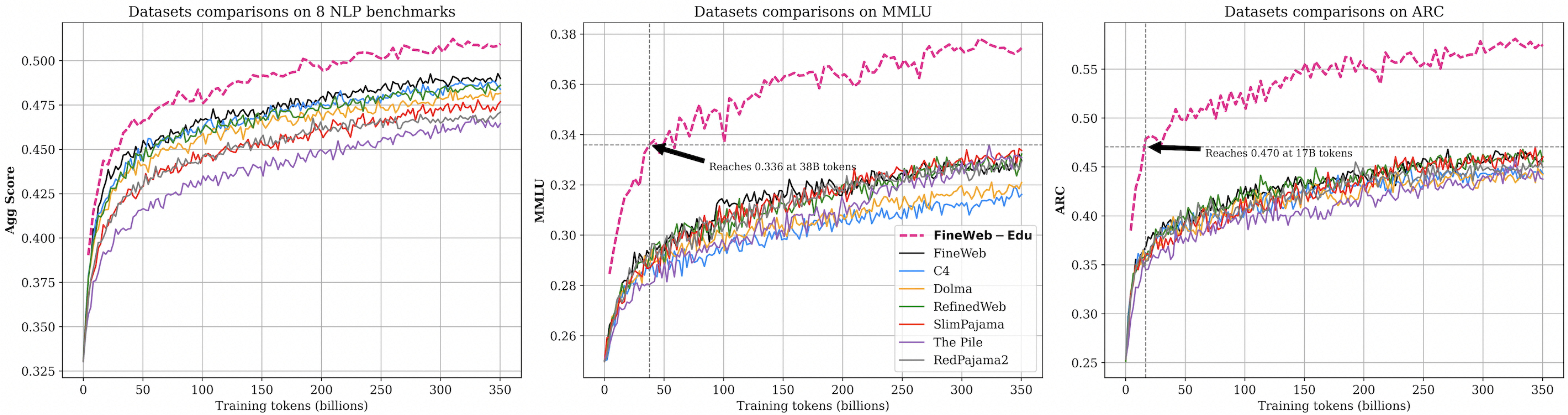
Gopher Rules (Rae et al., 2021)

```python
def gopher_rules_pass(sample) -> bool:
    """ function returns True if the sample complies with Gopher rules """
    signals = json.loads(sample["quality_signals"])

    # rule 1: number of words between 50 and 10'000
    word_count = signals["rps_doc_word_count"][0][2]
    if word_count < 50 or word_count > 10_000:
        return False

    # rule 2: mean word length between 3 and 10
    mean_word_length = signals["rps_doc_mean_word_length"][0][2]
    if mean_word_length < 3 or mean_word_length > 10:
        return False

    # rule 2: symbol to word ratio below 0.1
    symbol_word_ratio = signals["rps_doc_symbol_to_word_ratio"][0][2]
    if  symbol_word_ratio > 0.1:
        return False

    # rule 3: 90% of lines need to start without a bullet point
    n_lines = signals["ccnet_nlines"][0][2]
    n_lines_bulletpoint_start = sum(map(lambda ln: ln[2], signals["rps_lines_start_w:
    if n_lines_bulletpoint_start / n_lines > 0.9:
        return False

    # rule 4: the ratio between characters in the most frequent 2-gram and the total
    # of characters must be below 0.2
    top_2_gram_frac = signals["rps_doc_frac_chars_top_2gram"][0][2]
    if top_2_gram_frac > 0.2:
        return False

    # rule 5: ...
```

Model-based quality filtering

"To train a quality classifier based on **Llama 2**, we create a training set of cleaned web documents, **describe the quality requirements**, and **instruct Llama 2's chat model to determine if the documents meets these requirements**. We use **DistilRoberta** (Sanh et al., 2019) to generate quality scores for each document for efficiency reasons. We experimentally evaluate the efficacy of various quality filtering configurations."

FINEWEB-EDU They generate **450k annotations** by **llama-3-instruct** for identifying educational content



Model-based quality filtering

Below is an extract from a web page. Evaluate whether the page has a high educational value and could be useful in an educational setting for teaching from primary school to grade school levels using the additive 5-point scoring system described below. Points are accumulated based on the satisfaction of each criterion:

- Add 1 point if the extract provides some basic information relevant to educational topics, even if it includes some irrelevant or non-academic content like advertisements and promotional material.
- Add another point if the extract addresses certain elements pertinent to education but does not align closely with educational standards. It might mix educational content with non-educational material, offering a superficial overview of potentially useful topics, or presenting information in a disorganized manner and incoherent writing style.
- Award a third point if the extract is appropriate for educational use and introduces key concepts relevant to school curricula. It is coherent though it may not be comprehensive or could include some extraneous information. It may resemble an introductory section of a textbook or a basic tutorial that is suitable for learning but has notable limitations like treating concepts that are too complex for grade school students.
- Grant a fourth point if the extract highly relevant and beneficial for educational purposes for a level not higher than grade school, exhibiting a clear and consistent writing style. It could be similar to a chapter from a textbook or a tutorial, offering substantial educational content, including exercises and solutions, with minimal irrelevant information, and the concepts aren't too advanced for grade school students. The content is coherent, focused, and valuable for structured learning.
- Bestow a fifth point if the extract is outstanding in its educational value, perfectly suited for teaching either at primary school or grade school. It follows detailed reasoning, the writing style is easy to follow and offers profound and thorough insights into the subject matter, devoid of any non-educational or complex content.

The extract: <extract>.

After examining the extract:

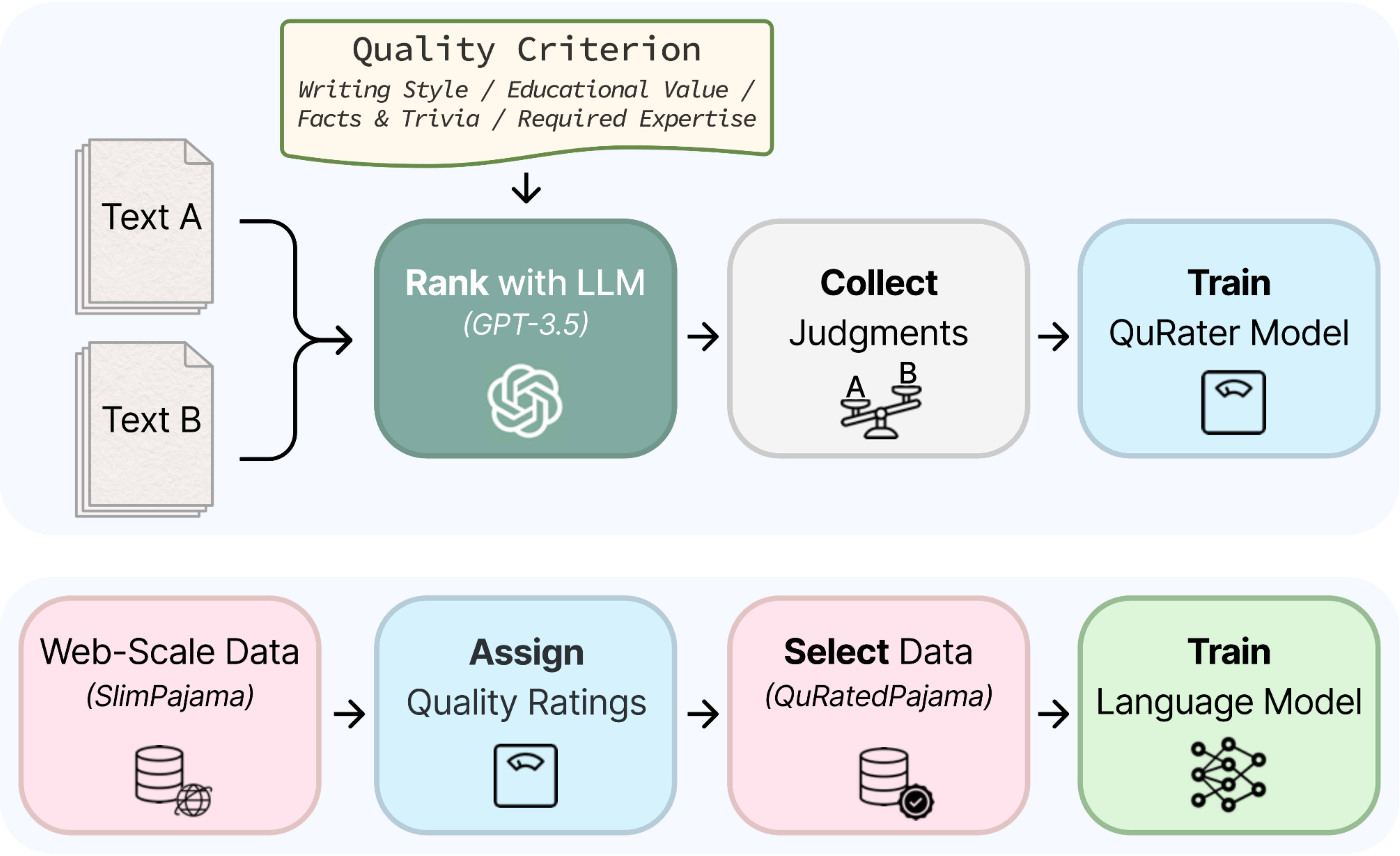
- Briefly justify your total score, up to 100 words.
- Conclude with the score using the format: "Educational score: <total points>"

https://huggingface.co/datasets/HuggingFaceFW/fineweb-edu

QuRating: Selecting high-quality data with LM signals

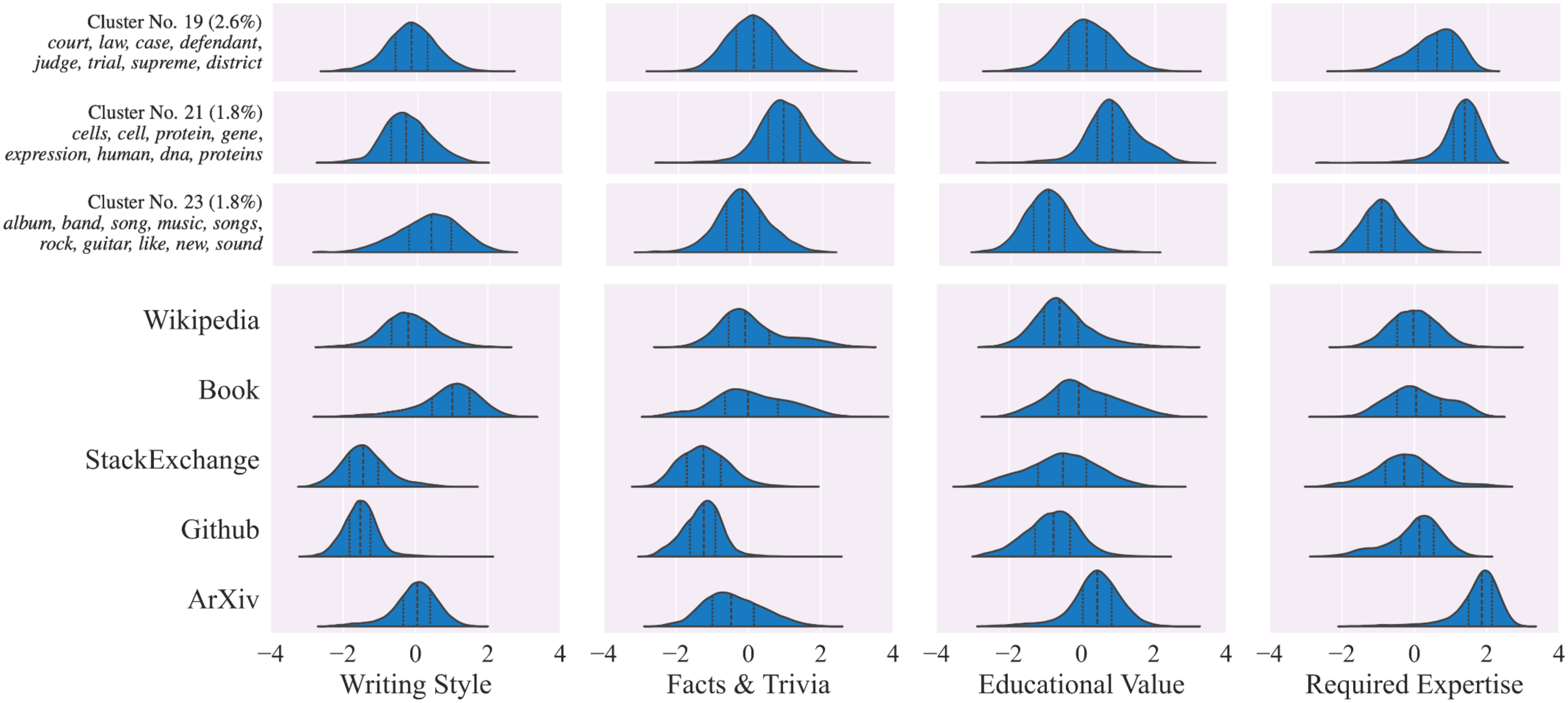
Part I
measure quality

Part II
utilize quality



QuRating: Selecting High-Quality Data for Training Language Models (2024)

QuRating: Selecting high-quality data with LM signals



QuRating: Selecting High-Quality Data for Training Language Models (2024)

Code and math data

- Common wisdom: code and math data are very important for pre-training
- They build **domain-specific pipelines** that extract code and math-relevant web pages

Published as a conference paper at ICLR 2024

AT WHICH TRAINING STAGE DOES CODE DATA HELP LLMS REASONING?

Yingwei Ma[1,2]*, Yue Liu[1]*, Yue Yu[1,2]†, Yuanliang Zhang[1], Yu Jiang[3], Changjian Wang[1], Shanshan Li[1]†
[1]National University of Defense Technology
[2]Peng Cheng Laboratory
[3]Tsinghua University

To Code, or Not To Code? Exploring Impact of Code in Pre-training

Viraat Aryabumi[1], Yixuan Su[2], Raymond Ma[2], Adrien Morisot[2], Ivan Zhang[2], Acyr Locatelli[2], Marzieh Fadaee[1], Ahmet Üstün[1], and Sara Hooker[1]

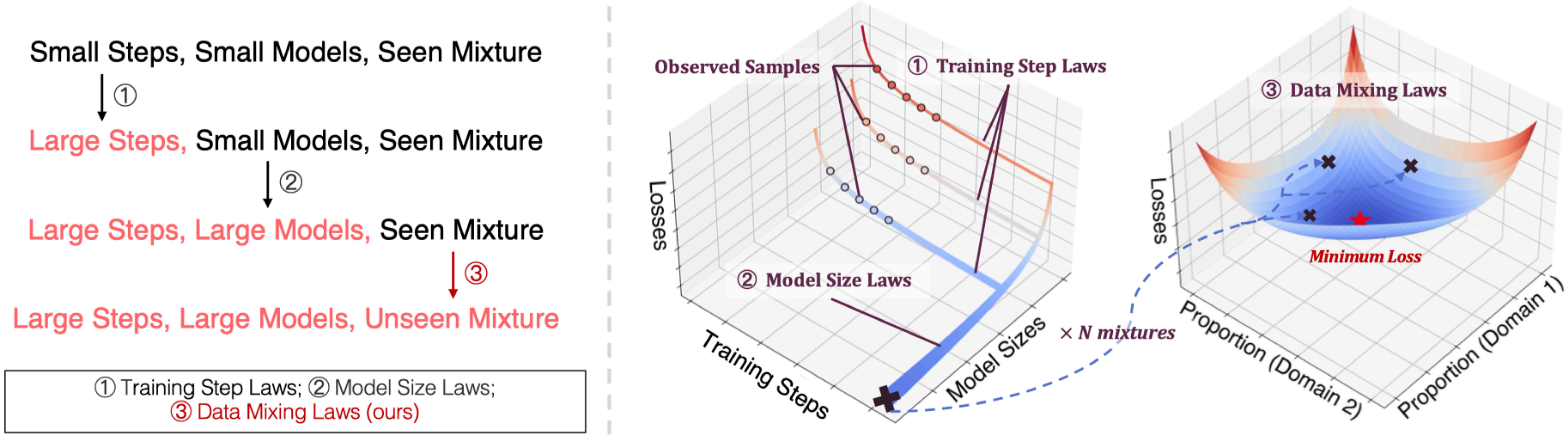
[1]Cohere For AI, [2]Cohere

- Code is a critical building block for generalization far beyond coding tasks
  - Compared to text-only pre-training, 8.2% in NL reasoning, 4.2% in world knowledge, 6.6% in general win rates, 12x in code performance
- The quality of code data has an outsized impact in downstream tasks

Determining data mix

"Roughly **50%** of tokens corresponding to general knowledge, **25%** of mathematical and reasoning tokens, **17%** code tokens, **8%** multilingual tokens"

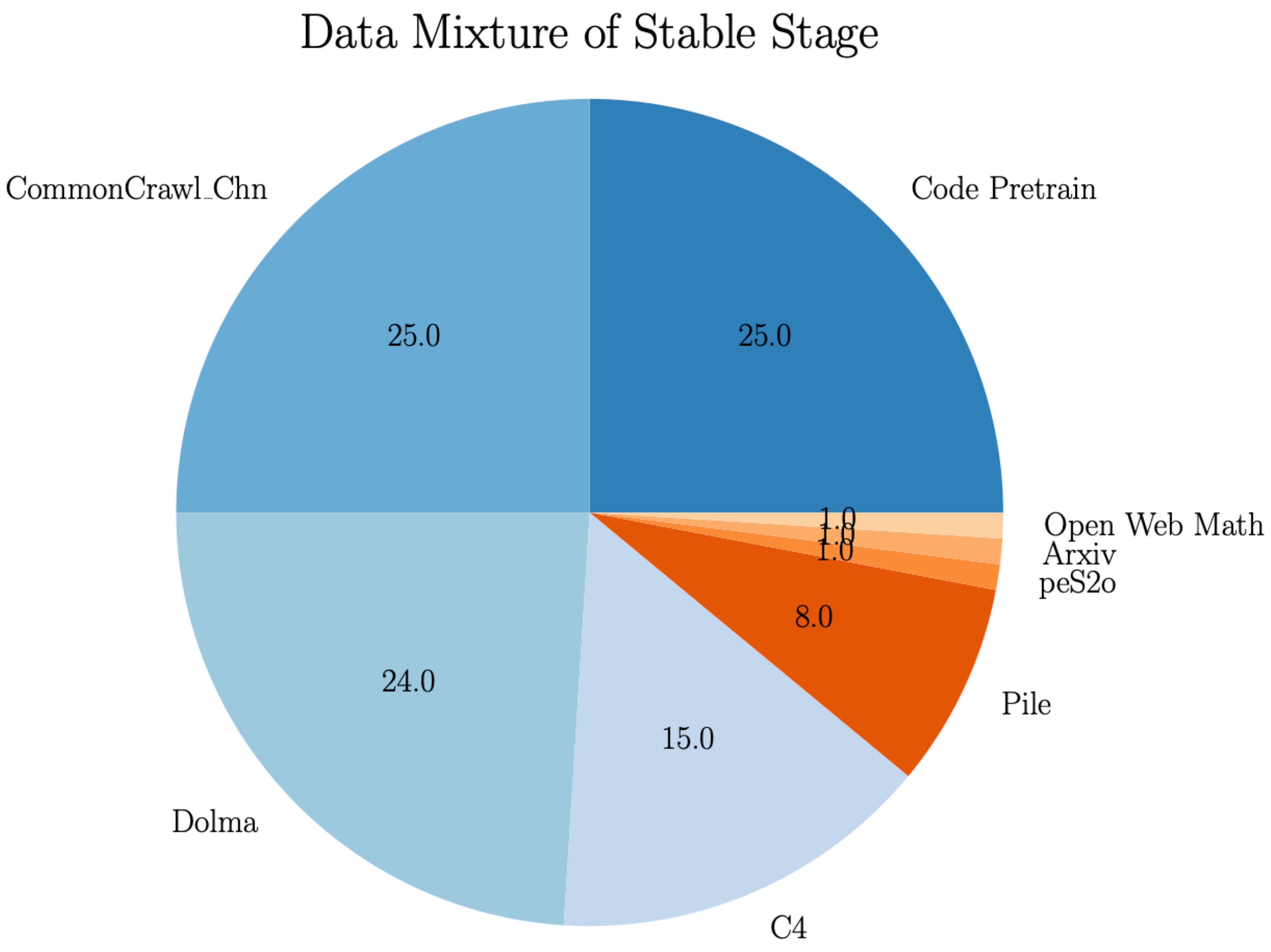
- **Scaling laws for data mix**: "train several smaller models on a data mix and use that to predict the performance on that mix", "repeat this process for different data mixes to select a new data mix candidate"

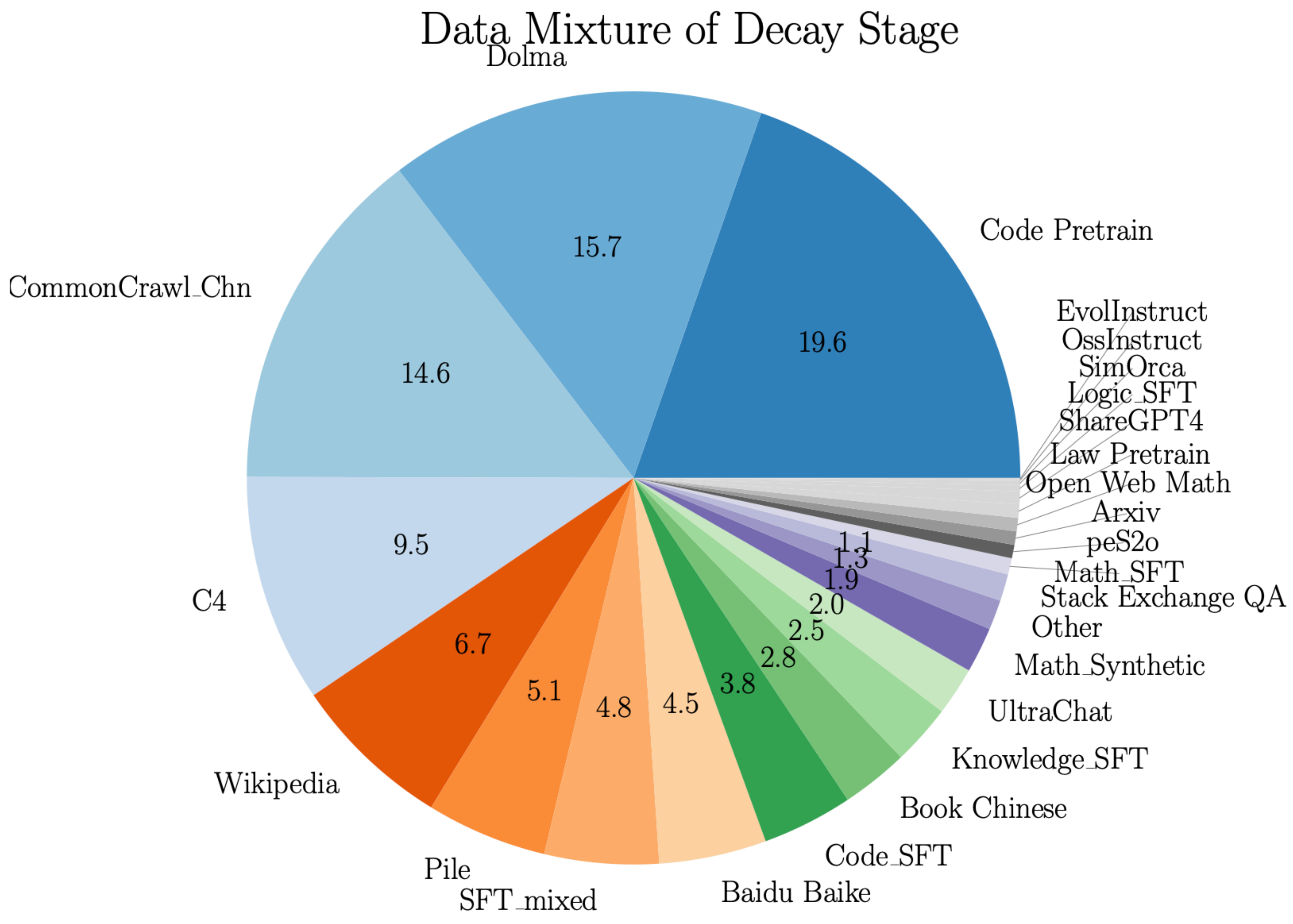


Data Mixing Laws: Optimizing Data Mixtures by Predicting Language Modeling Performance (2024)

Determining data mix

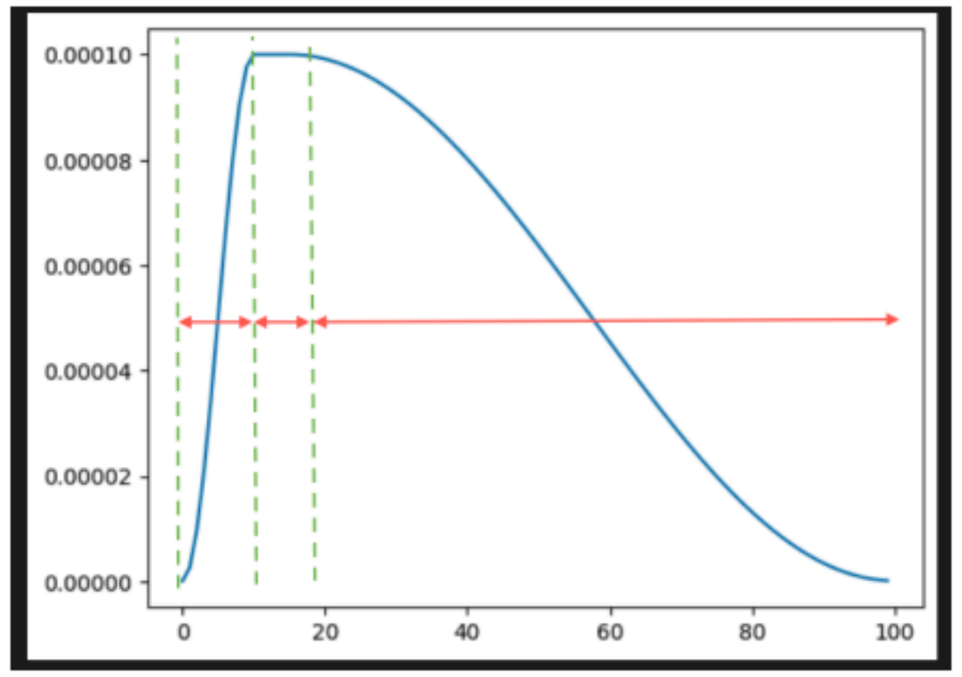
Domains: Common Crawl, CC, Github, Wikipedia, Books, arXiv, …





Training recipe

- AdamW: learning rate of 8×10^{-5}, a linear warm up of 8000 steps, and a cosine learning rate schedule decaying to 8×10^{-7} over 1,200,000 steps
- They adjusted the pre-training mix during training
  - Increased percentage of non-English data
  - Upsample mathematical data to improve the model's knowledge cut-off
  - Downsampled subsets of pre-training data that were later identified as lower quality
- **Long-context pre-training**: first train on 8k, and increase context length to 128k in six stages (800B training tokens)
  - Challenges: scarcity of real long-context pre-training data
  - The performance on short-context tasks will degrade drastically



Cosine LR schedule with linear warmup

Data annealing

- They upsample on data sources of very high-quality at the end of training (final 40M tokens; no benchmark datasets used in annealing)

Does your data spark joy? Performance gains from domain upsampling at the end of training

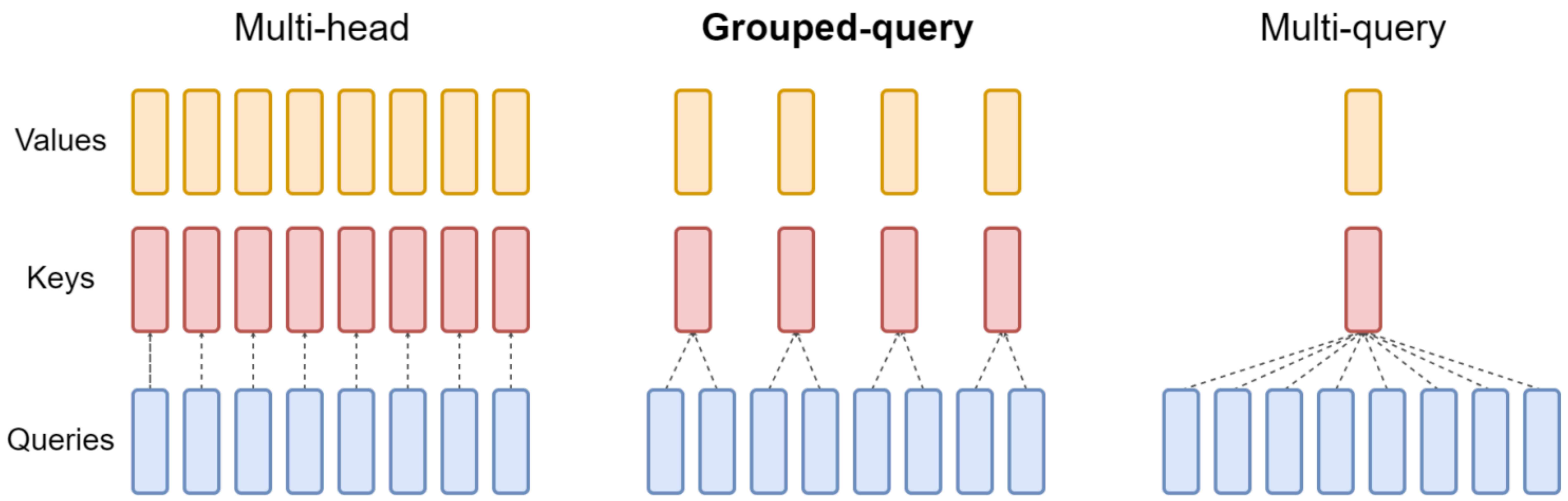
Cody Blakeney*, Mansheej Paul*, Brett W. Larsen*, Sean Owen, and Jonathan Frankle

Databricks Mosaic Research

- They view data annealing as a cheap way to measure the impact of domain-specific datasets on model capabilities

Model architecture

- Standard dense Transformers, the same architecture as Llama-2
- **Grouped query attention (GQA):** 8 key-value heads to improve inference speed

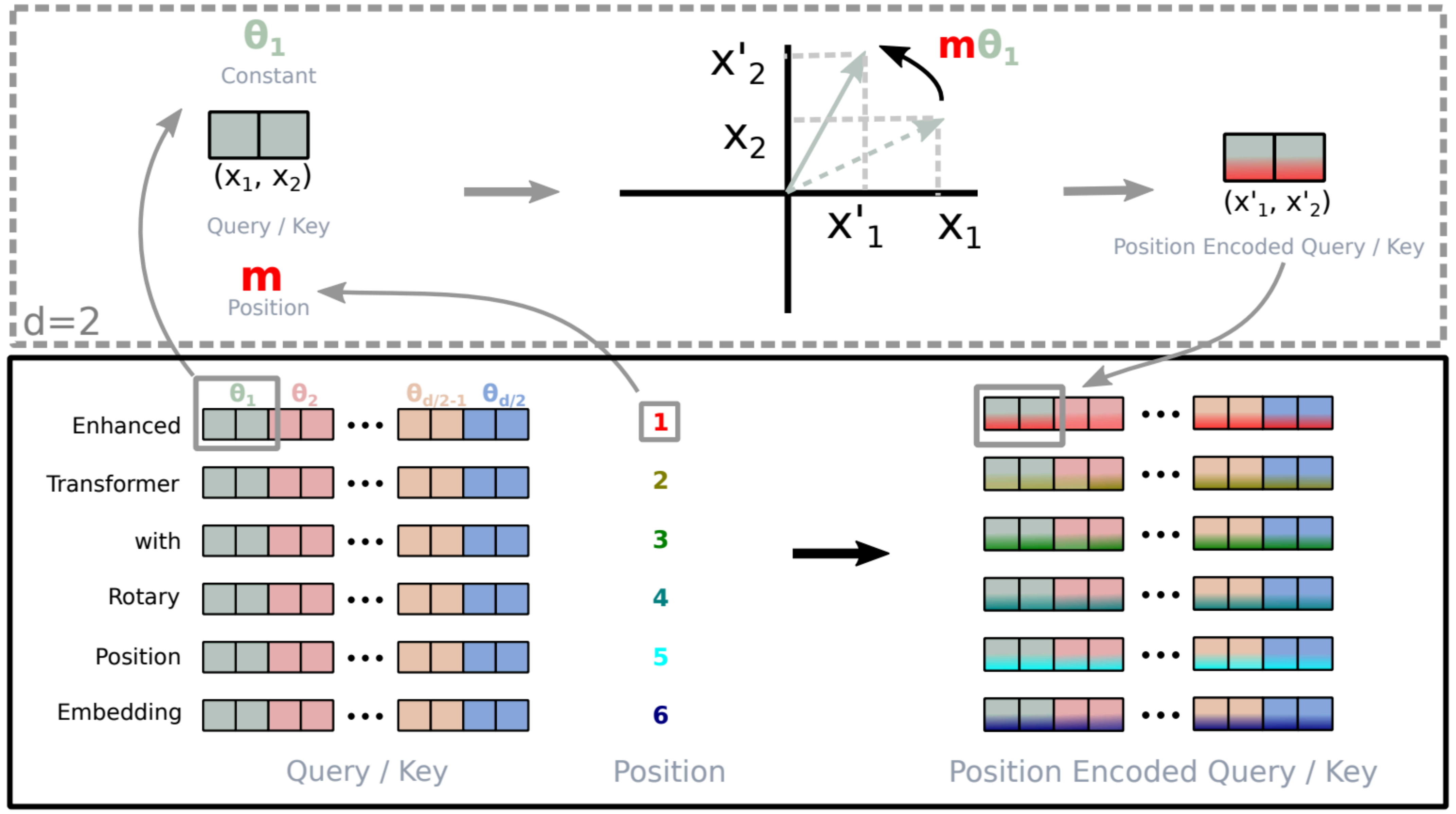


(Ainslie et al., 2023) GQA: Training generalized multi-query transformer models from multi-head checkpoints.

Model architecture

- Standard dense Transformers, the same architecture as Llama-2
- **Grouped query attention (GQA):** 8 key-value heads to improve inference speed
- Prevents self-attention between documents within the same sequence
- A much larger vocabulary: 128K
- **RoPE positional embeddings**: base frequency = 500,000

Rope positional embeddings



$$\theta_i = 10000^{-2(i-1)/d}, i \in [1, 2, ..., d/2]$$

Base frequency

$$f_{\{q,k\}}(\boldsymbol{x}_m, m) = \begin{pmatrix} \cos m\theta & -\sin m\theta \\ \sin m\theta & \cos m\theta \end{pmatrix} \begin{pmatrix} W^{(11)}_{\{q,k\}} & W^{(12)}_{\{q,k\}} \\ W^{(21)}_{\{q,k\}} & W^{(22)}_{\{q,k\}} \end{pmatrix} \begin{pmatrix} x^{(1)}_m \\ x^{(2)}_m \end{pmatrix}$$

$$f_{\{q,k\}}(\boldsymbol{x}_m, m) = \boldsymbol{R}^d_{\Theta,m} \boldsymbol{W}_{\{q,k\}} \boldsymbol{x}_m$$

where

$$\boldsymbol{R}^d_{\Theta,m} = \begin{pmatrix} \cos m\theta_1 & -\sin m\theta_1 & 0 & 0 & \cdots & 0 & 0 \\ \sin m\theta_1 & \cos m\theta_1 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & \cos m\theta_2 & -\sin m\theta_2 & \cdots & 0 & 0 \\ 0 & 0 & \sin m\theta_2 & \cos m\theta_2 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & \cos m\theta_{d/2} & -\sin m\theta_{d/2} \\ 0 & 0 & 0 & 0 & \cdots & \sin m\theta_{d/2} & \cos m\theta_{d/2} \end{pmatrix}$$

(Su et al., 2021) RoFormer: Enhanced Transformer with Rotary Position Embedding

Evaluation

Reading Comprehension	SQuAD V2 (Rajpurkar et al., 2018), QuaC (Choi et al., 2018), RACE (Lai et al., 2017),
Code	HumanEval (Chen et al., 2021), MBPP (Austin et al., 2021),
Commonsense reasoning/understanding	CommonSenseQA (Talmor et al., 2019), PiQA (Bisk et al., 2020), SiQA (Sap et al., 2019), OpenBookQA (Mihaylov et al., 2018), WinoGrande (Sakaguchi et al., 2021)
Math, reasoning, and problem solving	GSM8K (Cobbe et al., 2021), MATH (Hendrycks et al., 2021b), ARC Challenge (Clark et al., 2018), DROP (Dua et al., 2019), WorldSense (Benchekroun et al., 2023)
Adversarial	Adv SQuAD (Jia and Liang, 2017), Dynabench SQuAD (Kiela et al., 2021), GSM-Plus (Li et al., 2024c) PAWS (Zhang et al., 2019)
Long context	QuALITY (Pang et al., 2022), many-shot GSM8K (An et al., 2023a)
Aggregate	MMLU (Hendrycks et al., 2021a), MMLU-Pro (Wang et al., 2024b), AGIEval (Zhong et al., 2023), BIG-Bench Hard (Suzgun et al., 2023)

Performance: reading comprehension

	Reading Comprehension		
	SQuAD	QuAC	RACE
Llama 3 8B	77.0 ±0.8	**44.9** ±1.1	**54.3** ±1.4
Mistral 7B	73.2 ±0.8	44.7 ±1.1	53.0 ±1.4
Gemma 7B	**81.8** ±0.7	42.4 ±1.1	48.8 ±1.4
Llama 3 70B	81.8 ±0.7	**51.1** ±1.1	59.0 ±1.4
Mixtral 8×22B	**84.1** ±0.7	44.9 ±1.1	**59.2** ±1.4
Llama 3 405B	**81.8** ±0.7	**53.6** ±1.1	**58.1** ±1.4
GPT-4	–	–	–
Nemotron 4 340B	–	–	–
Gemini Ultra	–	–	–

	Math and Reasoning	
	ARC-C	DROP
Llama 3 8B	**79.7** ±2.3	**59.5** ±1.0
Mistral 7B	78.2 ±2.4	53.0 ±1.0
Gemma 7B	78.6 ±2.4	56.3 ±1.0
Llama 3 70B	**92.9** ±1.5	**79.6** ±0.8
Mixtral 8×22B	91.9 ±1.6	77.5 ±0.8
Llama 3 405B	96.1 ±1.1	**84.8** ±0.7
GPT-4	**96.3** ±1.1	80.9 ±0.8
Nemotron 4 340B	94.3 ±1.3	–
Gemini Ultra	–	$82.4^{\triangle}$ ±0.8

DROP: 3-shot, SQuAD: 1-shot, RACE: 0-shot, QuAC: 1-shot, ARC-C: 25-shot..

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

Setting	ARC (Challenge)
Fine-tuned SOTA	**78.5**[KKS+20]
GPT-3 Zero-Shot	51.4
GPT-3 One-Shot	53.2
GPT-3 Few-Shot	51.5

Performance: commonsense reasoning

	Commonsense Understanding		
	PiQA	OpenBookQA	Winogrande
Llama 3 8B	81.0 ±1.8	45.0 ±4.4	75.7 ±2.0
Mistral 7B	**83.0** ±1.7	47.8 ±4.4	**78.1** ±1.9
Gemma 7B	81.5 ±1.8	**52.8** ±4.4	74.7 ±2.0
Llama 3 70B	83.8 ±1.7	47.6 ±4.4	83.5 ±1.7
Mixtral 8×22B	**85.5** ±1.6	**50.8** ±4.4	**84.7** ±1.7
Llama 3 405B	**85.6** ±1.6	**49.2** ±4.4	82.2 ±1.8
GPT-4	–	–	87.5 ±1.5
Nemotron 4 340B	–	–	**89.5** ±1.4

PiQA: 0-shot, OpenBookQA: 0-shot, Winogrande: 5-shot

Setting	PIQA	OpenBookQA
Fine-tuned SOTA	79.4	**87.2**[KKS+20]
GPT-3 Zero-Shot	**80.5***	57.6
GPT-3 One-Shot	**80.5***	58.8
GPT-3 Few-Shot	**82.8***	65.4

Setting	Winogrande (XL)
Fine-tuned SOTA	**84.6**[b]
GPT-3 Zero-Shot	70.2
GPT-3 One-Shot	73.2
GPT-3 Few-Shot	77.7

Performance: code and math

HUMANEVAL

```python
def incr_list(l: list):
    """Return list with elements incremented by 1.
    >>> incr_list([1, 2, 3])
    [2, 3, 4]
    >>> incr_list([5, 3, 5, 2, 3, 3, 9, 0, 123])
    [6, 4, 6, 3, 4, 4, 10, 1, 124]
    """
    return [i + 1 for i in l]
```

```python
def solution(lst):
    """Given a non-empty list of integers, return the sum of all of the odd elements
    that are in even positions.

    Examples
    solution([5, 8, 7, 1]) =⇒12
    solution([3, 3, 3, 3, 3]) =⇒9
    solution([30, 13, 24, 321]) =⇒0
    """
    return sum(lst[i] for i in range(0,len(lst)) if i % 2 == 0 and lst[i] % 2 == 1)
```

```python
def encode_cyclic(s: str):
    """
    returns encoded string by cycling groups of three characters.
    """
    # split string to groups. Each of length 3.
    groups = [s[(3 * i):min((3 * i + 3), len(s))] for i in range((len(s) + 2) // 3)]
    # cycle elements in each group. Unless group has fewer elements than 3.
    groups = [(group[1:] + group[0]) if len(group) == 3 else group for group in groups]
    return "".join(groups)

def decode_cyclic(s: str):
    """
    takes as input string encoded with encode_cyclic function. Returns decoded string.
    """
    # split string to groups. Each of length 3.
    groups = [s[(3 * i):min((3 * i + 3), len(s))] for i in range((len(s) + 2) // 3)]
    # cycle elements in each group.
    groups = [(group[-1] + group[:-1]) if len(group) == 3 else group for group in groups]
    return "".join(groups)
```

GSM8K

Problem

The battery charge in Mary's cordless vacuum cleaner lasts ten minutes. It takes her four minutes to vacuum each room in her house. Mary has three bedrooms, a kitchen, and a living room. How many times does Mary need to charge her vacuum cleaner to vacuum her whole house?

Solution

Mary has 3 + 1 + 1 = 5 rooms in her house.
At 4 minutes a room, it will take her 4 * 5 = 20 minutes to vacuum her whole house.
At 10 minutes a charge, she will need to charge her vacuum cleaner 20 / 10 = 2 times to vacuum her whole house.

Final Answer

2

Performance: code and math

	Code	
	HumanEval	MBPP
Llama 3 8B	**37.2** ±7.4	**47.6** ±4.4
Mistral 7B	30.5 ±7.0	47.5 ±4.4
Gemma 7B	32.3 ±7.2	44.4 ±4.4
Llama 3 70B	**58.5** ±7.5	66.2 ±4.1
Mixtral 8×22B	45.1 ±7.6	**71.2** ±4.0
Llama 3 405B	61.0 ±7.5	**73.4** ±3.9
GPT-4	67.0 ±7.2	–
Nemotron 4 340B	57.3 ±7.6	–
Gemini Ultra	**74.4** ±6.7	–

	GSM8K	MATH
Llama 3 8B	**57.2** ±2.7	20.3 ±1.1
Mistral 7B	52.5 ±2.7	13.1 ±0.9
Gemma 7B	46.4 ±2.7	**24.3** ±1.2
Llama 3 70B	83.7 ±2.0	41.4 ±1.4
Mixtral 8×22B	**88.4** ±1.7	**41.8** ±1.4
Llama 3 405B	89.0 ±1.7	**53.8** ±1.4
GPT-4	**92.0** ±1.5	–
Nemotron 4 340B	–	–
Gemini Ultra	88.9$^{\diamond}$ ±1.7	53.2±1.4

Contamination analysis

	Contam.	Performance gain est.		
		8B	70B	405B
AGIEval	98	8.5	19.9	16.3
BIG-Bench Hard	95	26.0	36.0	41.0
BoolQ	96	4.0	4.7	3.9
CommonSenseQA	30	0.1	0.8	0.6
DROP	–	–	–	–
GSM8K	41	0.0	0.1	1.3
HellaSwag	85	14.8	14.8	14.3
HumanEval	–	–	–	–
MATH	1	0.0	-0.1	-0.2
MBPP	–	–	–	–
MMLU	–	–	–	–
MMLU-Pro	–	–	–	–
NaturalQuestions	52	1.6	0.9	0.8
OpenBookQA	21	3.0	3.3	2.6
PiQA	55	8.5	7.9	8.1
QuaC	99	2.4	11.0	6.4
RACE	–	–	–	–
SiQA	63	2.0	2.3	2.6
SQuAD	0	0.0	0.0	0.0
Winogrande	6	-0.1	-0.1	-0.2
WorldSense	73	-3.1	-0.4	3.9

- How to decide which examples are contaminated?
  - "An example of a dataset D to be contaminated if a ratio T_D of its tokens are part of an 8-gram occurring at least once in the pre-training corpus”
- How to decide estimated performance gains from contamination?
  - Compare the performance on the “clean” subset vs entire dataset