

Lecture 2: Pre-training Data for LLMs and Evaluation of LLMs

Reinhard Heckel

Outline

- Popular open-source pre-training datasets and how they are obtained
- Discussion of popular LLM benchmarks

Training a GPT

Training sequence:

Longest_German_word_you_know



20, 30, 1, 12, 11, 1, 3, 1, 56, 1, 2



next token	predicted probability									
30	0.6	0
1	0	0.5
12	0.01

Loss: cross-entropy between coming token and predicted probability, for each position in the sequence

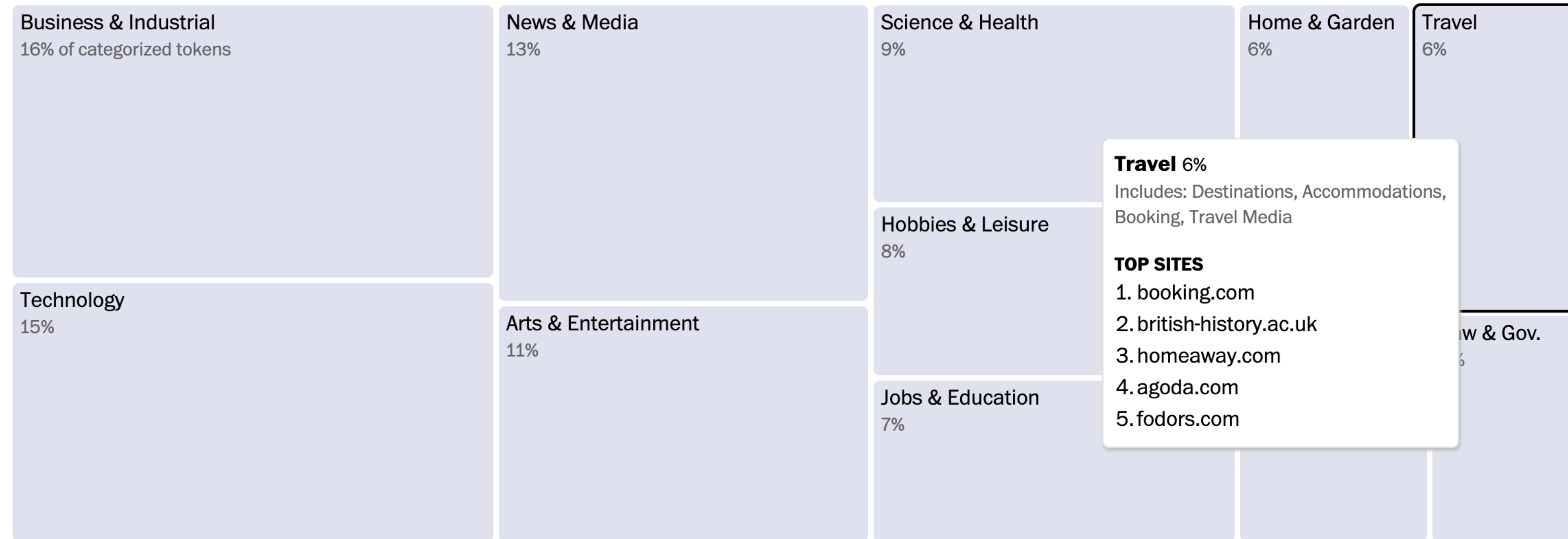
Language model training

stage	pre training	post training
method	next token prediction	next token prediction, reinforcement learning
data	huge amounts of high-quality text	few high quality examples and/or comparison data/and reward functions
outcome	base model: GPT, LLaMa	ChatGPT, ChatLLaMa, thinking models, agentic models

Pretraining data is often web filtered data

LLaMA pretraining data

Dataset	Sampling prop.	Epochs	Disk size
CommonCrawl	67.0%	1.10	3.3 TB
C4	15.0%	1.06	783 GB
Github	4.5%	0.64	328 GB
Wikipedia	4.5%	2.45	83 GB
Books	4.5%	2.23	85 GB
ArXiv	2.5%	1.06	92 GB
StackExchange	2.0%	1.03	78 GB



Touvron et al., 'LLaMA...', 2023

<https://www.washingtonpost.com/>

Majority of training data is filtered web-data

Curation of pretraining data


- Training data of closed (OpenAI, Anthropic) and many open models (Llama, Mistral, DeepSeek) is not public
 - Reasons: Data is critical ingredient and copyright concerns
- Curation of training data is critical for performance
- Curation strategies:
 - filtering
 - removing duplicates
 - mixing and weighting data sources
 - generating synthetic data



WebText and OpenWebText

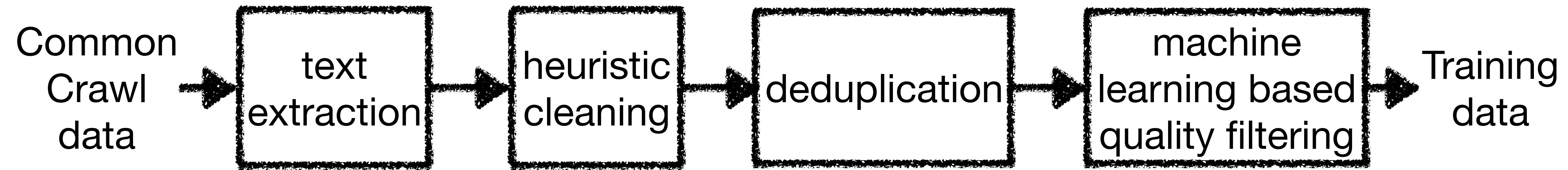
- Very early, OpenAI internal Dataset of millions of webpages that GPT2 has been trained on
- “We scraped all outbound links from Reddit, a social media platform, which received at least 3 karma. This can be thought of as a heuristic indicator for whether other users found the link interesting, educational, or just funny”
- Pages are de-duplicated and light-cleaned with some heuristics
- 8M documents, 40GB of text

Common Crawl maintains
a *free, open repository* of
web crawl data that can be
used by anyone.










Feb/March 2024 snapshot: ca. 100 TB of compressed data

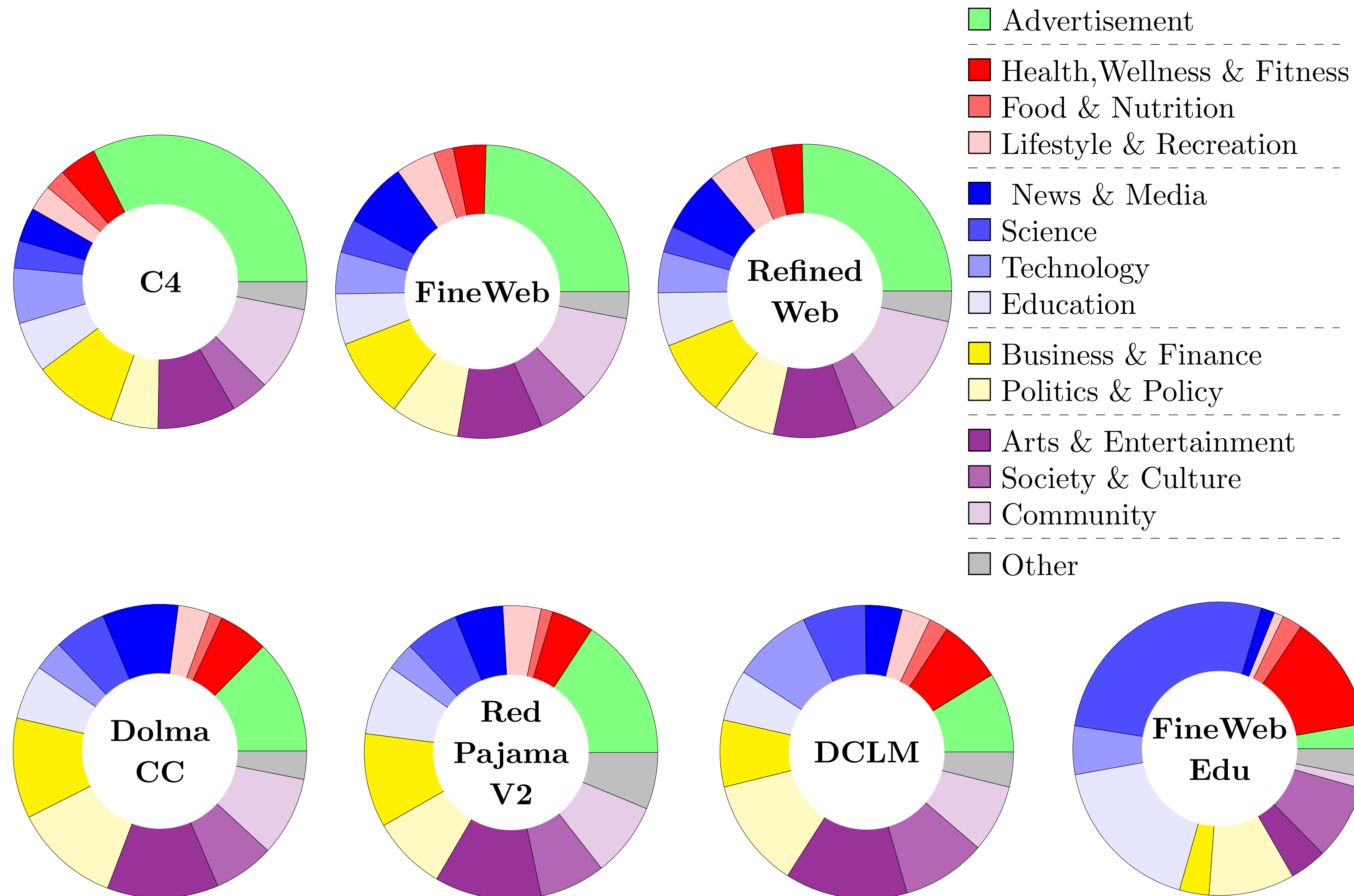
Data curation pipeline



Popular Pretraining datasets

	heuristic filtering & deduplication	machine-learning based quality filtering
C4		none
FineWeb		none
RefinedWeb		none
Dolma CC		some wiki based
RedPajama		some wiki based
DCLM-Baseline		extensive based on instruction fine-tuning data
FineWeb-Edu		extensive based on educational data

Thematic categorization



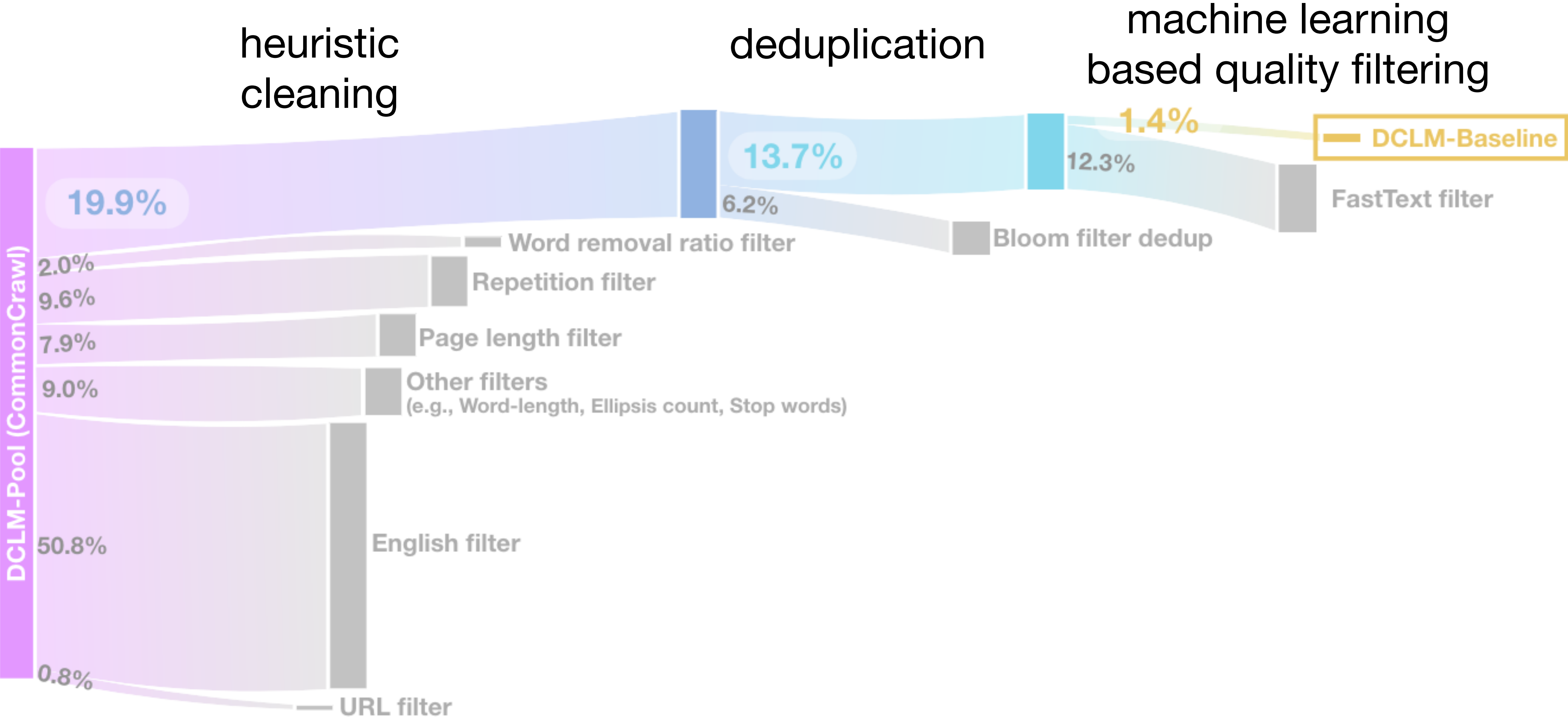
DCLM-Baseline pipeline

DataComp-LM: In search of the next generation of training sets for language models

Jeffrey Li*^{1,2} Alex Fang*^{1,2} Georgios Smyrnis*⁴ Maor Ivgi*⁵
Matt Jordan⁴ Samir Gadre^{3,6} Hritik Bansal⁸ Etash Guha^{1,15} Sedrick Keh³ Kushal Arora³
Saurabh Garg¹³ Rui Xin¹ Niklas Muennighoff²² Reinhard Heckel¹² Jean Mercat³ Mayee
Chen⁷ Suchin Gururangan¹ Mitchell Wortsman¹ Alon Albalak^{19,20} Yonatan Bitton¹⁴
Marianna Nezhurina^{9,10} Amro Abbas²³ Cheng-Yu Hsieh¹ Dhruva Ghosh¹ Josh Gardner¹
Maciej Kilian¹⁷ Hanlin Zhang¹⁸ Rulin Shao¹ Sarah Pratt¹ Sunny Sanyal⁴ Gabriel Ilharco¹
Giannis Daras⁴ Kalyani Marathe¹ Aaron Gokaslan¹⁶ Jieyu Zhang¹ Khyathi Chandu¹¹
Thao Nguyen¹ Igor Vasiljevic³ Sham Kakade¹⁸ Shuran Song^{6,7} Sujay Sanghavi⁴ Fartash
Faghri² Sewoong Oh¹ Luke Zettlemoyer¹ Kyle Lo¹¹ Alaaeldin El-Nouby² Hadi
Pouransari² Alexander Toshev² Stephanie Wang¹ Dirk Groeneveld¹¹ Luca Soldaini¹¹
Pang Wei Koh¹ Jenia Jitsev^{9,10} Thomas Kollar³ Alexandros G. Dimakis^{4,21}
Yair Carmon⁵ Achal Dave^{†3} Ludwig Schmidt^{†1,7} Vaishaal Shankar^{†2}

¹University of Washington, ²Apple, ³Toyota Research Institute, ⁴UT Austin, ⁵Tel Aviv University, ⁶Columbia University, ⁷Stanford, ⁸UCLA, ⁹JSC, ¹⁰LAION, ¹¹AI2, ¹²TUM, ¹³CMU, ¹⁴Hebrew University, ¹⁵SambaNova, ¹⁶Cornell, ¹⁷USC, ¹⁸Harvard, ¹⁹UCSB, ²⁰SynthLabs, ²¹Bespokelabs.AI, ²²Contextual AI, ²³DatologyAI

Construction of DCLM-BASELINE



Step 3: Machine learning-based quality filtering

Binary classification, keep top 10%:

- **Negative data: pretraining data**
- **Positive Data: instruction fine-tune data**

Dataset	Threshold	CORE	MMLU	EXTENDED
OH-2.5 + ELI5	10%	41.0	29.2	21.4
Wikipedia	10%	35.7	27.0	19.1
OpenWebText2	10%	34.7	25.0	18.7
GPT-3 Approx	10%	37.5	24.4	20.0
OH-2.5 + ELI5	15%	39.8	27.2	21.5
OH-2.5 + ELI5	20%	38.7	24.2	20.3

Example instruction finetuning data from OpenHermes 2.5

Human: A factory produces 250 widgets every day. How many widgets will the factory produce in a year, assuming it operates 365 days a year?

GPT: To find the total number of widgets produced in a year, we can multiply the daily production rate by the number of days in a year:

Total widgets = Daily production * Number of days
$$= 250 * 365$$

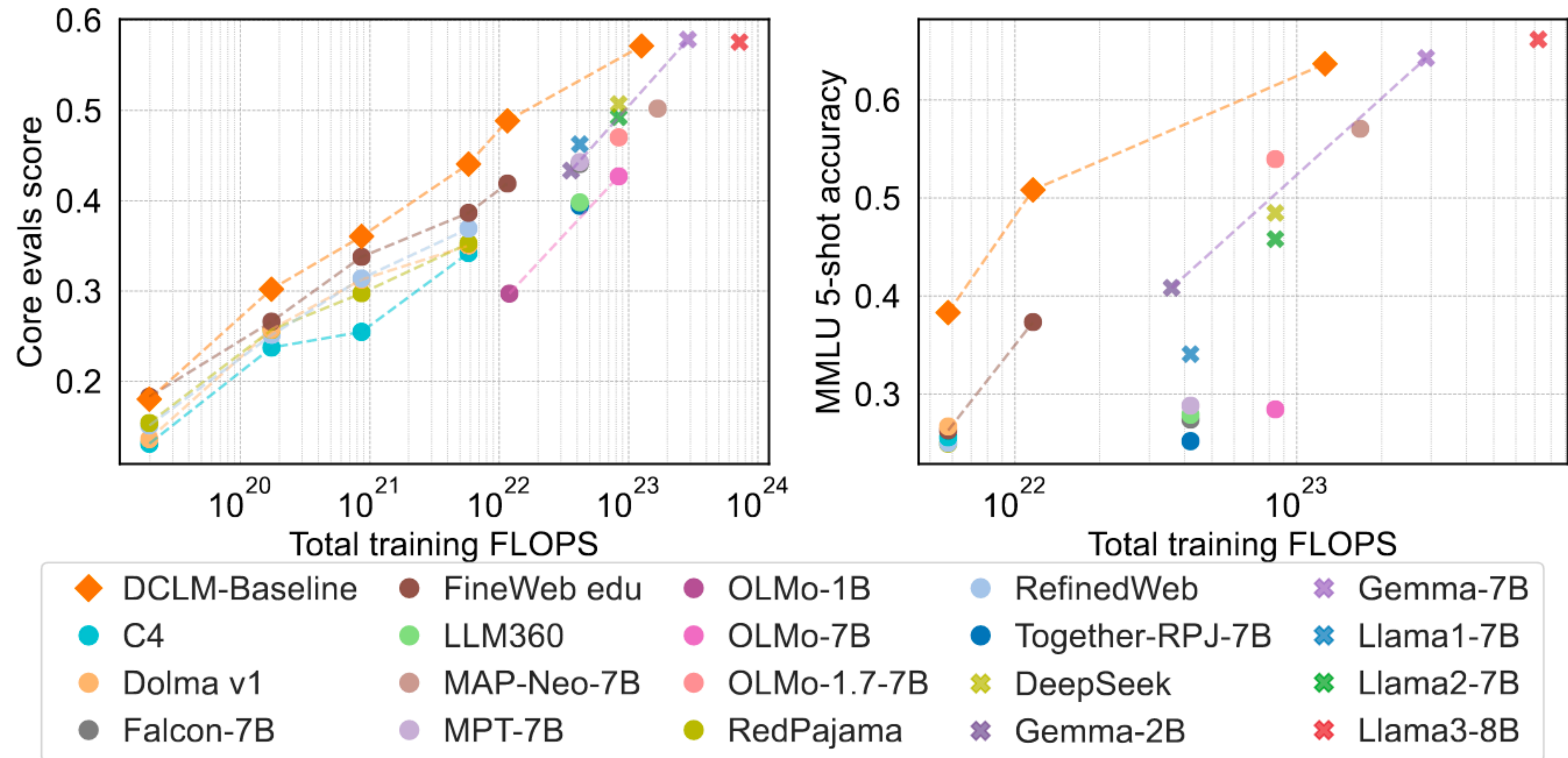
Calculate the result:

Total widgets = 91,250

So, the factory will produce 91,250 widgets in a year if it operates 365 days a year.

Improved data yields better models and/or cheaper to train models

- Better training data = better model performance when trained on the data



Beyond web-filtered data

- Many other data sources exist and are useful
- Often data sources are combined, but how to weight them is difficult

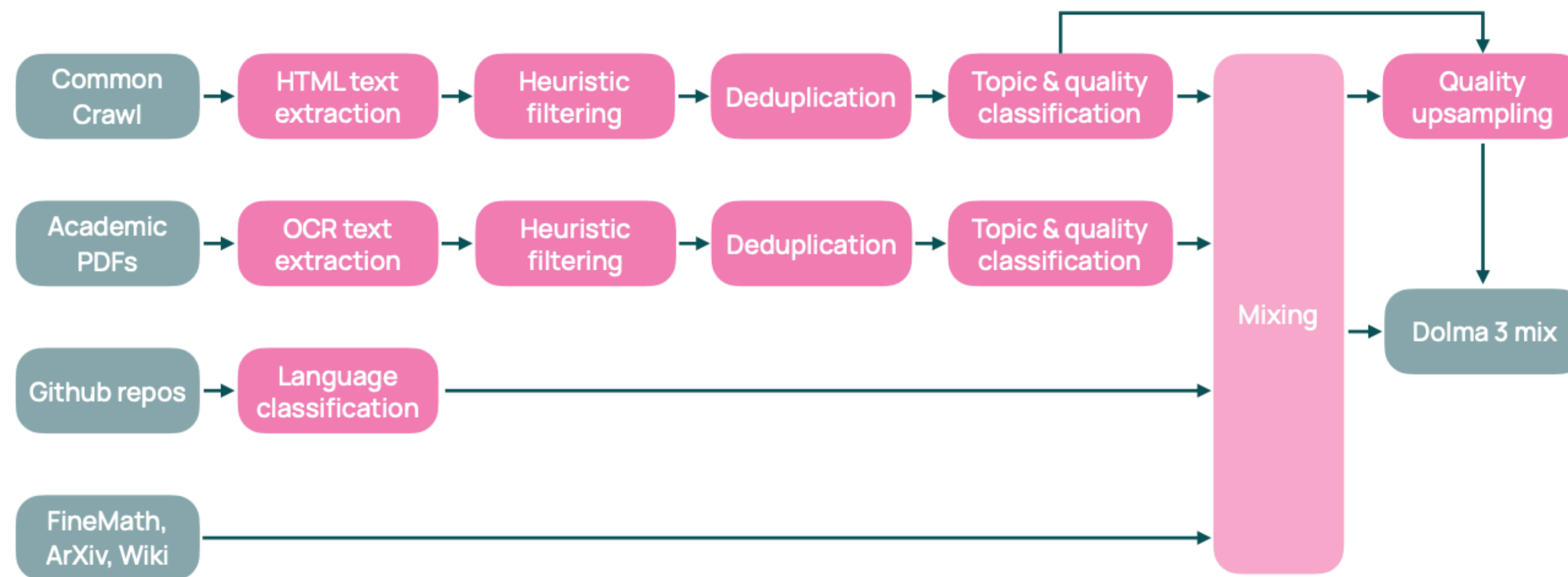


Figure 8 Data curation flow for pretraining data sources in DOLMA 3 MIX.

Olmo 3, 2025

Stages of pre-training

- Pre-training on web-scale data
- Mid-training on high quality tokens
- Long-context extension: Continued pre-training on a mix of long and regular length sequences

Type	Source	2T Pool		100B Mix	
		Tokens	Docs	Tokens	Docs
Math (synth)	TinyMATH Mind**	899M	1.42M	898M (0.9%)	1.52M
Math (synth)	TinyMATH PoT**	241M	729K	241M (0.24%)	758K
Math (synth)	CraneMath*	5.62B	6.55M	5.62B (5.63%)	7.24M
Math (synth)	MegaMatt*	3.88B	6.79M	1.73B (1.73%)	3.23M
Math (synth)	Dolmino Math^^	10.7B	21M	10.7B (10.7%)	22.3M
Code	StackEdu (FIM)^	21.4B	32M	10.0B (10.0%)	16.2M
Python (synth)	CraneCode*	18.8B	19.7M	10.0B (10.0%)	11.7M
QA (synth)	Reddit To Flashcards**	21.6B	370M	5.90B (5.9%)	101M
QA (synth)	Wiki To RCQA**	4.22B	22.3M	3.0B (3.0%)	16.3M
QA (synth)	Nemotron Synth QA^	487B	972M	5.0B (5.0%)	10.6M
Thinking (synth)	Math Meta-Reasoning**	1.05B	984K	381M (0.38%)	401K
Thinking (synth)	Code Meta-Reasoning**	1.27B	910K	459M (0.46%)	398K
Thinking (synth)	Program-Verifiable**	438M	384K	159M (0.16%)	158K
Thinking (synth)	OMR Rewrite FullThoughts^	850M	291K	850M (0.85%)	394K
Thinking (synth)	QWQ Reasoning Traces^	4.77B	438K	1.87B (1.87%)	401K
Thinking (synth)	General Reasoning Mix^	2.48B	668K	1.87B (1.87%)	732K
Thinking (synth)	Gemini Reasoning Traces^	246M	55.2K	246M (0.25%)	85.1K
Thinking (synth)	Llama Nemotron Reasoning Traces^	20.9B	3.91M	1.25B (1.25%)	368K
Thinking (synth)	OpenThoughts2 Reasoning Traces^	5.6B	1.11M	1.25B (1.25%)	402K
Instruction (synth)	Tulu 3 SFT^^	1.61B	1.95M	1.1B (1.1%)	1.45M
Instruction (synth)	Dolmino 1 Flan^^	16.8B	56.9M	5.0B (5.0%)	14.8M
PDFs	OLMOOCR science PDFs (HQ subset)^	240B	28.7M	4.99B (5.0%)	1.20M
Web pages	STEM-Heavy Crawl^	5.21B	5.16M	4.99B (5.0%)	5.53M
Web pages	Common Crawl (HQ subset)^	1.32T	965M	22.4B (22.5%)	18.3M
Total		2.19T	2.52B	99.95B (100%)	236M

Table 5 Composition of the midtraining data (Dolma 3 Dolmino Mix). Here we show the full composition of the midtraining data mix. **=newly-introduced synthetic dataset. *=novel recreation of existing data. ^^=reuse of previously-introduced data. ^=filtering or light transformation of existing external data. *Olmo 3, 2025*

Dataset classification experiment for pretraining data

Text 1: Made it back, can I come inside for a change? Made of glass and falling fast all the way! Thanks for correcting Tokyo Police Club - Miserable lyrics!

Task: C4 or FineWeb? C4

Text 2: Jamie Oliver is a famous CHEF from the UK. Here you can learn how to make scramble eggs in three different ways: English, French and American way! ENJOY IT!

Task: C4 or FineWeb? C4

Text 3: Short-term and long-term changes in the strength of synapses in neural networks underlie working memory and long-term memory storage in the brain.

Task: C4 or FineWeb? FineWeb

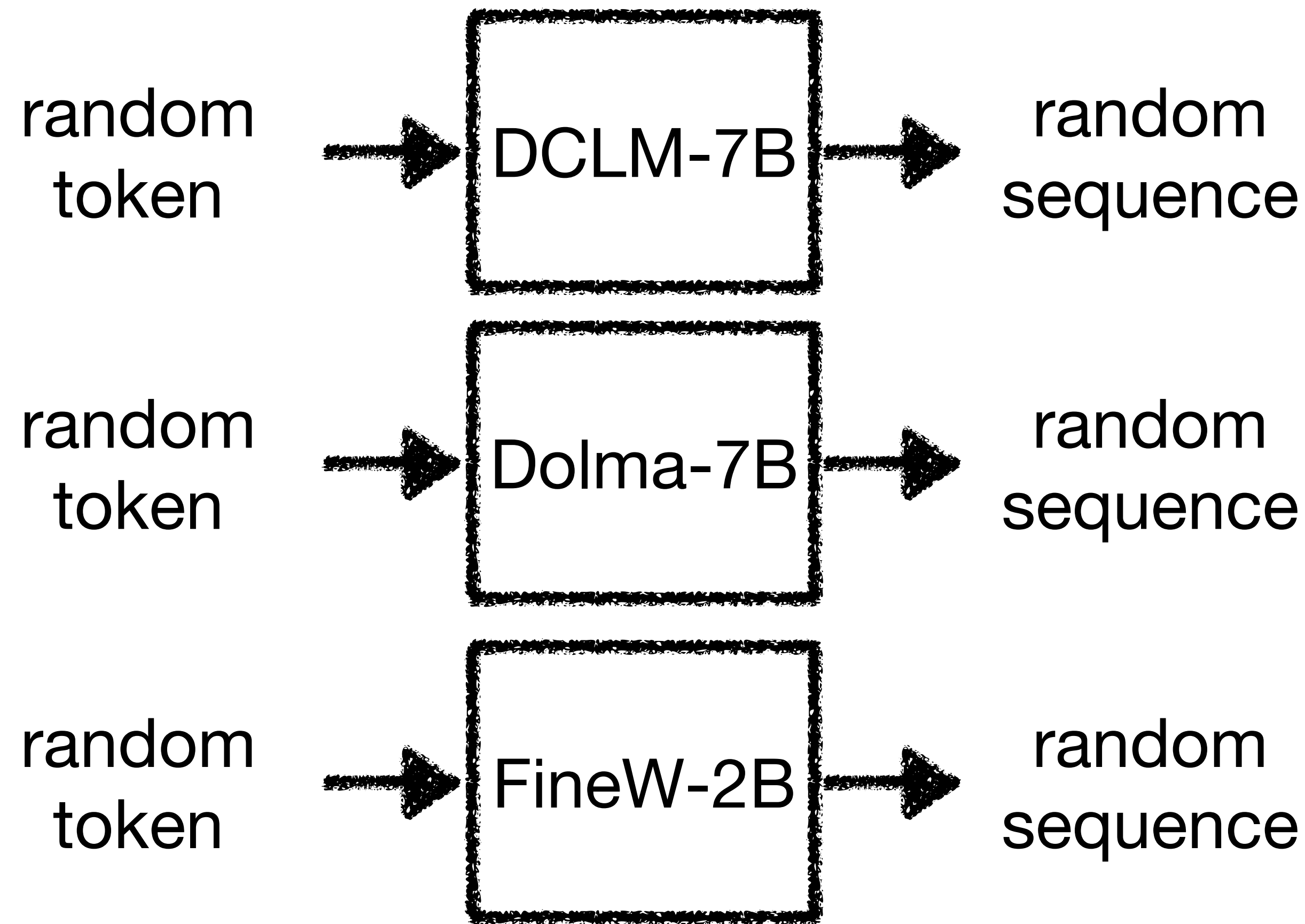
Text 4: Yesterday, we indulged in all the goodness of sweets, so I thought it only appropriate that we feature the other side of the coin: Salty. Now, I'm a girl who loves her potato chips.

Task: C4 or FineWeb? FineWeb

Dataset classification

- Text sequences can be well classified according to datasets
- For humans this task is very difficult. C4 vs FineWeb:
 - Human: 63%
 - GPT: 85.3%

Bias propagation through models



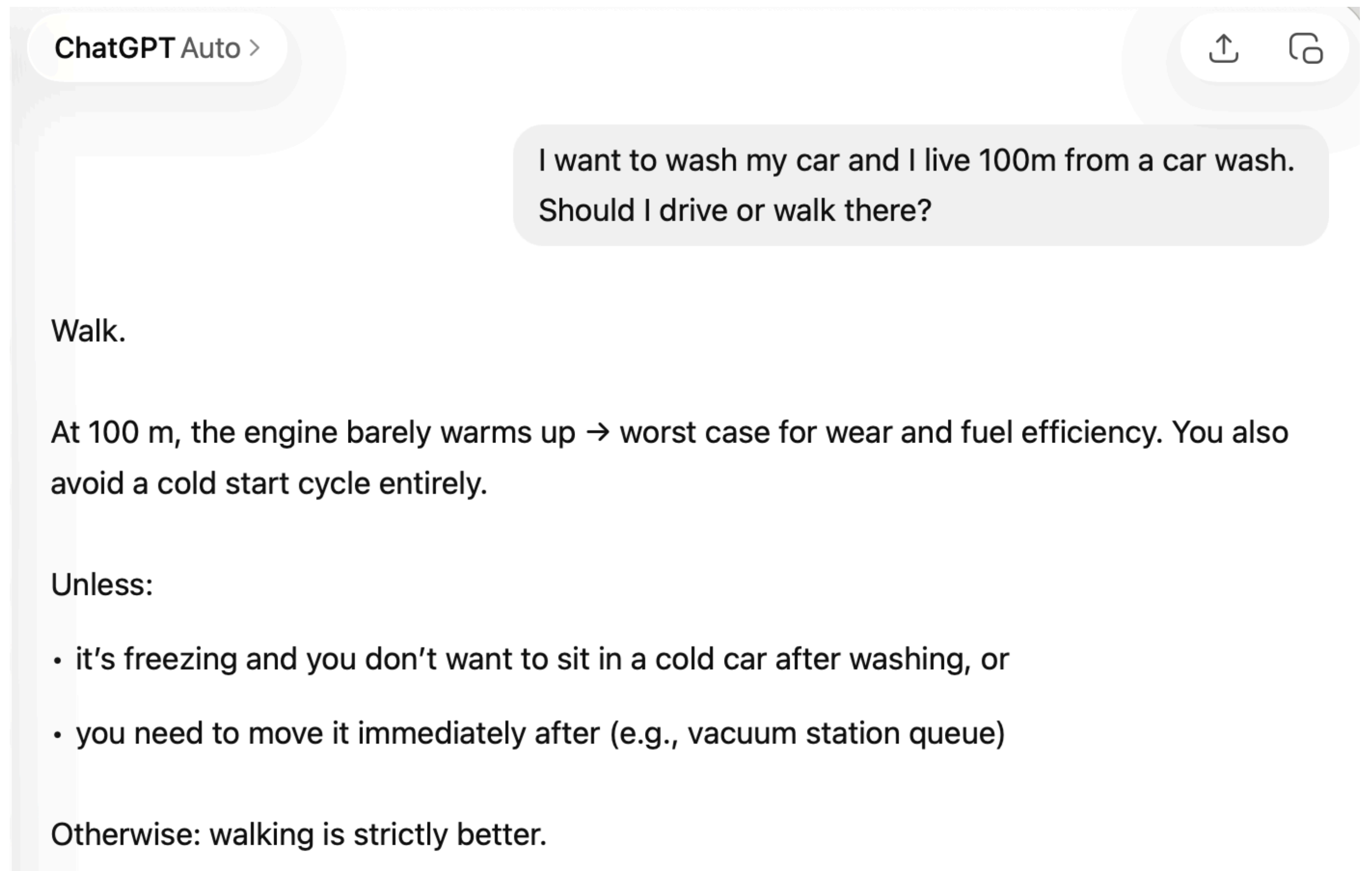
84% accuracy of classifying generated sequences with classifier trained on pertaining data!

vs. 93% on original data

Discussion

Q1: GPT2 (early LLM) was already good at cooking recipes - why?

Q2: Why does GPT5.4 make this naive mistake?



ChatGPT Auto >

I want to wash my car and I live 100m from a car wash. Should I drive or walk there?

Walk.

At 100 m, the engine barely warms up → worst case for wear and fuel efficiency. You also avoid a cold start cycle entirely.

Unless:

- it's freezing and you don't want to sit in a cold car after washing, or
- you need to move it immediately after (e.g., vacuum station queue)

Otherwise: walking is strictly better.

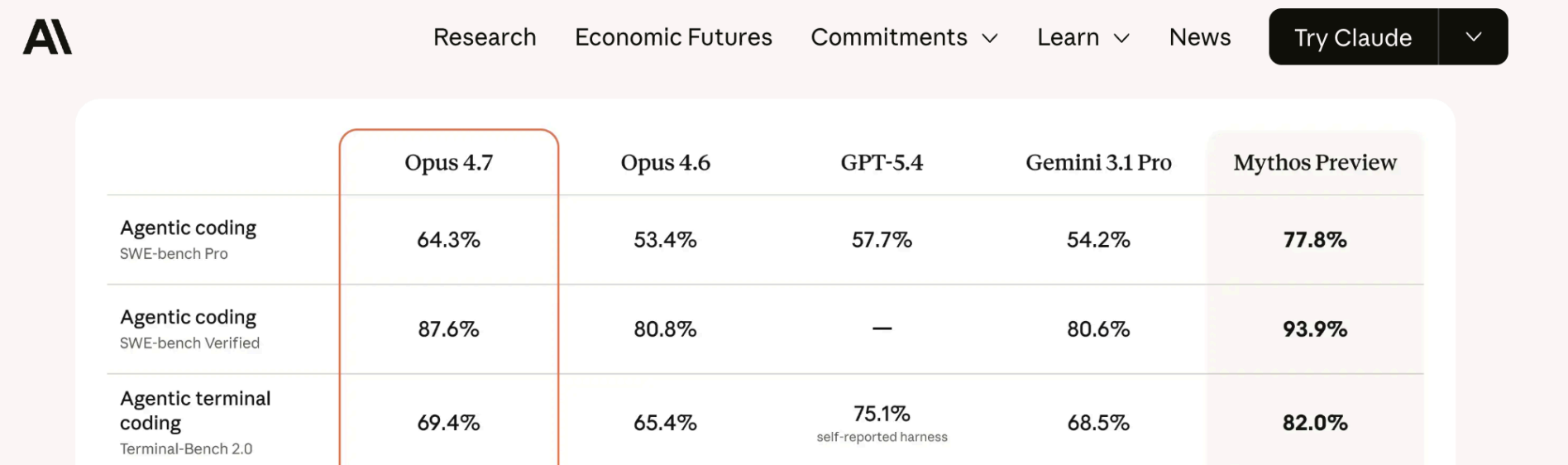
Summary pre-training data for LLMs

- Pre-training data is primarily curated web data
- Data curation is critical for LLM performance
- Better data means a model performs better when trained on it
- Synthetic data is becoming increasingly relevant
- Further reading:
 - DataComp-LM: In search of the next generation of training sets for language models, 2024
 - Olmo 3, 2025

Evaluation

Why evaluate models?

- Model selection: Customer [individual, company] decides which model to use
- Evaluation is critical for progress in AI: Improving on benchmarks has been the main way machine learning progresses
 - Developers need to measure the comparative capabilities of model architecture, training data, etc
- Safety and risk assessment, scientific understanding



The screenshot shows a benchmarking table for AI models. The table compares five models: Opus 4.7, Opus 4.6, GPT-5.4, Gemini 3.1 Pro, and Mythos Preview across three categories: Agentic coding (SWE-bench Pro), Agentic coding (SWE-bench Verified), and Agentic terminal coding (Terminal-Bench 2.0). Opus 4.7 is highlighted with a red border.

	Opus 4.7	Opus 4.6	GPT-5.4	Gemini 3.1 Pro	Mythos Preview
Agentic coding SWE-bench Pro	64.3%	53.4%	57.7%	54.2%	77.8%
Agentic coding SWE-bench Verified	87.6%	80.8%	—	80.6%	93.9%
Agentic terminal coding Terminal-Bench 2.0	69.4%	65.4%	75.1% self-reported harness	68.5%	82.0%



Evaluation approaches

- **Perplexity:** How well does the model predict the next token on a sequence?
- **Benchmarks**
 - Usually a verifiable task: Problem, model answers, automatic verification (correct/incorrect), report average of results (e.g., 70%)
 - Examples:
 - Reasoning: Hellaswag, GPQA Diamond, Humanity's last exam
 - General: MMLU
 - Math: GSM8K, MATH, AIME
 - Code, agentic: SWE-Bench, Terminal Bench
- **Evaluating open-ended outputs:** LLM as a judge, grading against rubrics
- **Human preference-based evaluation:** Chatbot arena

Perplexity

- The perplexity measures how well a LLM predicts the next token
- Given a test set with N tokens t_1, t_2, \dots

- perplexity = $\exp \left(-\frac{1}{N} \sum_{i=1}^N \log P(t_i) \right)$

- Where $P(t_i)$: Probability of i -th token according to model

: A BENCHMARK FOR EVALUATING LANGUAGE MODEL FIT

Ian Magnusson[♣] Akshita Bhagia[♣] Valentin Hofmann[♣] Luca Soldaini[♣]
Ananya Harsh Jha[♣] Oyvind Tafjord[♣] Dustin Schwenk[♣] Evan Pete Walsh[♣]
Yanai Elazar^{♣◇} Kyle Lo[♣] Dirk Groeneveld[♣] Iz Beltagy[♣] Hannaneh Hajishirzi^{♣◇}
Noah A. Smith^{♣◇} Kyle Richardson[♣] Jesse Dodge[♣]
[♣]Allen Institute for Artificial Intelligence
[◇]Paul G. Allen School of Computer Science & Engineering, University of Washington
{ianm, jessed}@allenai.org

- Cheap and dense signal
- Measures prediction, not capability

ABSTRACT

Language models (LMs) commonly report perplexity on monolithic data held out from training. Implicitly or explicitly, this data is composed of domains—varying distributions of language. Rather than assuming perplexity on one distribution extrapolates to others, PERPLEXITY ANALYSIS FOR LANGUAGE MODEL ASSESSMENT (PALOMA),¹ measures LM fit to 585 text domains, ranging from *nytimes.com* to *r/depression* on Reddit. We invite submissions to our benchmark and organize results by comparability based on compliance with guidelines such as removal of benchmark contamination from pretraining. Submissions can also record parameter and training token count to make comparisons of Pareto efficiency for performance as a function of these measures of cost. We populate our benchmark with results from 6 baselines pretrained on popular corpora. In case studies, we demonstrate analyses that are possible with PALOMA, such as finding that pretraining without data beyond Common Crawl leads to inconsistent fit to many domains.

Benchmarks on verifiable tasks

- Format:
 - Problem
 - Model answers
 - Automatic verification (correct/incorrect)
 - Report average of results (e.g., 70%)
- Historically often from [school, college, professional] exams
- Historically often multiple choice or very simple verification (e.g., is the solution to a math problem X)

HellaSwag: Can a Machine Really Finish Your Sentence?

Rowan Zellers[^] Ari Holtzman[^] Yonatan Bisk[^] Ali Farhadi[^][♡] Yejin Choi[^][♡]

[^]Paul G. Allen School of Computer Science & Engineering, University of Washington

[♡]Allen Institute for Artificial Intelligence

<https://rowanzellers.com/hellaswag>

- 70k multiple choice questions, focused on common sense reasoning
- A question and 4 possible continuations.
- Situation: "A person is mixing ingredients in a bowl."
Option A: "They then pour the mixture into a pan."
Option B: "They turn on the faucet."
Option C: "They plug in the kettle."
Option D: "They sit down and start watching TV."
- Situation and options are concatenated, and we evaluate what's most likely according to the LLM

Measuring massive multitask language understanding (MMLU) benchmark

- Benchmark proposed by Hendrycks et al., 2020
- Multiple choice questions from various fields: Humanities, social sciences, hard sciences, and other areas
- Questions collected by graduate and undergraduate students from sources online
- 15908 questions split into development, validation, and test set, test set has 14079 questions
- Inspection of the dataset: https://huggingface.co/datasets/cais/mmlu/viewer/college_medicine/dev?row=0

Microeconomics

One of the reasons that the government discourages and regulates monopolies is that

- (A) producer surplus is lost and consumer surplus is gained.
- (B) monopoly prices ensure productive efficiency but cost society allocative efficiency.
- (C) monopoly firms do not engage in significant research and development.
- (D) consumer surplus is lost with higher prices and lower levels of output.



**Conceptual
Physics**

When you drop a ball from rest it accelerates downward at 9.8 m/s^2 . If you instead throw it downward assuming no air resistance its acceleration immediately after leaving your hand is

- (A) 9.8 m/s^2
- (B) more than 9.8 m/s^2
- (C) less than 9.8 m/s^2
- (D) Cannot say unless the speed of throw is given.



**Conceptual
Physics**

When you drop a ball from rest it accelerates downward at 9.8 m/s^2 . If you instead throw it downward assuming no air resistance its acceleration immediately after leaving your hand is

- (A) 9.8 m/s^2
- (B) more than 9.8 m/s^2
- (C) less than 9.8 m/s^2
- (D) Cannot say unless the speed of throw is given.



**College
Mathematics**

In the complex z -plane, the set of points satisfying the equation $z^2 = |z|^2$ is a

- (A) pair of points
- (B) circle
- (C) half-line
- (D) line



A 33-year-old man undergoes a radical thyroidectomy for thyroid cancer. During the operation, moderate hemorrhaging requires ligation of several vessels in the left side of the neck. Postoperatively, serum studies show a calcium concentration of 7.5 mg/dL, albumin concentration of 4 g/dL, and parathyroid hormone concentration of 200 pg/mL. Damage to which of the following vessels caused the findings in this patient?

- (A) Branch of the costocervical trunk
- (B) Branch of the external carotid artery
- (C) Branch of the thyrocervical trunk
- (D) Tributary of the internal jugular vein



- Typically prompted with 5-shot evaluation, i.e., we repeat 5 questions with answer and then phrase the actual question

Question: ...
Choices: ...
Correct answer: B

Question: ...
Choices: ...
Correct answer: B

Question: Glucose is transported into the muscle cell:

Choices:
A. via protein transporters called GLUT4.
B. only in the presence of insulin.
C. via hexokinase.
D. via monocarbylic acid transporters.

Correct answer:

Grade School Math 8k (GSM8K)

- 8.5K high quality grade school math word problems written by human problem writers
- Test multi-step reasoning abilities

Question:

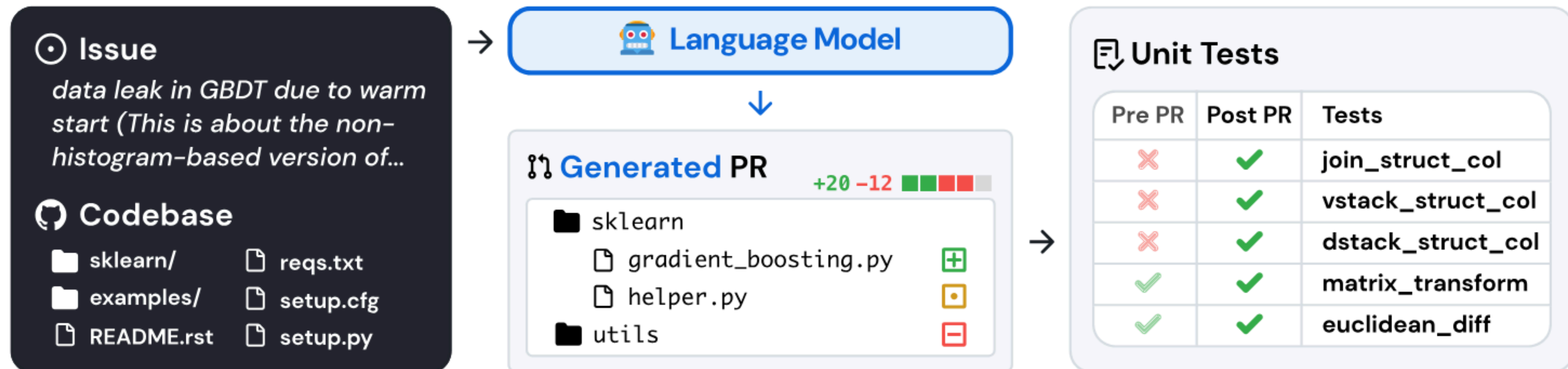
Natalia sold clips to 48 of her friends in April, and then she sold half as many clips in May. How many clips did Natalia sell altogether in April and May?

Answer:

Natalia sold $48/2 = \langle\langle 48/2=24 \rangle\rangle 24$ clips in May. Natalia sold $48+24 = \langle\langle 48+24=72 \rangle\rangle 72$ clips altogether in April and May. ##### 72

SWE Bench

- Given codebase and issue description, write pull request
- Evaluation: Unit tests



Task familiarity

- Evaluation performance depends strongly on how well prepared the model is for the given task
- See Dominguez-Olmedo et al. “Training on the Test Task Confounds Evaluation and Emergence”, 2024

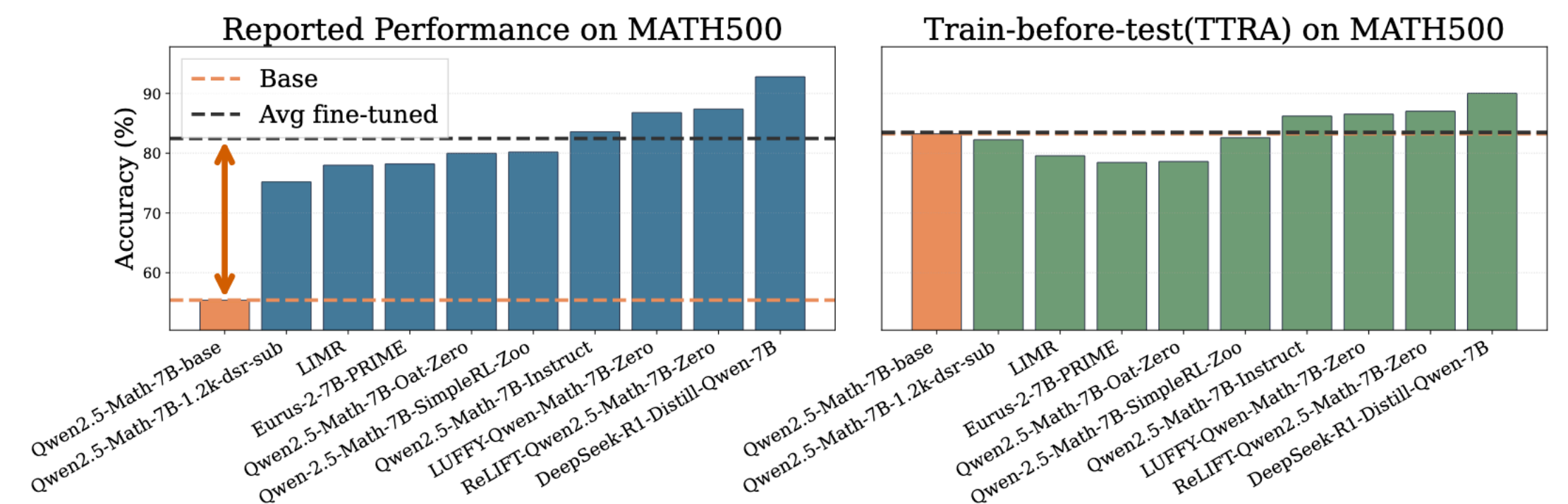


Figure 1: Fine-tuning gains are largely diminished after task alignment. **Left:** The reported performance reveals substantial accuracy gaps between base models and their fine-tuned variants. **Right:** After applying TTRA to all models, the base model’s performance increases significantly, nearly matching that of the fine-tuned models, suggesting that many gains from RLVR/SFT reported in the literature are not a difference in reasoning capability, but rather artifacts of task familiarity.

Test-time RL alignment exposes task familiarity artifacts in LLM benchmarks, 2026

Evaluating open-ended outputs

- **Problem:** Most benchmark questions/tasks have one clear answer or can be verified, but many tasks are open-ended: E.g., legal report writing
- LLM-as-a-judge: Uses an LLM to score output like a human would
 - Issue: Is limited by the capabilities of the judge model
- Rubrics: Decompose quality into explicit criteria (e.g, has issue X been mentioned in the report)
 - LLM judge only needs to score the criterion; weaker judges can work
 - Issue: Constrained by rubrics, difficult if there are multiple solution paths

Human preference based evaluation

- Chatbot arena:
 - User types a query, gets two responses, rates which is better
- ELO ratings based on pairwise comparisons are computed

Chatbot Arena (lmarena.ai) is an open-source platform for evaluating AI through human preference, developed by researchers at [LMarena](https://lmarena.ai). With over 1,000,000 user votes, the platform ranks best LLM and AI chatbots using the Bradley-Terry model to generate live leaderboards. For technical details, check out our [paper](#).

Chatbot Arena thrives on community engagement — cast your vote to help improve AI evaluation!

[New Launch! Copilot Arena: VS Code Extension to compare Top LLMs](#)

Arena [NEW: Overview](#) [Arena \(Vision\)](#) [Arena-Hard-Auto](#) [Full Leaderboard](#)

Total #models: 234. Total #votes: 2,912,179. Last updated: 2025-05-05.

Code to recreate leaderboard tables and plots in this [notebook](#). You can contribute your vote at lmarena.ai!

Category

Overall

Apply filter

Style Control

Show Deprecated

Overall Questions

#models: 234 (100%) #votes: 2,912,179 (100%)

Rank* (UB)	Rank (StyleCtrl)	Model	Arena Score	95% CI	Votes	Organization	License	Knowledge Cutoff
1	1	Gemini-2.5-Pro-Preview-05-06	1448	+7/-12	3545	Google	Proprietary	Unknown
1	1	Gemini-2.5-Pro-Exp-03-25	1437	+6/-4	12720	Google	Proprietary	Unknown
3	1	o3-2025-04-16	1411	+11/-7	5844	OpenAI	Proprietary	Unknown

Summary evaluation

- Evaluation is critical for progress in AI: Improving on benchmarks has been the main way machine learning progresses
- Level of task-preparation is important and can confound evaluations
- Benchmarks are useful for model comparisons, less so as absolute performance measures
- Test/training set overlap can confound evaluations
- Further reading: Moritz Hardt “The Emerging Science of Machine Learning Benchmarks”, 2026.