Data

Where the training mixture comes from, and how anyone knows it is good

Required

Penedo et al., The FineWeb Datasets · arXiv:2406.17557

Optional

Li et al., DataComp-LM · arXiv:2406.11794 Xie et al., DoReMi · arXiv:2305.10429 Wang et al., Self-Instruct · arXiv:2212.10560
Shumailov et al., The Curse of Recursion · arXiv:2305.17493 Gunasekar et al., Textbooks Are All You Need · arXiv:2306.11644

Christoffer Heckman
Department of Computer Science, University of Colorado Boulder

September 8, 2026

Quick overview

Where Thursday stopped. The judge’s biases measured, contamination, error bars, the harness.

Then the levers, and lever one is today

The mixture. What a model eats, and when the human-written web runs out.

The mixture is a choice

The pipeline. FineWeb, stage by stage, with your vote on four real documents.

Dedup, filters, a learned classifier

Synthesis. You already trained on synthetic text. Then collapse, on your codebase.

TinyStories, textbooks, Self-Instruct, recursion

The meter. Who grades the data: humans, judges, classifiers.

What each costs, what each cannot see

1 WHERE THURSDAY STOPPED

Six Ways a Number Lies

Failure Modes

The judge’s

Position bias. Swap A and B, and the verdict changes.

Verbosity bias. Longer scores higher, content held equal.

Self-preference. The judge favors text in its own voice.

The benchmark's

Contamination. The test set leaked into the training data.

Format fragility. Reformat the questions, and the scores move.

Saturation. Scores crowd the ceiling; the gaps left are noise.

Judging the Judge

Claude v1, 23.8: below random; the order IS the verdict

Judge	Prompt	Consistency	Biased toward first	Biased toward second	Error
Claude-v1	default	23.8%	**75.0%**	0.0%	1.2%
	rename	56.2%	11.2%	**28.7%**	**3.8%**
GPT-3.5	default	46.2%	**50.0%**	1.2%	2.5%
	rename	51.2%	38.8%	6.2%	**3.8%**
GPT-4	default	**65.0%**	30.0%	5.0%	0.0%
	rename	**66.2%**	28.7%	5.0%	0.0%

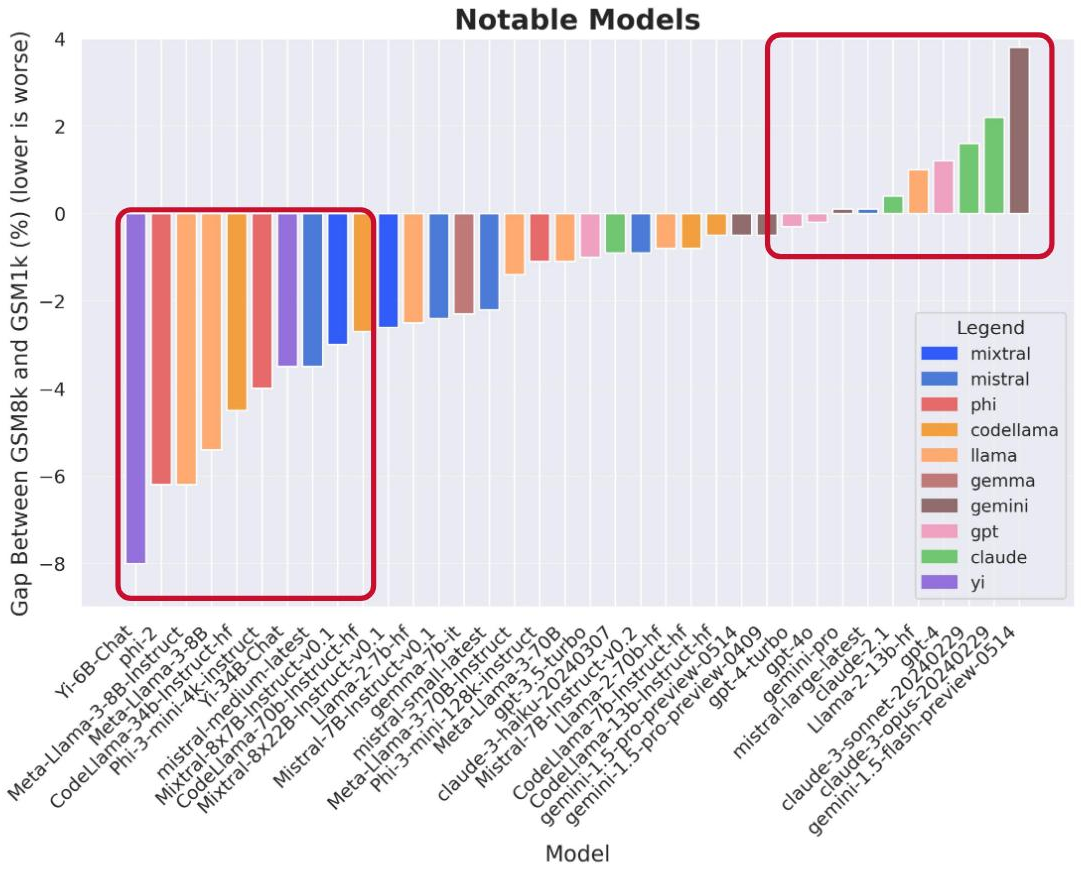
GPT-4: same winner after the swap, 65% of pairs
judged twice, A and B trading places; random verdicts agree 33%

Zheng et al., "Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena," NeurIPS 2023, arXiv:2306.05685, Table 2.

Contamination

the setup: write 1,250 NEW problems at GSM8k difficulty; run every model on both tests; each bar = old score minus new score

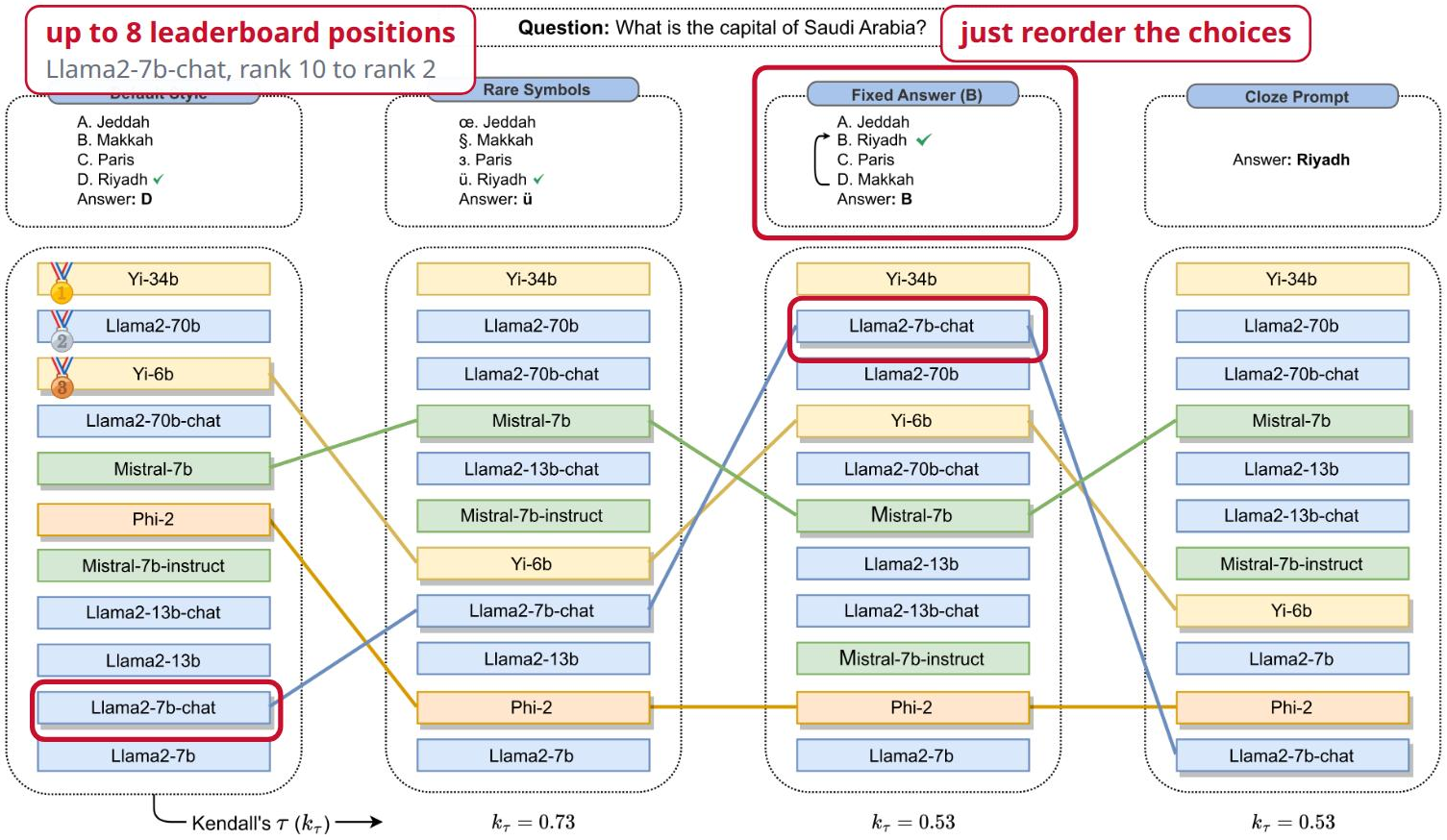
Phi and Mistral families: up to 8 points lower on the fresh problems
that gap is memorized test data, not math



frontier models: bar near zero, the skill is real

Zhang et al., "A Careful Examination of Language Model Performance on Grade School Arithmetic," 2024, arXiv:2405.00332, Figure 1.

Fragile Rankings



Alzahrani et al., "When Benchmarks are Targets: Revealing the Sensitivity of Large Language Model Leaderboards," 2024, arXiv:2402.01781, Figure 1.

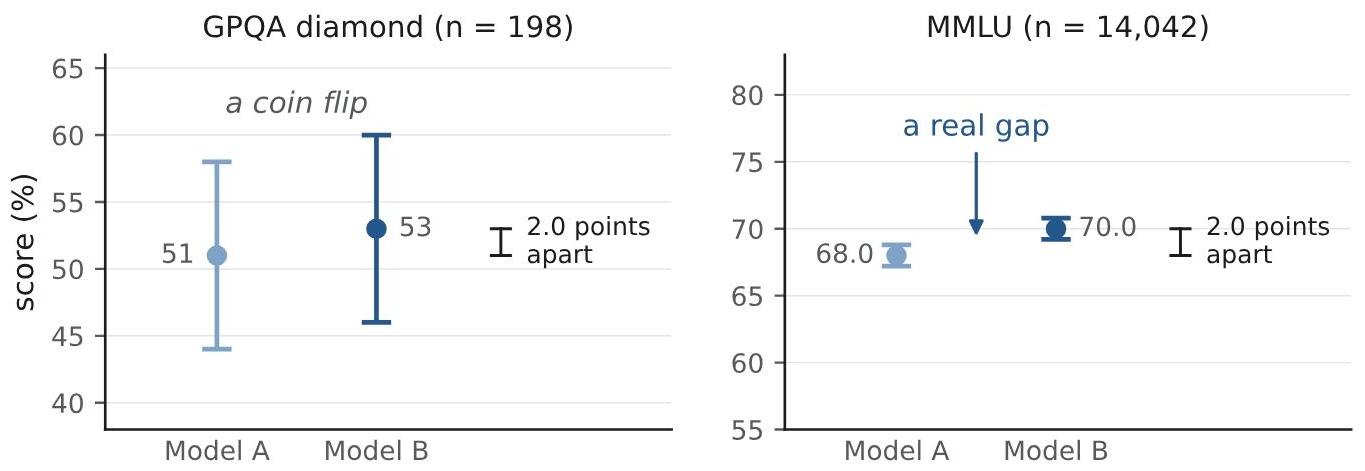
Error Bars

A benchmark score is an average over n questions, so it carries a standard error.

$$\mathrm{SE} = \sqrt{\frac{p(1-p)}{n}}$$

MMLU, $14{,}042$ questions: $\pm 1.96\sqrt{\frac{0.7 \cdot 0.3}{14{,}042}} = 0.8$ percentage points

GPQA diamond, 198 questions: $\pm 1.96\sqrt{\frac{0.5 \cdot 0.5}{198}} = 7.0$ percentage points



A two-point lead on a small benchmark is a coin flip. Paired, per-question comparison tightens it.

2 THE HARNESS

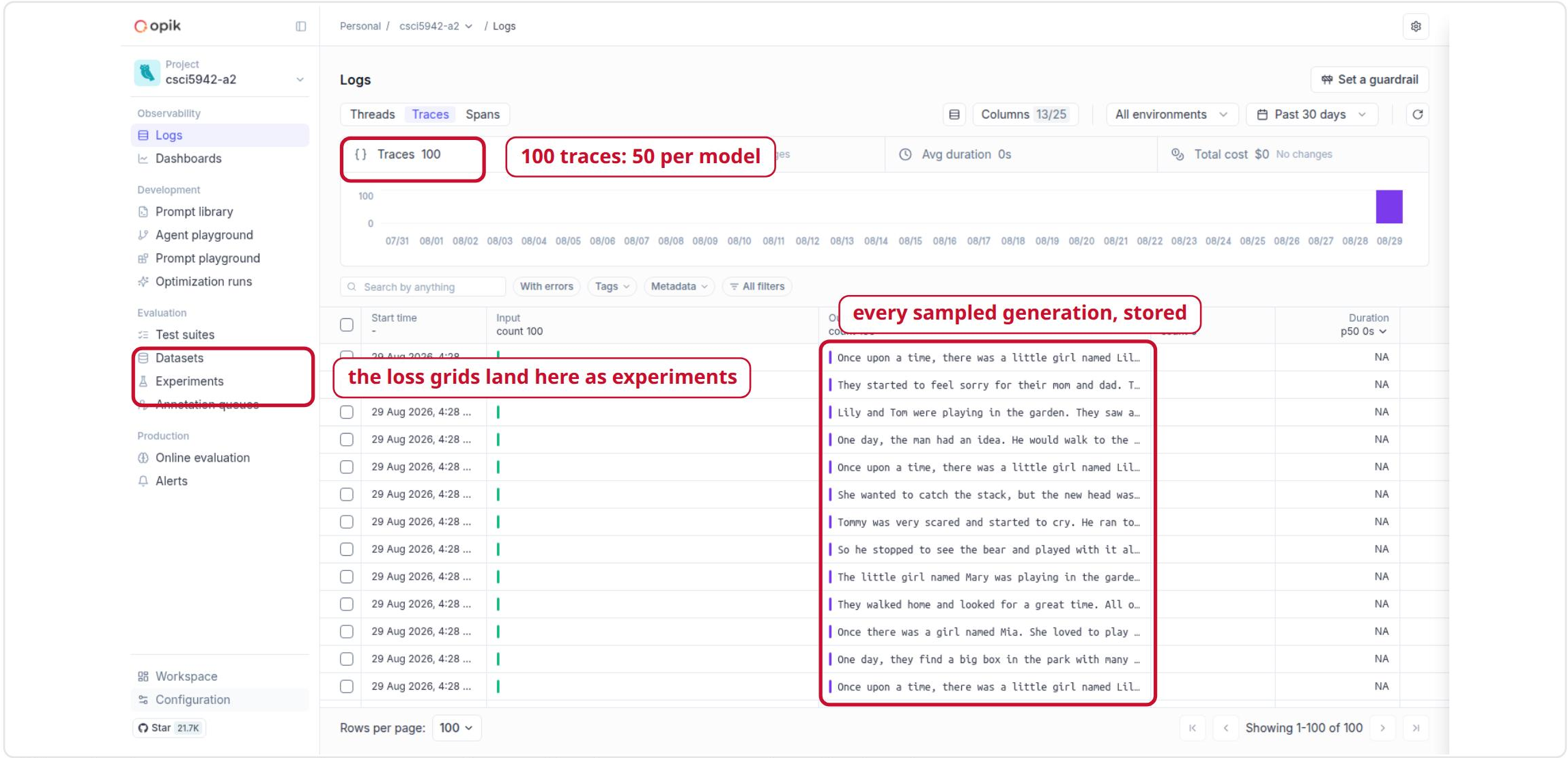
Evaluation as Infrastructure

Harnesses

```
task: mmlu_abstract_algebra
  dataset_path: cais/mmlu
  num_fewshot: 5
  doc_to_choice: ["A", "B", "C", "D"]
  metric: acc
```

The prompt, the shots, the scoring and the aggregation are pinned in config, or the number is not comparable. The same LLaMA 65B weights scored 63.7 on MMLU in one harness and 48.8 in another.

Opik



Self-hosted Opik, project csci5942-a2: the Assignment 1 base model and its TinyStories twin, logged by the A2 harness.

Assignment 2

Train the base config, unchanged, on a second corpus: TinyStories.

Same 82M characters of budget, two minutes on your A1 GPU

Evaluate both models on four domains: the loss grid from this lecture, yours.

Shakespeare, TinyStories, Wikipedia, Python, with the OOV accounting

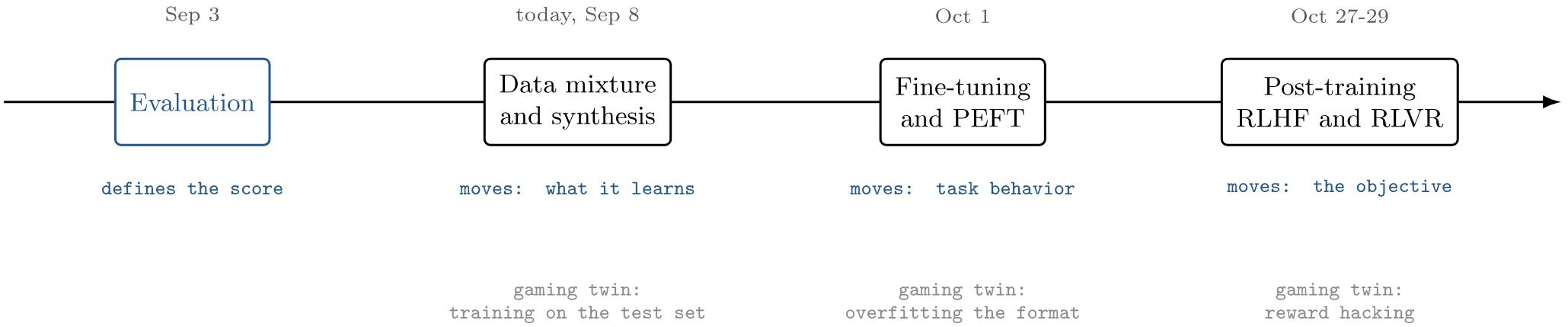
Judge both models’ generations in Opik, with the order-swap control.

Gemini as judge, per-corpus rubric, error bars required

Deliverable: which model is better? Defend the answer per exam.

The correct answer names the exam it is true on

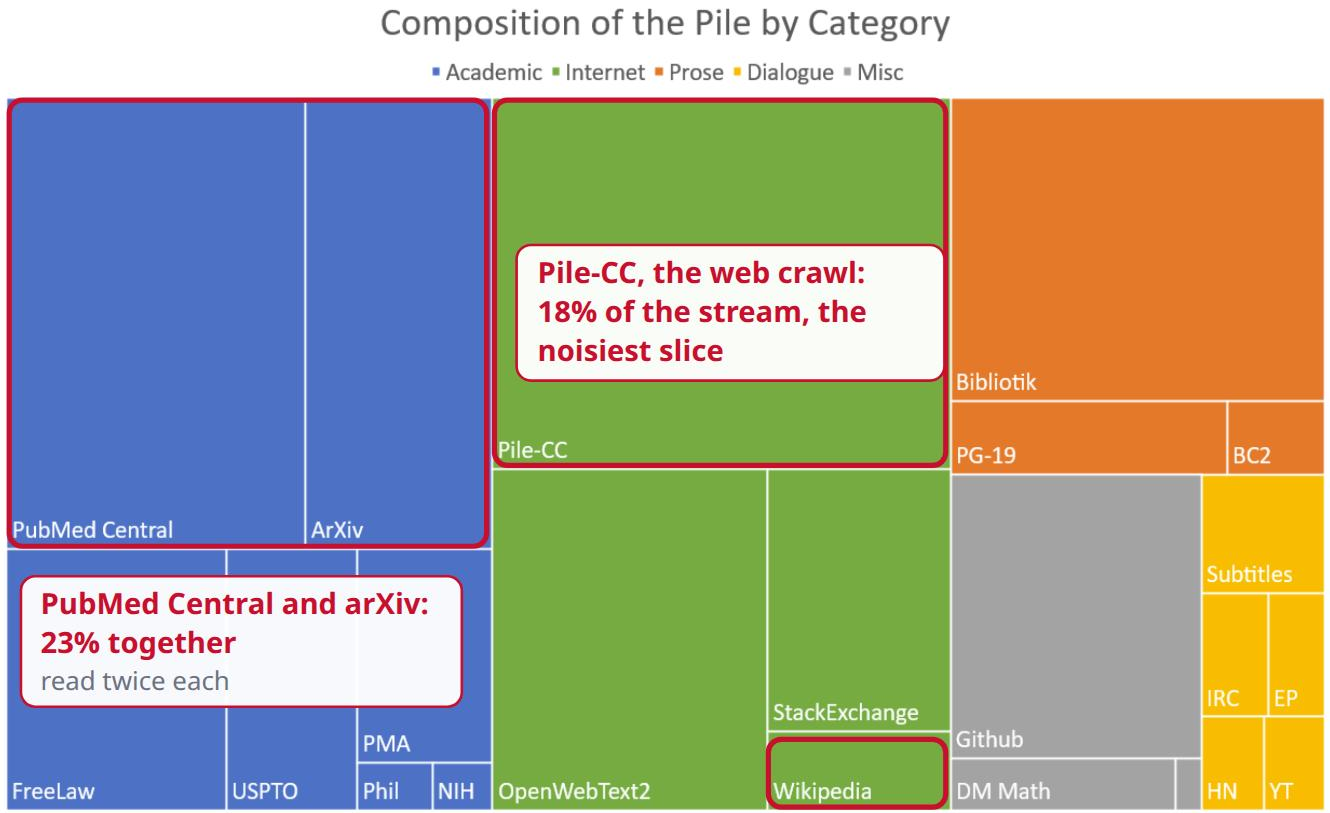
Making the Number Move



3 THE LEVER

The Mixture Decides

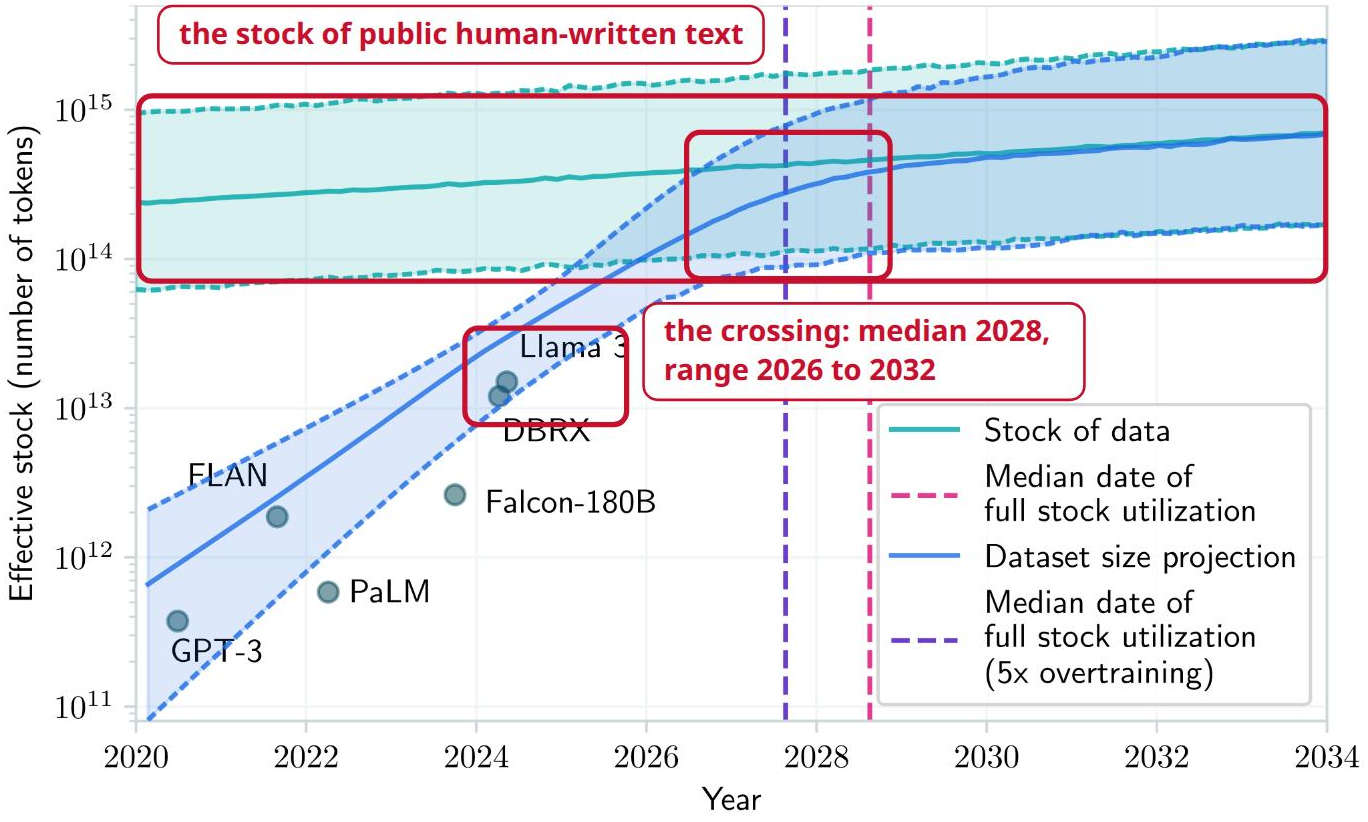
What a Model Eats



Thursday's chain, with the first arrow's domain named: $C \xrightarrow{\text{power law}} L_{\text{mixture}} \xrightarrow{\text{link}} \text{score}$

Gao et al., "The Pile: An 800GB Dataset of Diverse Text for Language Modeling," 2020, arXiv:2101.00027, Figure 1; shares from Table 1 (bytes in the final dataset: raw size times epochs).

Running Out

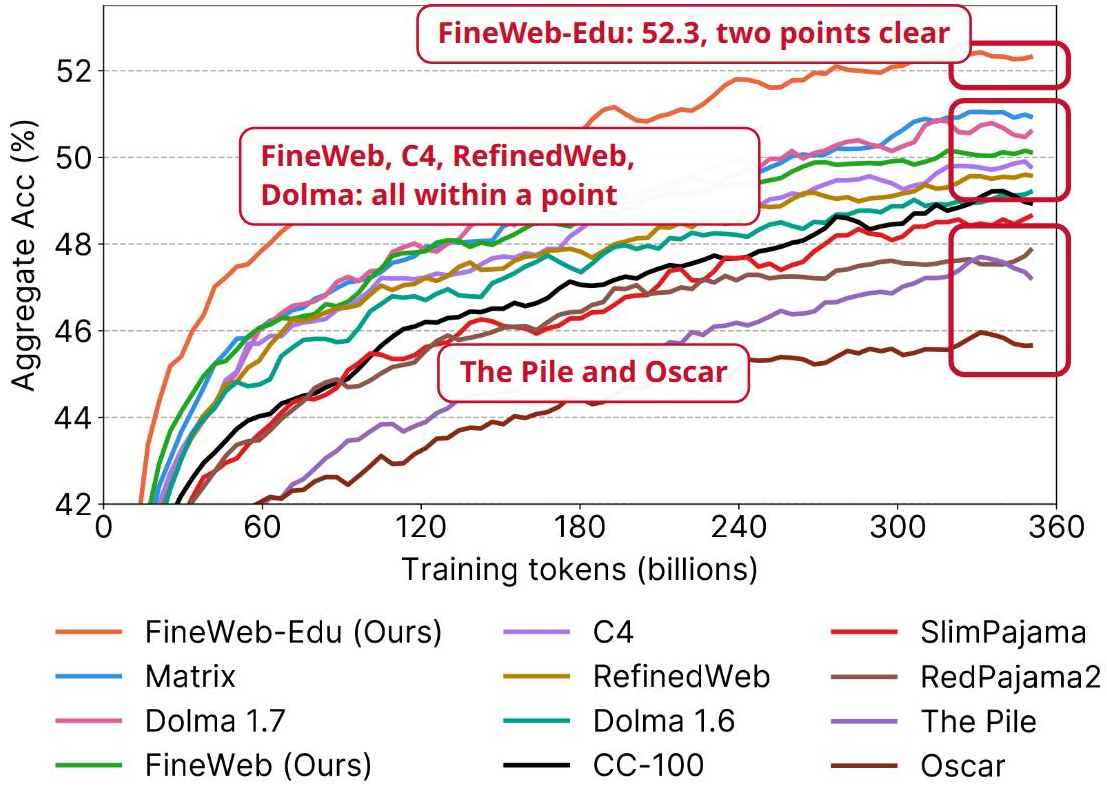


one frontier run: Llama 3 pretrained on 15T tokens, one twentieth of the 320T effective stock

the exits: repeat what exists, keep only the good part, or write more

Villalobos et al., "Will We Run Out of Data? Limits of LLM Scaling Based on Human-Generated Data," 2022, revised 2024, arXiv:2211.04325, Figures 1 and 3. Llama 3 token count: Grattafiori et al., "The Llama 3 Herd of Models," 2024, arXiv:2407.21783.

The Ablation



`the instrument: 1.71B parameters, same architecture and tokenizer, a fixed token budget; only the training data changes`

score: the mean over eight benchmarks (HellaSwag, ARC, MMLU, five more)

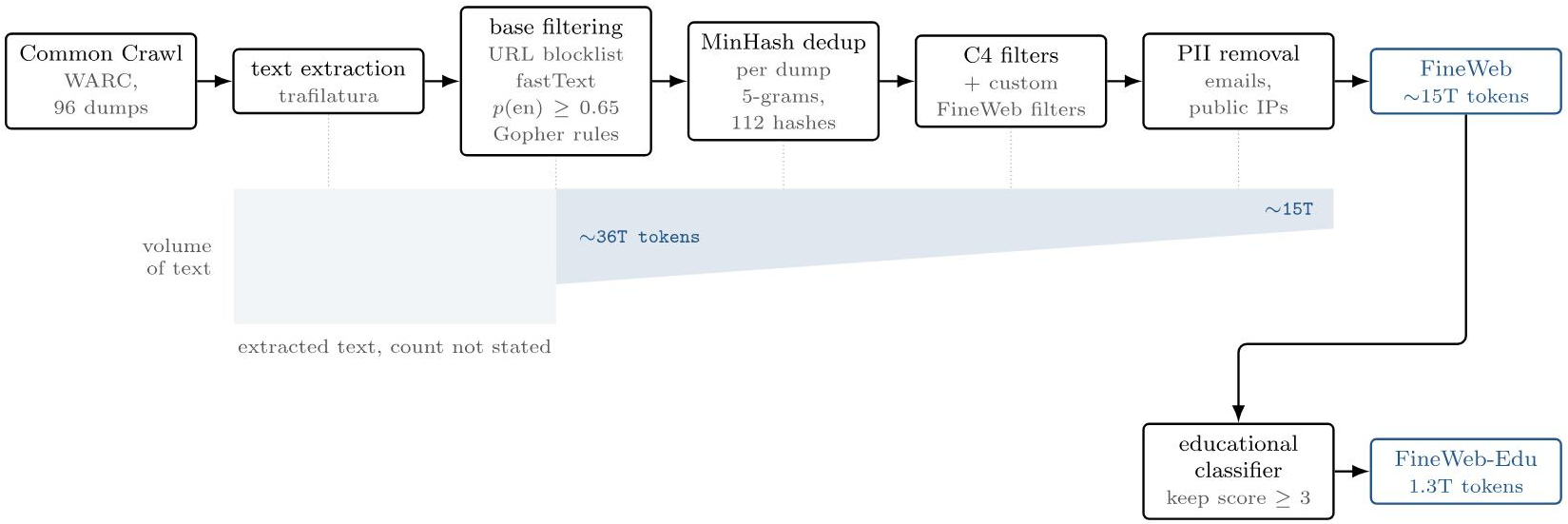
a filter helped if this curve rose at matched compute; nothing else counts

Penedo et al., "The FineWeb Datasets: Decanting the Web for the Finest Text Data at Scale," NeurIPS 2024 Datasets and Benchmarks, arXiv:2406.17557, Figure 10 and Section 3.1.

4 THE PIPELINE

Decanting the Web

The Pipeline



Extraction. trafilatura on the raw HTML, not the crawl's own text files.

Base filtering. URL blocklist, English at $p \geq 0.65$, Gopher's repetition rules.

Deduplication. MinHash over 5-grams, inside each dump.

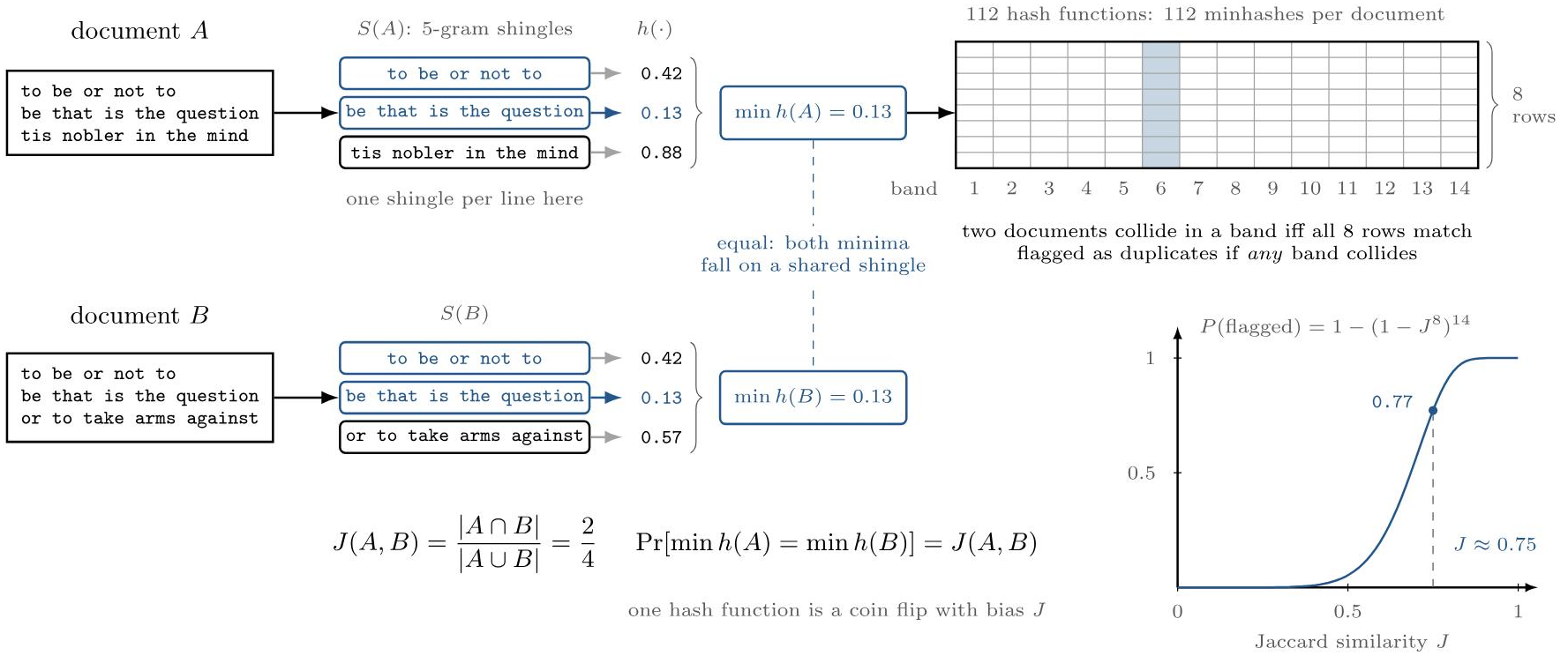
C4 and custom filters. Format rules, plus three FineWeb found itself.

Scrubbing. Emails and IP addresses anonymized for release.

Classifier. An educational score; keep 3 and up (FineWeb-Edu).

Stage order, thresholds and token counts from Penedo et al., 2024, arXiv:2406.17557, Sections 3 and 4.

Deduplication



$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} = \frac{2}{4} \qquad \Pr[\min h(A) = \min h(B)] = J(A,B)$$

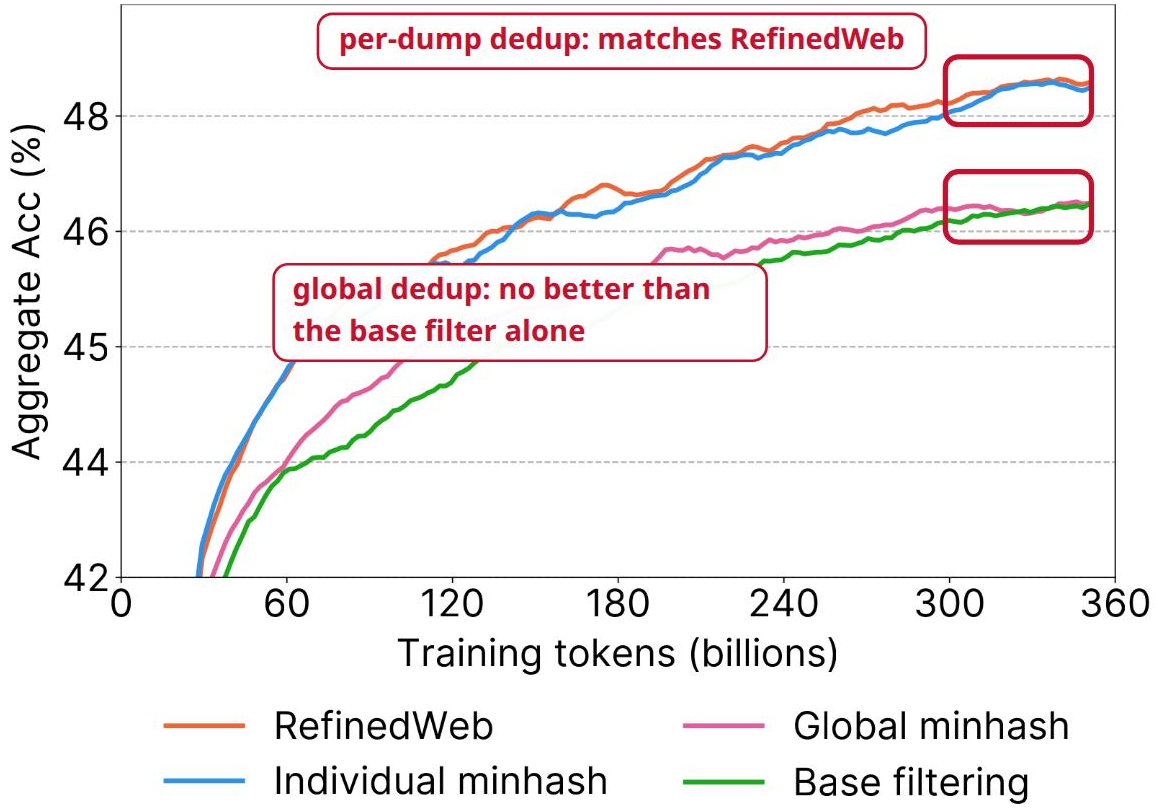
one hash function is a coin flip with bias J

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} \qquad \Pr\left[\min h(A) = \min h(B)\right] = J(A,B) \qquad \Pr[\text{flag}] = 1 - \left(1 - J^8\right)^{14}$$

$J = 0.5$: flagged 5% of the time. $J = 0.75$: 77%. $J = 0.9$: 99.96%.

MinHash: Broder, "On the Resemblance and Containment of Documents," 1997. Settings (5-gram shingles, 112 hash functions, 14 buckets of 8 rows, 75% target) from Penedo et al., 2024, arXiv:2406.17557, Section 3.4 and Appendix E.1. Dedup benefits: Lee et al., "Deduplicating Training Data Makes Language Models Better," ACL 2022, arXiv:2107.06499.

Deduplication

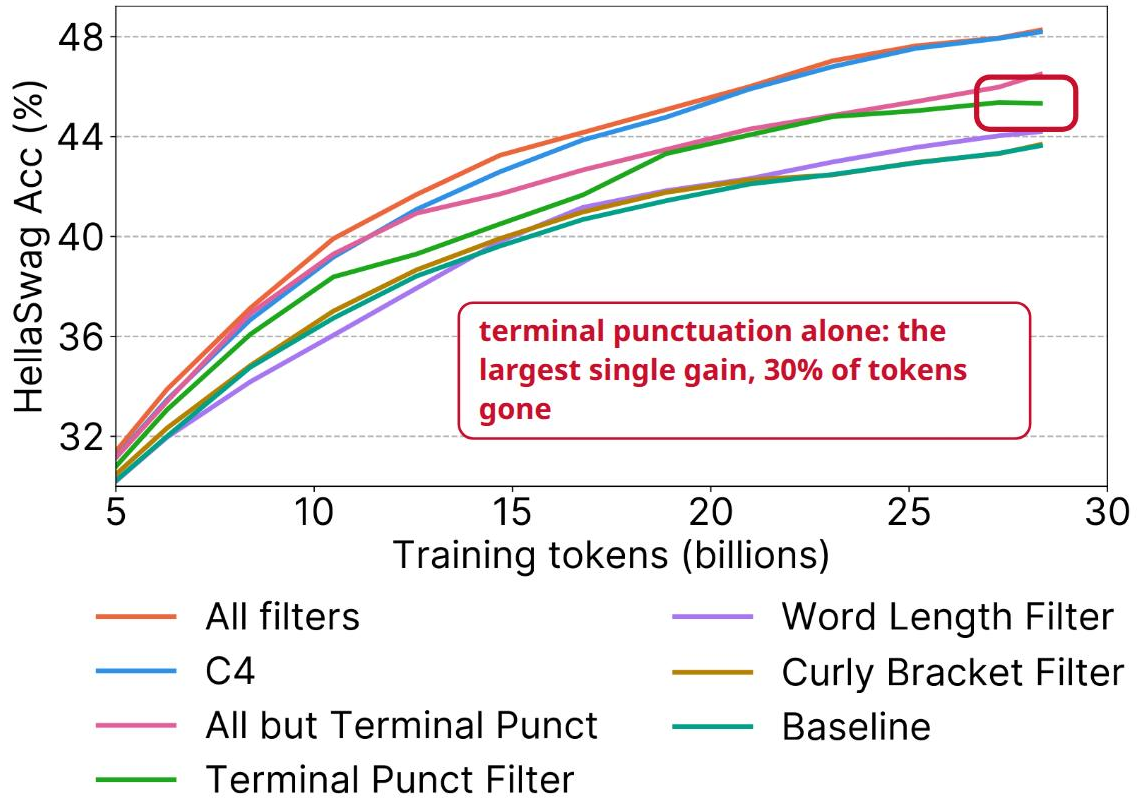


the experiment: MinHash across all 96 dumps at once, against MinHash inside each dump on its own

what global dedup did: 36T tokens down to 4T; on the oldest dump the 10% it kept scored worse than the 90% it removed

Penedo et al., "The FineWeb Datasets," 2024, arXiv:2406.17557, Figures 3 to 5 and Section 3.4.

Heuristic Filters



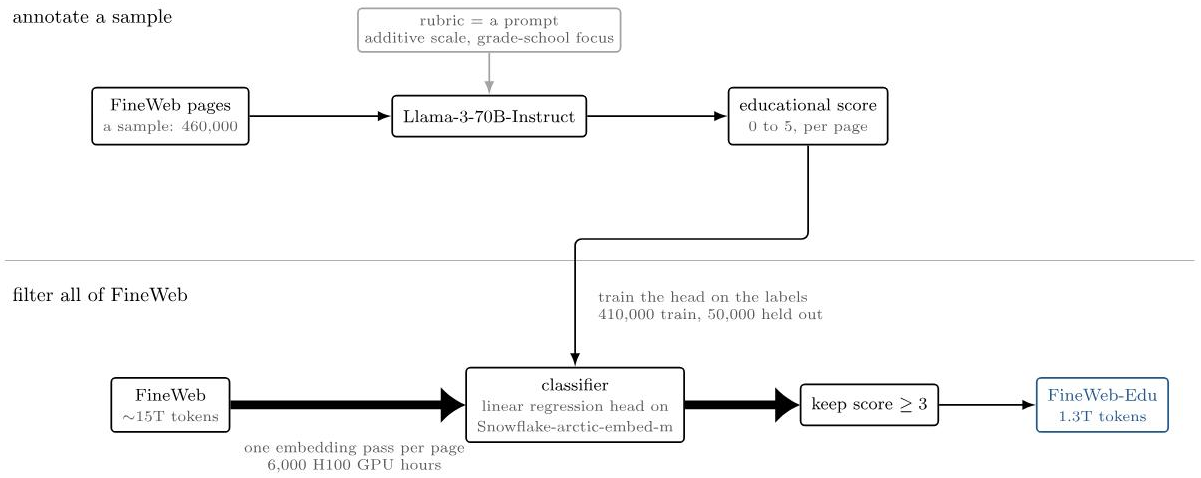
C4: lines end in punctuation, 5 words per line, 3 sentences per page, no lorem ipsum, no curly braces

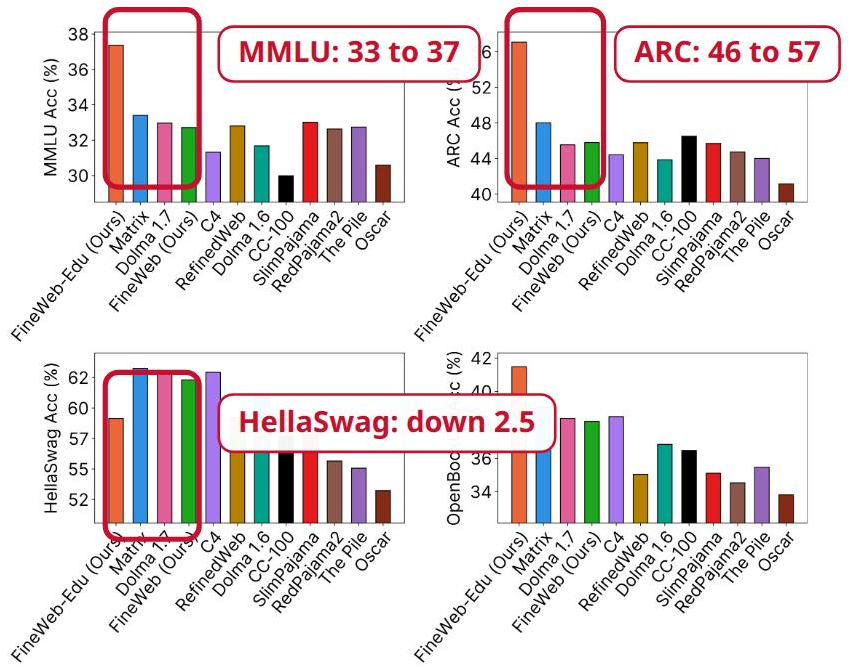
Gopher: caps on repeated lines, paragraphs, and n-grams (30% of lines, 20% of characters)

FineWeb's own three: statistics that separate the good and bad sets of the last slide

Penedo et al., "The FineWeb Datasets," 2024, arXiv:2406.17557, Figure 6 and Sections 3.5 to 3.6. C4 rules: Raffel et al., "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer," JMLR 2020, arXiv:1910.10683, Section 2.2. Gopher rules: Rae et al., 2021, arXiv:2112.11446, Appendix A.1.1.

A Learned Filter





annotator: Llama-3-70B-Instruct scores 460,000 pages 0 to 5 against a written rubric

classifier: a regression head on an embedding model, trained on those scores

threshold: keep score ≥ 3; 1.3T tokens survive of 15T

Penedo et al., "The FineWeb Datasets," 2024, arXiv:2406.17557, Section 4 and Figure 16 (1.71B models, 350B tokens). Classifier: HuggingFaceFW/fineweb-edu-classifier.

Keep or Drop

A bigcreations.nl

... to the online portfolio of BIG creations. Take a look around and get inspired! When a question rises about the possibilities or service, don't be afraid to contact by phone!

score 0.08 · dropped

B abcnews.go.com/Health/texas-man-pioneers-beatless-heart

Craig Lewis, a 55-year-old Texas man who was about 12 hours away from death, became the first human to receive what doctors are calling the "beatless heart."

score 2.76 · dropped

C ni.com/documentation/en/daqexpress/3.2/node-ref/path-type

Returns the type of the specified path, indicating whether it is an absolute, relative, or <Not A Path>. This node does not verify that the path exists on the computer.

score 2.62 · dropped

D mathforum.org/library/drmath/sets/high_circles.html

why 360 degrees? See also the Dr. Math FAQ: segments of circles. Browse High School Conic Sections, Circles. Selected answers: Find the center of a circle. Is a circle a polygon?

score 3.71 · kept

Four documents from HuggingFaceFW/fineweb (sample-10BT shard), scored locally with the released HuggingFaceFW/fineweb-edu-classifier (figures-src/fineweb_edu_score.py). FineWeb-Edu keeps a rounded score of 3 or more. Of 600 documents sampled the same way, the median score was 1.07 and 32 cleared the bar.

5 SYNTHESIS

Past the Data Wall

You Already Did This

Write a short story (3-5 paragraphs) which only uses very simple words that a 3 year old child would likely understand. The story should use the verb "decorate", the noun "thunder" and the adjective "ancient". The story should have the following features: the story should contain at least one dialogue, the story has a bad ending. Remember to only use simple words!

three required words, a toddler vocabulary, a required feature

Once upon a time, in an ancient house, there lived a girl named Lily. She loved to decorate her room with pretty things. One day, she found a big box in the attic. She opened it and saw many shiny decorations. Lily was very happy and decided to use them in her room.
As Lily was decorating her room, the sky outside became dark. There was a loud thunder sound, and Lily got scared. She ran to her mom and said, "Mommy, the thunder is so loud!" Her mom hugged her and said, "Don't worry, it will pass soon."
But the thunder did not stop. It got louder and louder, and the ancient house started to shake. Suddenly, the roof fell down on the pretty decorations. Lily was sad because her room was not pretty anymore. The end.

```
He asked Lily if she could play with Tom. Lily said yes and told her
dad about the pond. She said yes and he put her toy car and her ball
inside.

Tom and Lily were glad they had made her feel better.
```

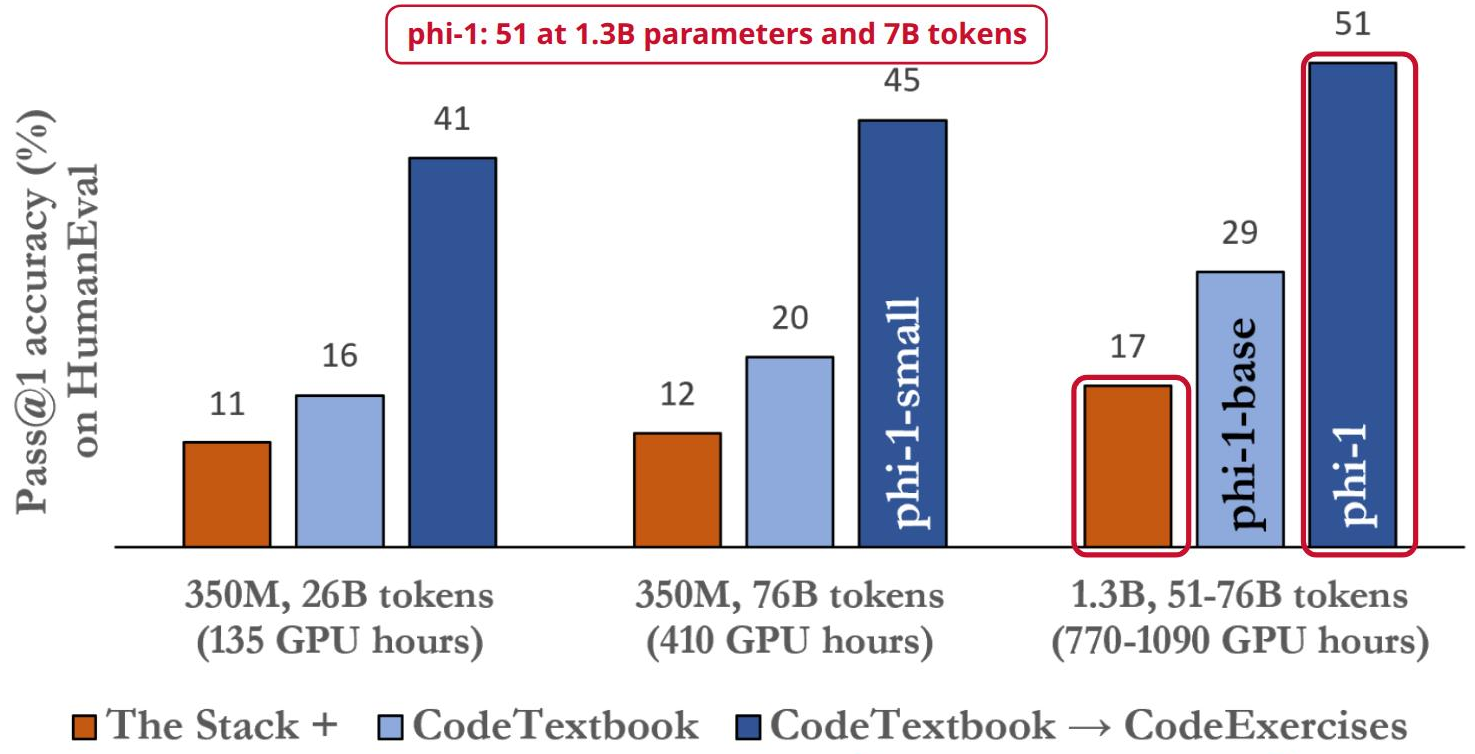
your twin, 10.6M parameters, 82M characters of TinyStories

the authors: GPT-3.5 and GPT-4, from about 1,500 words a 3-year-old knows

your grid row: twin 0.70 on TinyStories, 2.62 on Shakespeare; base 1.43 and 2.13

Eldan and Li, "TinyStories: How Small Can Language Models Be and Still Speak Coherent English?" 2023, arXiv:2305.07759, Section 2. Sample and losses (nats per character): Assignment 2 codebase, results/base-tinystories.json, temperature 0.8, top-k 40.

Textbooks



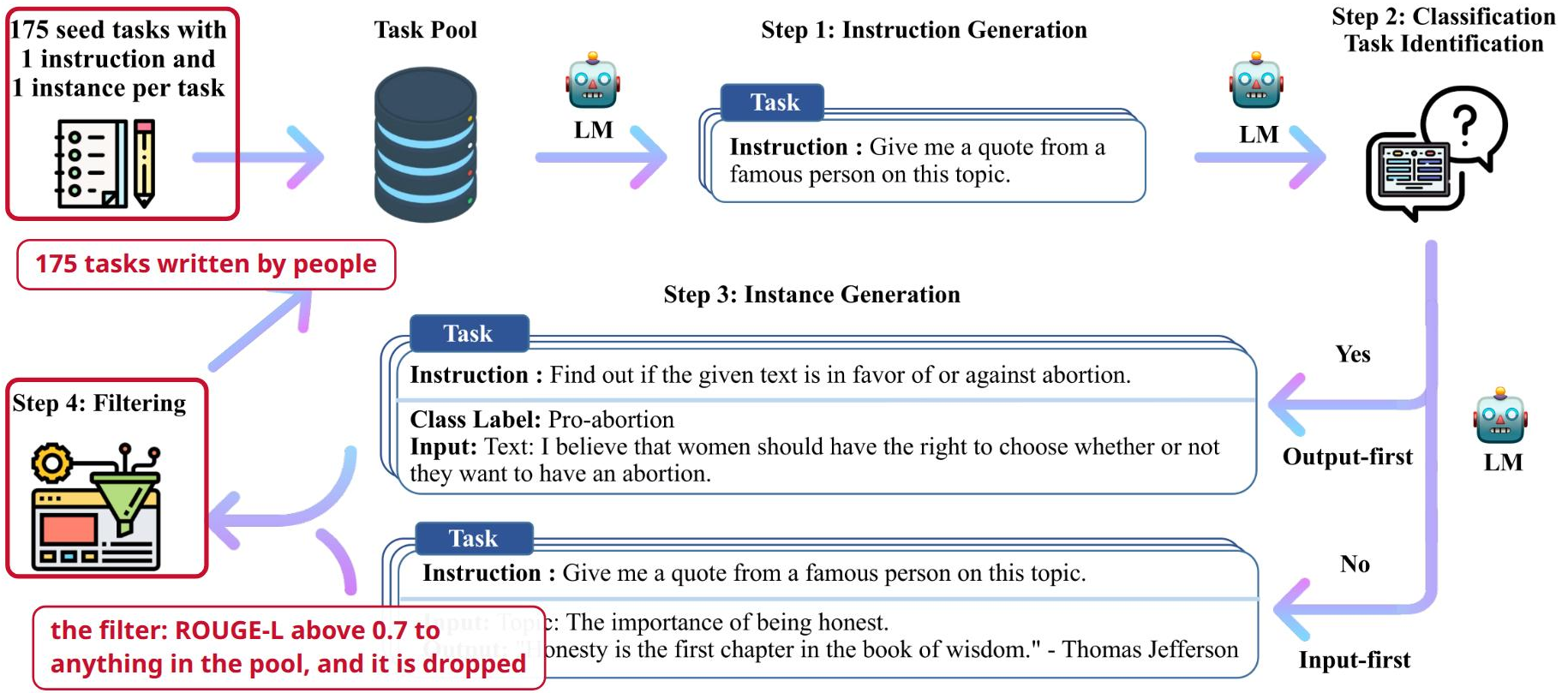
6B tokens of code: The Stack and StackOverflow, kept by a classifier trained on 100k GPT-4 quality labels

under 1B tokens of textbook: written by GPT-3.5 topic by topic, plus 180M tokens of exercises

for scale: StarCoder, 15.5B parameters on 1T tokens, scores 33.6 on the same test

Gunasekar et al., "Textbooks Are All You Need," 2023, arXiv:2306.11644, Figure 2.1 and Table 1 (phi-1: 50.6 pass@1; StarCoder: Li et al., 2023, arXiv:2305.06161).

Self-Instruct



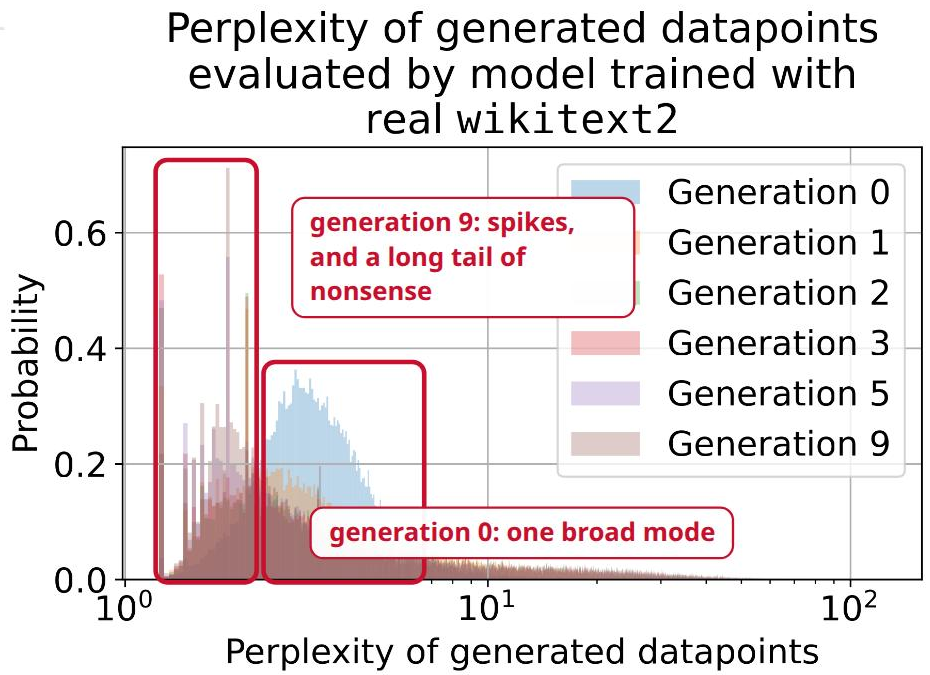
Evol-Instruct: rewrite each instruction harder: add a constraint, deepen, concretize (WizardLM)

Persona Hub: a billion web-mined personas as the axis of variation

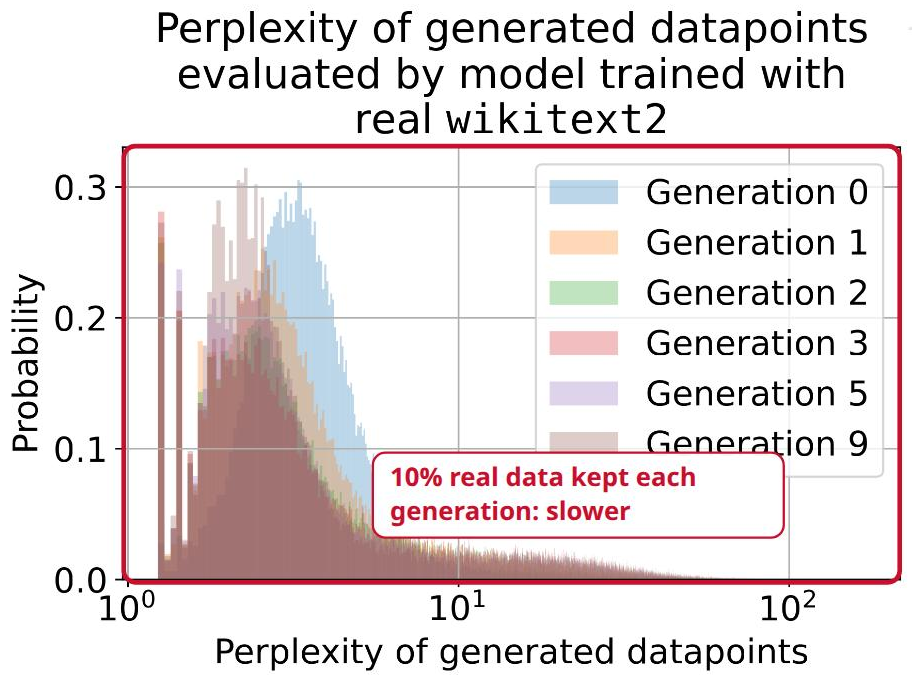
at industrial scale: over 98% of Nemotron-4 340B's alignment data is synthetic

Wang et al., "Self-Instruct: Aligning Language Models with Self-Generated Instructions," ACL 2023, arXiv:2212.10560, Figure 2. Xu et al., WizardLM, 2023, arXiv:2304.12244. Ge et al., Scaling Synthetic Data Creation with 1,000,000,000 Personas, 2024, arXiv:2406.20094. Nvidia, Nemotron-4 340B Technical Report, 2024, arXiv:2406.11704.

Collapse



(a) No data preserved



(b) 10% data preserved

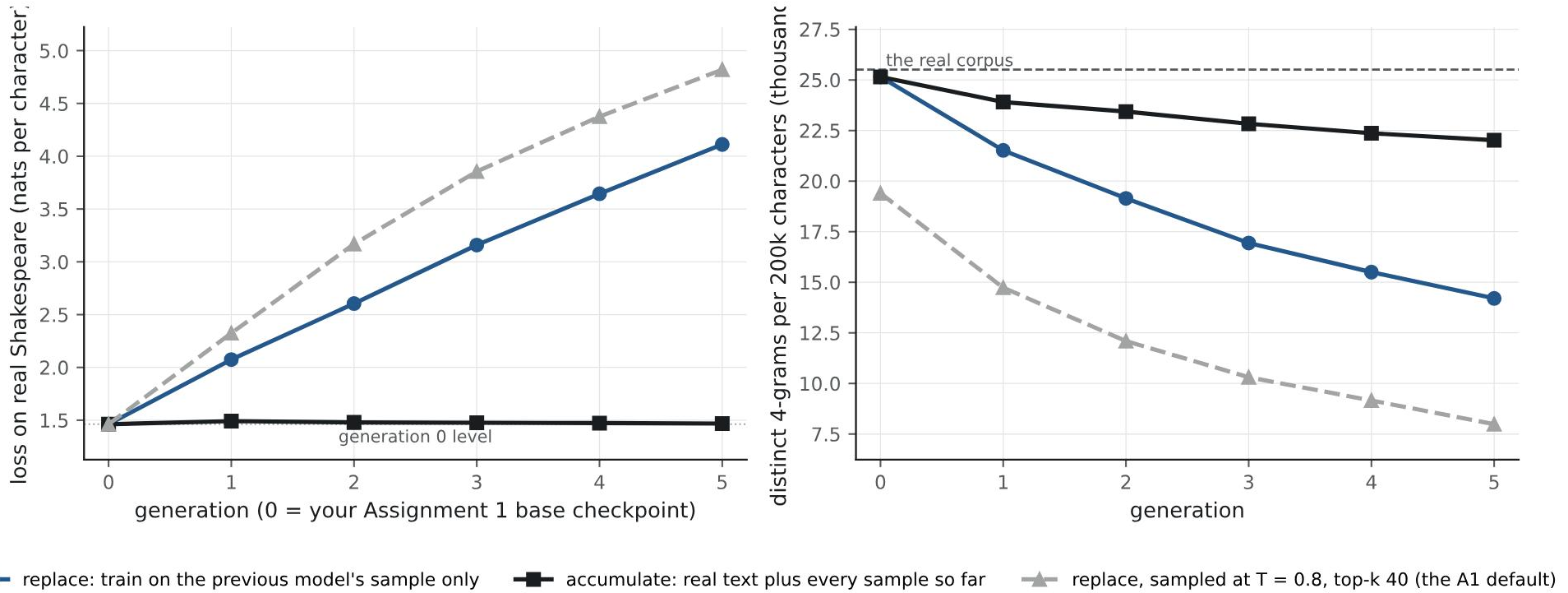
statistical error: a finite sample never contains the rare events; the tails go first

approximation error: the model cannot represent the whole distribution, and errs the same way each time

recursion: generation k learns generation $k-1$'s mistakes as truth

Shumailov et al., "The Curse of Recursion: Training on Generated Data Makes Models Forget," 2023, arXiv:2305.17493, Figure 11; the three error sources follow the published version, "AI models collapse when trained on recursively generated data," Nature 631, 2024.

Collapse

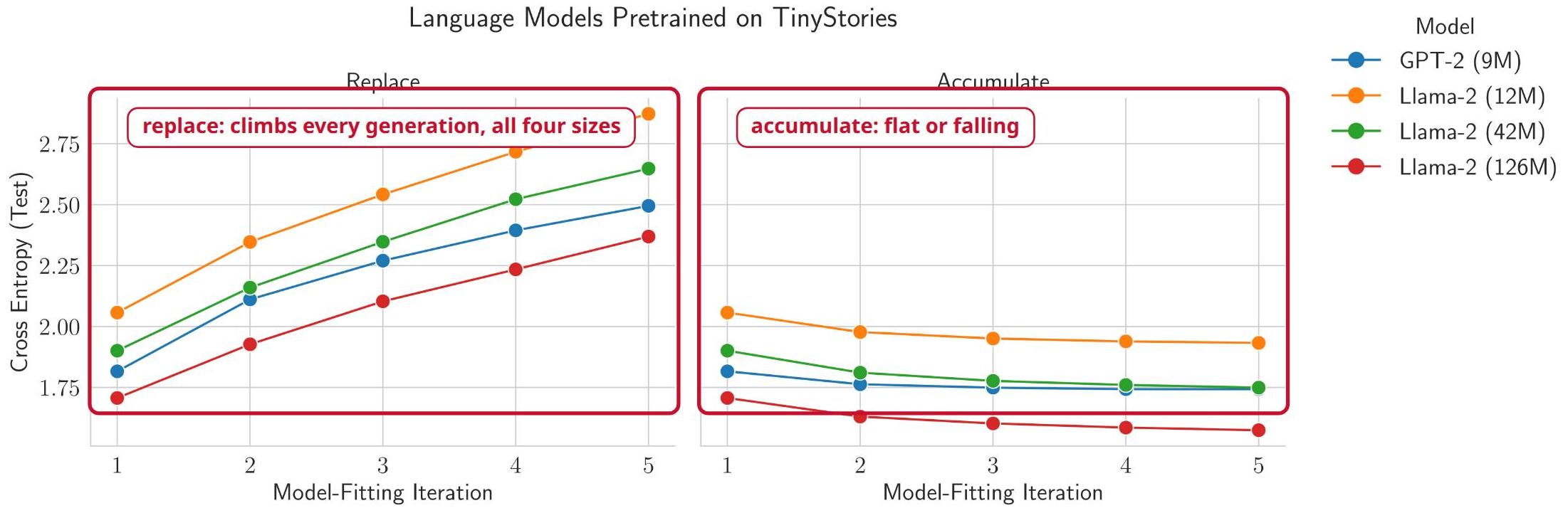


the run: generation 0 is your out/base checkpoint; each generation samples 1.0M characters from the previous model and retrains from scratch; the exam is the real held-out split

```
MENENIUS:
Who has no more teet than you shall appear him honour with him?
First Citizen:
Ay, woman:
I fear the soul of words.
Second Citizen:
```

Assignment 1 codebase, configs/base.json unchanged, 5,000 iterations per generation; figures-src/run_collapse.py and collapse-results.json. The first two arms sample at temperature 1.0 with no top-k, so the only truncation is the finite sample.

Accumulate



keep the real data: synthetic text joins the mixture; it never replaces it

verify what you keep: tests, checkers, judges; a filtered sample is a different object from a raw one

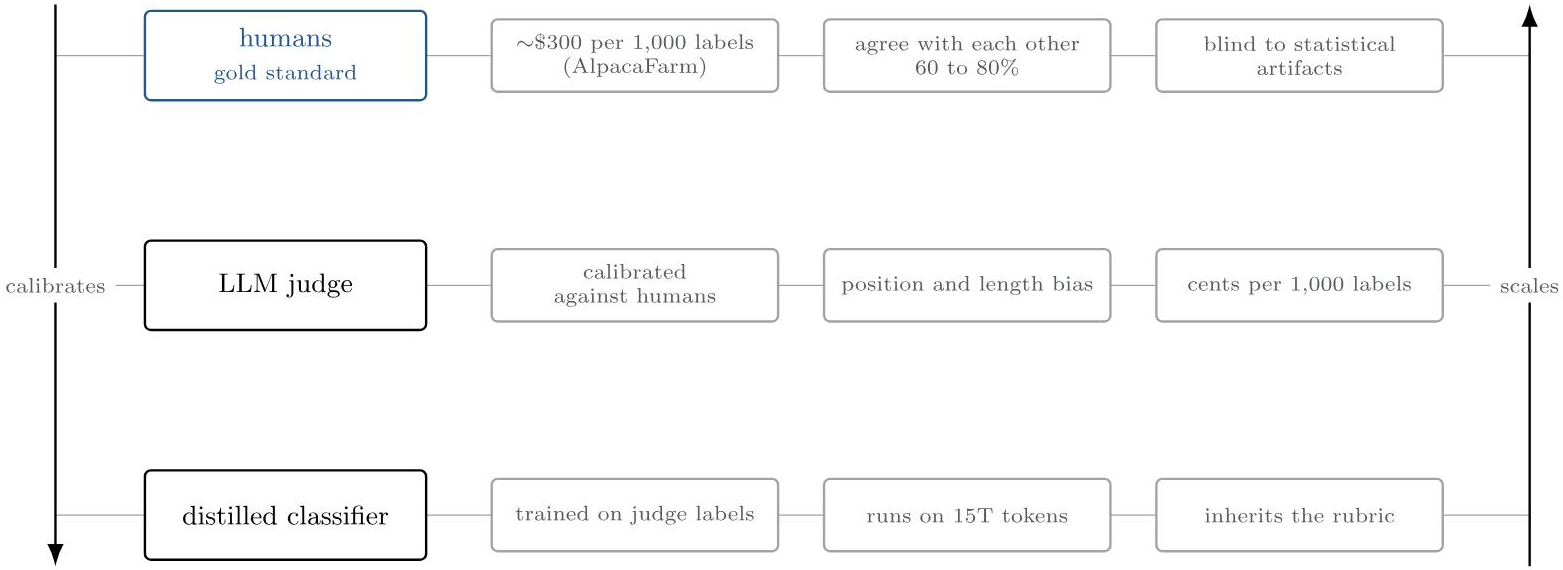
tag provenance: know which tokens a model wrote, so the next crawl can tell

Gerstgrasser et al., "Is Model Collapse Inevitable? Breaking the Curse of Recursion by Accumulating Real and Synthetic Data," 2024, arXiv:2404.01413, Figure 2 (language models pretrained on TinyStories).

6 JUDGES IN THE PIPELINE

Who Grades the Data

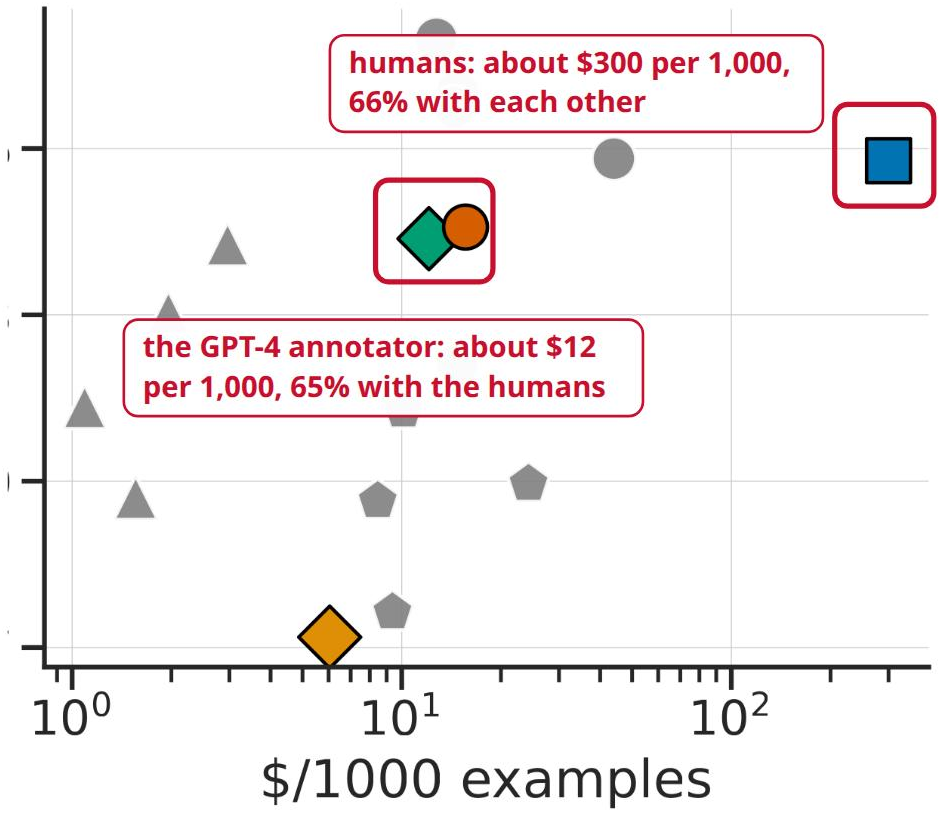
Three Meters



Humans set the meter, the judge reads it at scale, the classifier reads it on the whole web. Each inherits the biases above it.

LIMA: Zhou et al., 2023, arXiv:2305.11206. Judge agreement: Zheng et al., 2023, arXiv:2306.05685. Costs: Dubois et al., AlpacaFarm, 2023, arXiv:2305.14387. Classifier: Penedo et al., 2024, arXiv:2406.17557.

Calibrating the Judge



`y: agreement with the majority of three human votes; x: dollars per 1,000 pairwise labels, log scale`

the trade: 25x cheaper, at the human-human agreement level

the bias: the annotator prefers the longer answer 64% of the time; humans 62%; a policy trained against it learns to be long

Dubois et al., "AlpacaFarm: A Simulation Framework for Methods that Learn from Human Feedback," NeurIPS 2023, arXiv:2305.14387, Figure 4 (plot only; legend markers named in the callouts), Section 4.3 and Appendix C.2.

Reading

Required

Penedo et al., The FineWeb Datasets, arXiv:2406.17557. The pipeline of section 2, with the ablation behind every stage.

Optional

Li et al., DataComp-LM (2406.11794). Xie et al., DoReMi (2305.10429). Wang et al., Self-Instruct (2212.10560). Shumailov et al., The Curse of Recursion (2305.17493). Gunasekar et al., Textbooks Are All You Need (2306.11644).

Assignment 2 is due Tuesday, September 15. The judge you run in Part 4, with its order swap, is the meter from today. Zheng and Miller from Thursday's list still stand as its reading.