DATA 8005 Advanced Natural Language Processing

LLM - Data and Evaluation

Jianrui Wu, Tianle Li

Fall 2024

Outline

- Background
- The FineWeb Datasets
- DoReMi: Pre-training data reweighting
- Post-training Data in Llama 3
- Evaluations of LLama 3 and OpenAI o1
- Chatbot Arena: Human evaluation

Data

Pre-training | Post-training

Evaluation

Static benchmarks | Live questions

Background

- LLMs like Llama 3 show high performance on different benchmarks, such as:
  - MMLU: a general multitask benchmark
  - GPQA: a Q&A dataset for science domains
  - HumanEval: a benchmark for coding ability
- How to get these datasets to train and evaluate LLMs?

	Meta Llama 3 8B	Gemma 7B - It Measured	Mistral 7B Instruct Measured
MMLU 5-shot	**68.4**	53.3	58.4
GPQA 0-shot	**34.2**	21.4	26.3
HumanEval 0-shot	**62.2**	30.5	36.6
GSM-8K 8-shot, CoT	**79.6**	30.6	39.9
MATH 4-shot, CoT	**30.0**	12.2	11.0

	Meta Llama 3 70B	Gemini Pro 1.5 Published	Claude 3 Sonnet Published
MMLU 5-shot	**82.0**	81.9	79.0
GPQA 0-shot	39.5	**41.5** CoT	38.5 CoT
HumanEval 0-shot	**81.7**	71.9	73.0
GSM-8K 8-shot, CoT	**93.0**	91.7 11-shot	92.3 0-shot
MATH 4-shot, CoT	50.4	**58.5** Minerva prompt	40.5

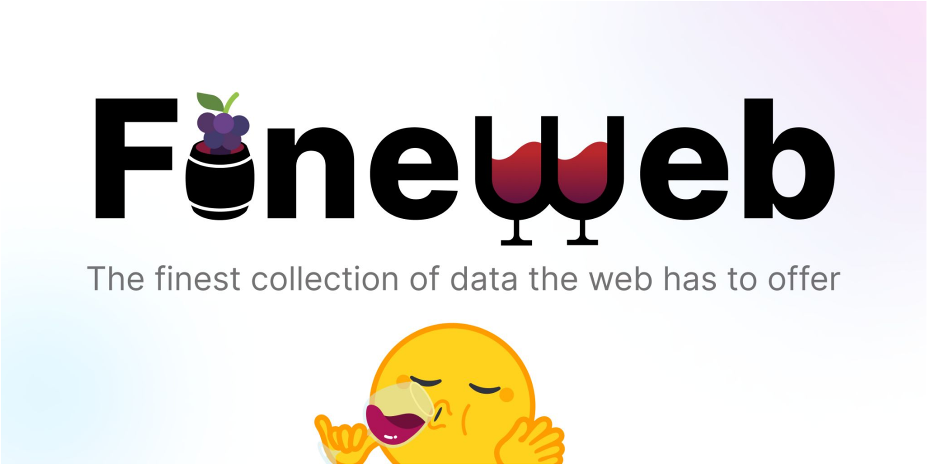
The FineWeb Datasets: Decanting the Web for the Finest Text Data at Scale

Introduction

- FineWeb
  - A web-based pre-training dataset derived from 96 Common Crawl
  - 15 trillion tokens
- Pipeline
  - Text extraction
  - Base filtering
  - Deduplication
  - C4's filters
  - heuristic filters
- FineWeb-Edu
  - An educational dataset filtered from FineWeb
  - 1.3 trillion tokens

Setup

- A series of ablation experiments
  - Models are identical apart from the data they are trained on
  - Evaluated on the same set of downstream task benchmark datasets
  - Train two models for each dataset version

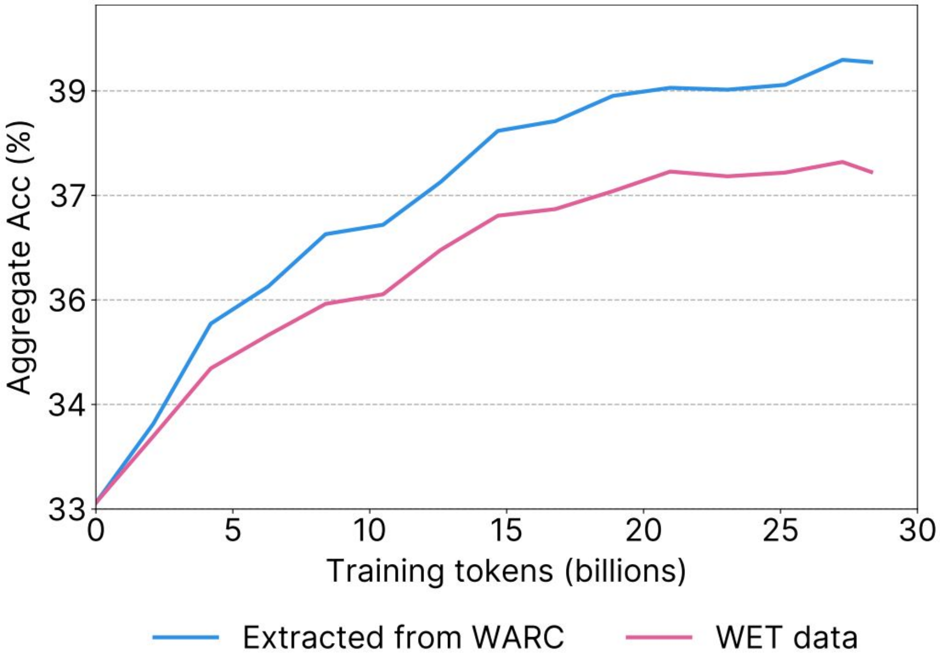


Pipeline I - Text Extraction

- Two formats in Common Crawl data
  - WARC (Web ARChive format): raw data, the full page HTML and request metadata
  - WET (WARC Encapsulated Text): a text-only version
- WET retained too much boilerplate and menu text
- Extracting the text content from the WARC files using trafilatura

Pipeline 1 - Text Extraction

- Trafilatura-extracted WARC vs WET

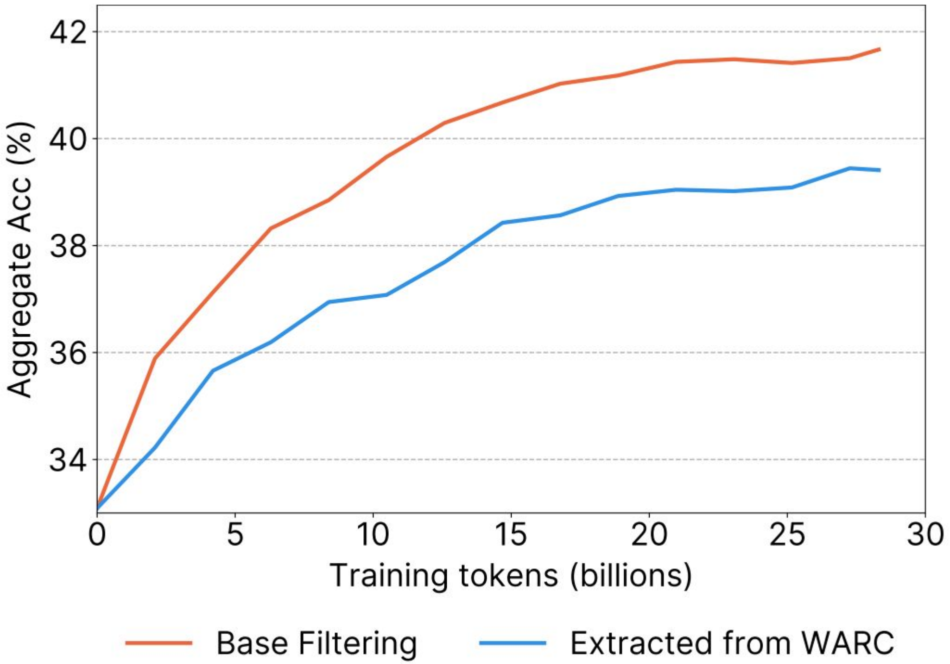


Pipeline 2 - Base Filtering

- URL filtering: using a blocklist to remove adult content
- A fastText language classifier: keep only English text with a score >= 0.65
- Quality and repetition filters from MassiveText

Pipeline 2 - Base Filtering

- Base filtered WARC vs Unfiltered WARC data

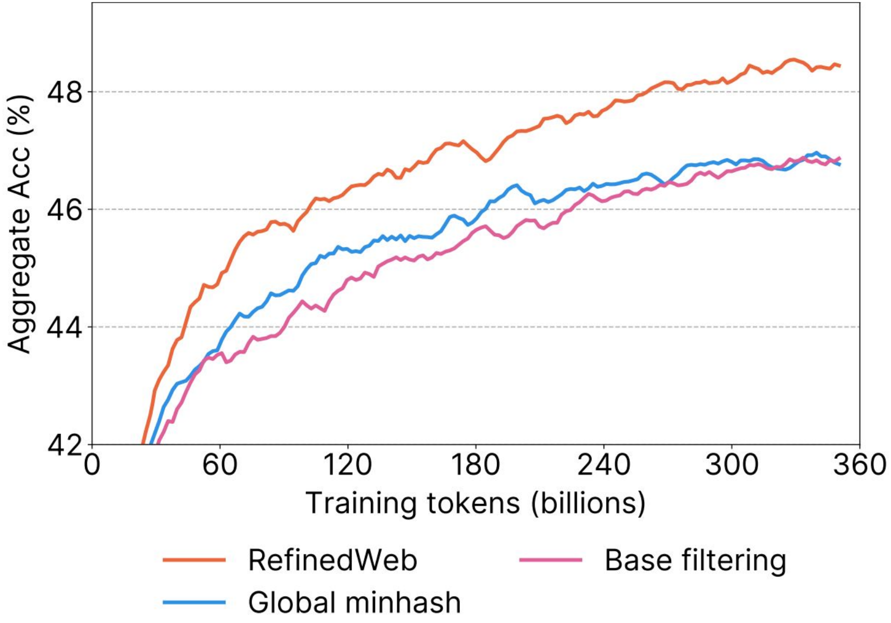


Pipeline 3 - Deduplication

- Duplicates: aggregators, mirrors, templated pages …
- Removing duplicates:
  - improve model performance
  - reduce model memorization
- MinHash: a fuzzy hash-based deduplication technique
  - collect each document's 5-grams
  - using 112 hash functions
  - split into 14 buckets of 8 hashes each
  - targeting documents that are at least 75% similar

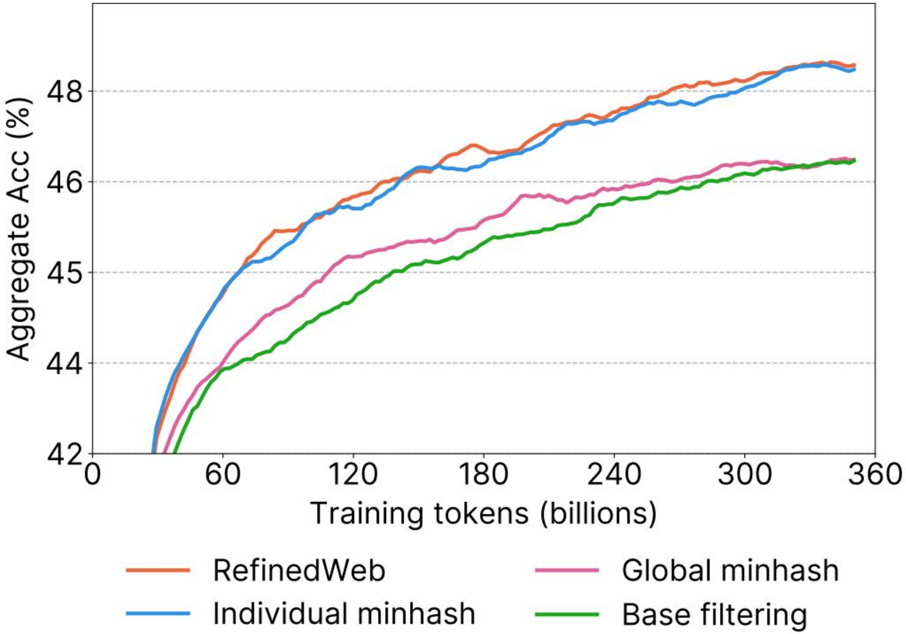
Pipeline 3 - Deduplication

- Global MinHash: apply MinHash to the entire dataset (all 96 snapshots)
- From the most recent snapshot to the oldest snapshot
- little improvement



Pipeline 3 - Deduplication

- Individual Minhash: individually deduplicating each snapshot
- Improve performance: remove large clusters of duplicates
- Harm performance: remove a small number of duplicates

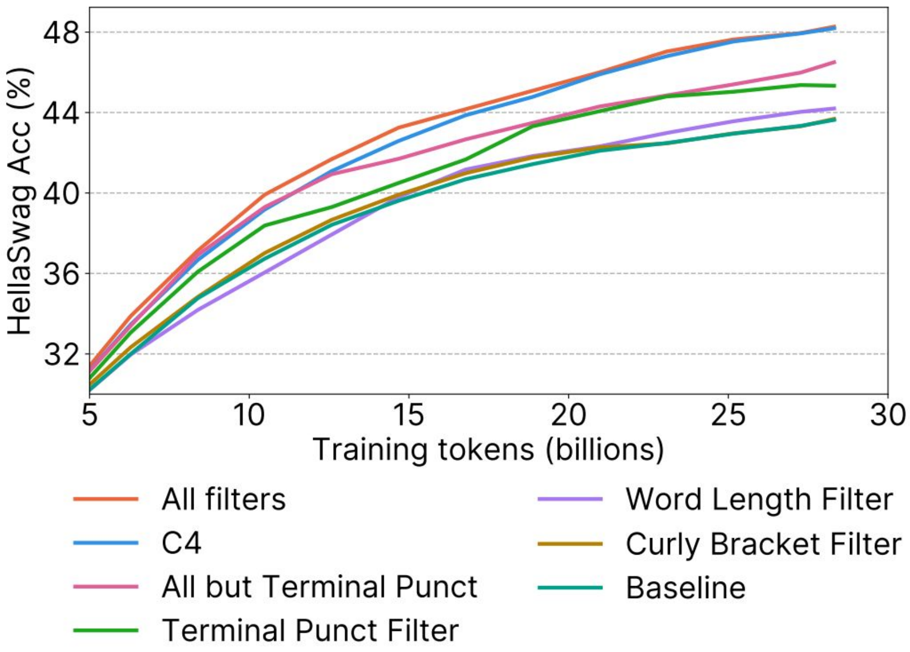


Pipeline 4 - C4's Filters

- C4 dataset: smaller but stronger. Why?
- Dropping lines that
  - without a terminal punctuation mark
  - mentioned javascript
  - had "terms-of-use"/"cookie policy" statements
- Dropping documents that were too short or that contained "lorem ipsum" or a curly bracket ({)

Pipeline 4 - C4's Filters

- Terminal punctuation filter gives the biggest boost but removes too much data (30%)

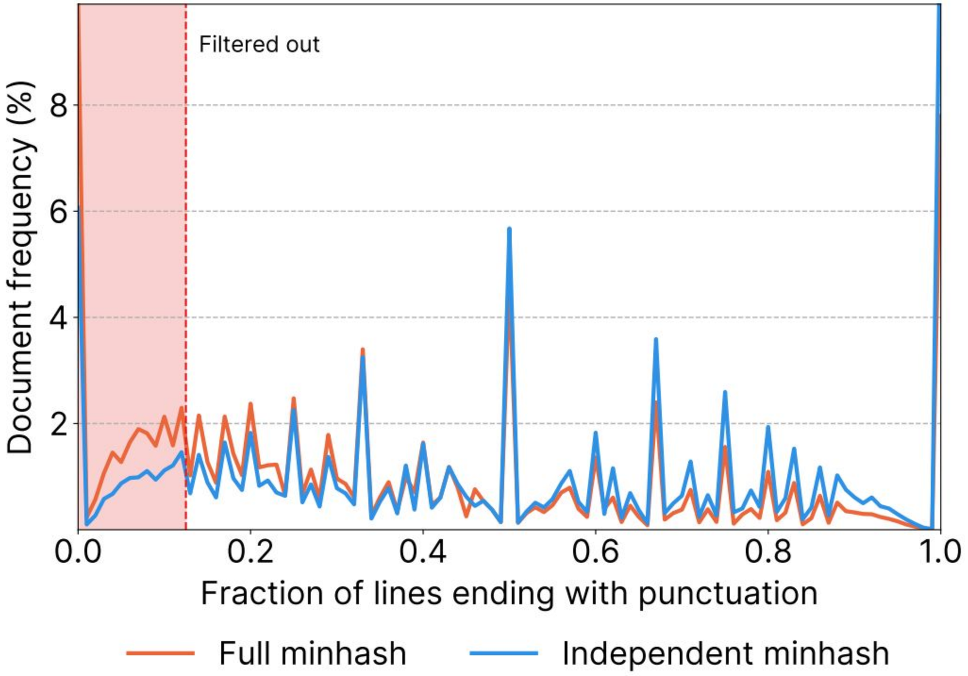


Pipeline 5 - Heuristic Filters

- A more systematic process for designing heuristic filters
  - collecting over 50 high-level statistics
  - "high-quality" and "low-quality" datasets
  - identified metrics for which the distribution of values differed significantly across the two datasets
- Three heuristic filters were chosen:
  - the fraction of lines ending with punctuation is <= 0.12
  - the fraction of characters in duplicated lines is >= 0.1
  - the fraction of lines shorter than 30 characters is >= 0.67

Pipeline 5 - Heuristic Filters

- Impact of Heuristic Filters on 2013-48 Crawl

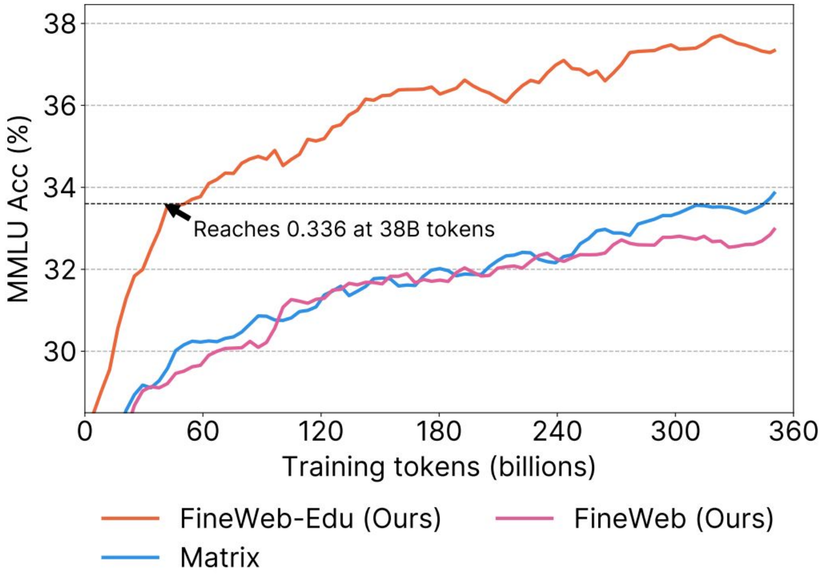


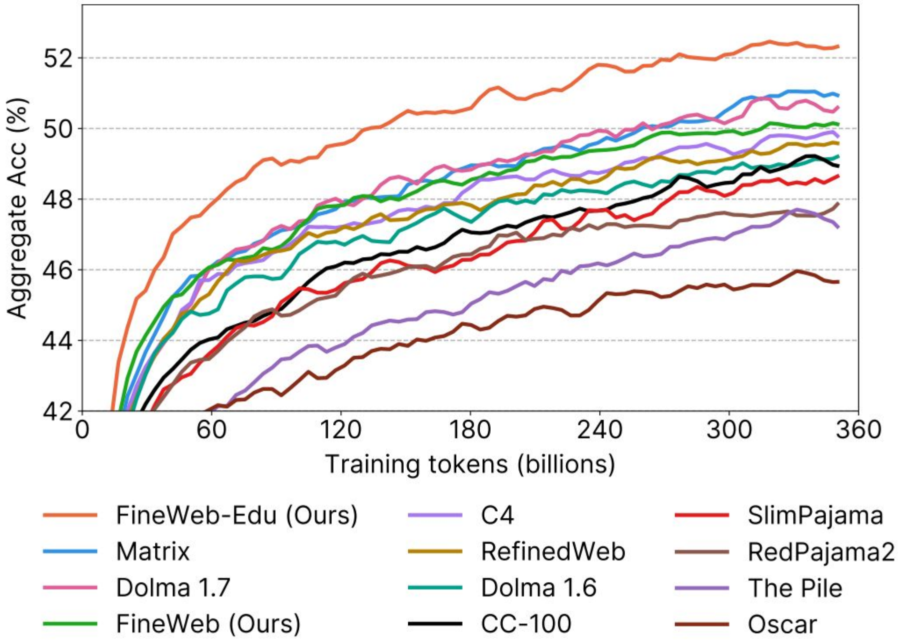
FineWeb-Edu

- Identifying educational content
  - synthetic annotations generated by Llama-3- 70B-Instruct
  - train a linear regression model as an educational quality classifier
  - determine the threshold

Performance

- Performance of FineWeb-Edu and FineWeb
- A 1.82B model trained on 350 billion tokens

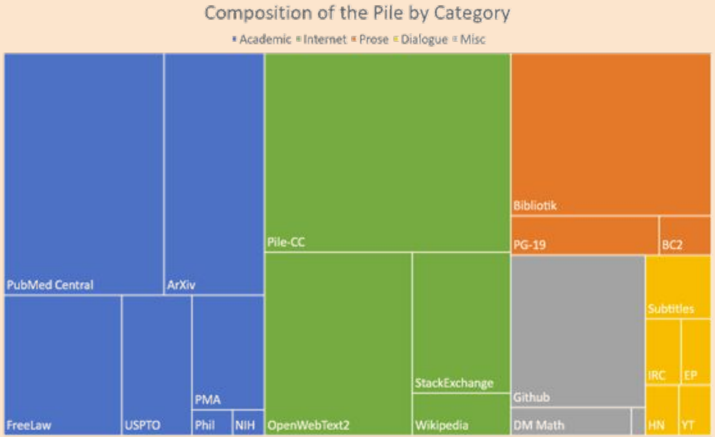




DoReMi: Optimizing Data Mixtures Speeds Up Language Model Pretraining

Pre-training data come from many sources…

Baseline: The Pile



(Effective Size = Raw Size x Epochs)

Domain	Baseline	DoReMi (280M)	Difference	Domain	Baseline	DoReMi (280M)	Difference
Pile-CC	0.1121	0.6057	+0.4936	DM Mathematics	0.0198	0.0018	-0.0180
YoutubeSubtitles	0.0042	0.0502	+0.0460	Wikipedia (en)	0.0919	0.0699	-0.0220
PhilPapers	0.0027	0.0274	+0.0247	OpenWebText2	0.1247	0.1019	-0.0228
HackerNews	0.0075	0.0134	+0.0059	Github	0.0427	0.0179	-0.0248
Enron Emails	0.0030	0.0070	+0.0040	FreeLaw	0.0386	0.0043	-0.0343
EuroParl	0.0043	0.0062	+0.0019	USPTO Backgrounds	0.0420	0.0036	-0.0384
Ubuntu IRC	0.0074	0.0093	+0.0019	Books3	0.0676	0.0224	-0.0452
BookCorpus2	0.0044	0.0061	+0.0017	PubMed Abstracts	0.0845	0.0113	-0.0732
NIH ExPorter	0.0052	0.0063	+0.0011	StackExchange	0.0929	0.0153	-0.0776
OpenSubtitles	0.0124	0.0047	-0.0077	ArXiv	0.1052	0.0036	-0.1016
Gutenberg (PG-19)	0.0199	0.0072	-0.0127	PubMed Central	0.1071	0.0046	-0.1025

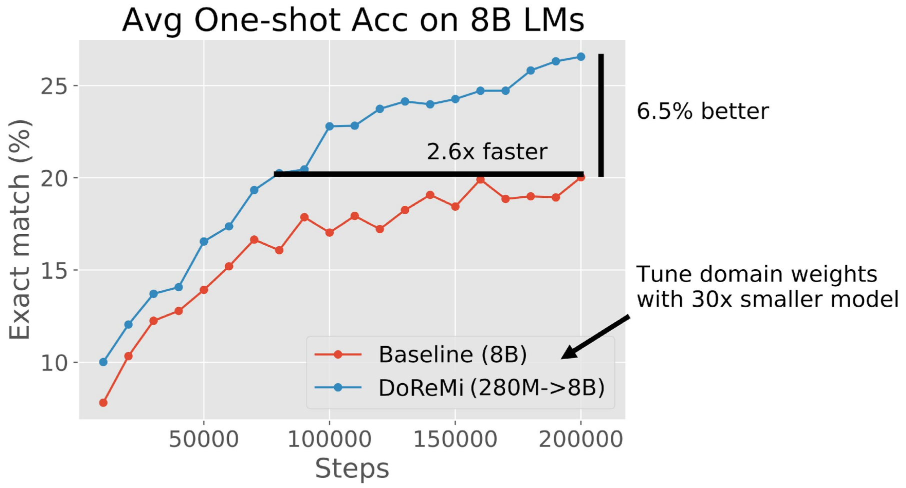
Left: arXiv:2101.00027

Data reweighting – by training a small proxy

DoReMi

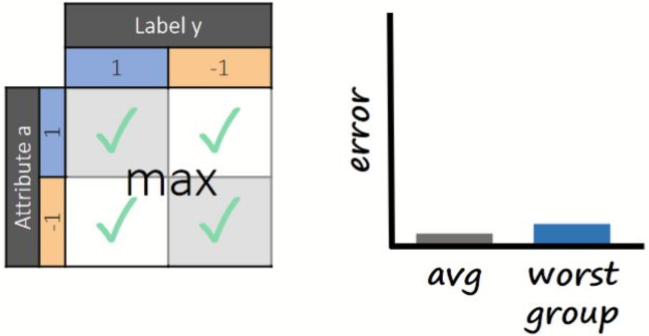
1. Train a small reference model $p_\theta(x)$ with default domain weights α_{ref}
2. Train a small proxy model $p_{\text{ref}}(x)$ with **<u>G</u>roup <u>D</u>istributionally <u>R</u>obust <u>O</u>ptimization** to obtain new domain weights $\bar{\alpha}$
3. Train large model with the new domain weights $\bar{\alpha}$

Domain	Baseline	DoReMi (280M)	Difference	Domain	Baseline	DoReMi (280M)	Difference
Pile-CC	0.1121	0.6057	+0.4936	DM Mathematics	0.0198	0.0018	-0.0180
YoutubeSubtitles	0.0042	0.0502	+0.0460	Wikipedia (en)	0.0919	0.0699	-0.0220
PhilPapers	0.0027	0.0274	+0.0247	OpenWebText2	0.1247	0.1019	-0.0228
HackerNews	0.0075	0.0134	+0.0059	Github	0.0427	0.0179	-0.0248
Enron Emails	0.0030	0.0070	+0.0040	FreeLaw	0.0386	0.0043	-0.0343
EuroParl	0.0043	0.0062	+0.0019	USPTO Backgrounds	0.0420	0.0036	-0.0384
Ubuntu IRC	0.0074	0.0093	+0.0019	Books3	0.0676	0.0224	-0.0452
BookCorpus2	0.0044	0.0061	+0.0017	PubMed Abstracts	0.0845	0.0113	-0.0732
NIH ExPorter	0.0052	0.0063	+0.0011	StackExchange	0.0929	0.0153	-0.0776
OpenSubtitles	0.0124	0.0047	-0.0077	ArXiv	0.1052	0.0036	-0.1016
Gutenberg (PG-19)	0.0199	0.0072	-0.0127	PubMed Central	0.1071	0.0046	-0.1025



Group DRO: To minimize the worst-group loss

$$\mathcal{R}_{\text{gDRO}}(w) = \underbrace{\max_{g'\in\mathcal{G}}}_{\text{worst-group}} \underbrace{\hat{\mathbb{E}}_{(x,y,g)}\left[\ell(w;(x,y)) \mid g = g'\right]}_{\text{average loss for each group } g'}$$



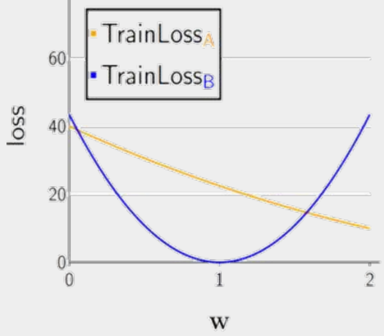
*In reality you also need proper regularization

Toy Example

Model: y'=wx
Loss: $(y'-y)^2$

x	y	g
1	4	A
2	8	A
5	5	B
6	6	B
7	7	B
8	8	B

per-group loss



Group DRO: To minimize the worst-group loss

$$\mathcal{R}_{\text{gDRO}}(w) = \underbrace{\max_{g' \in \mathcal{G}}}_{\text{worst-group}} \underbrace{\hat{\mathbb{E}}_{(x,y,g)} \left[\ell(w; (x,y)) \mid g = g'\right]}_{\text{average loss for each group } g'}$$

How to minimize it at training time (**online optimization**)?

$$\boxed{\min_{\theta \in \Theta}} \boxed{\sup_{q \in \Delta_m}} \sum_{g=1}^{m} \boxed{q_g} \mathbb{E}_{(x,y) \sim P_g} [\ell(\theta; (x,y))].$$

alternating optimization

Input: Step sizes η_q, η_θ; P_g for each $g \in \mathcal{G}$
Initialize $\theta^{(0)}$ and $q^{(0)}$
for $t = 1, \ldots, T$ **do**
  $g \sim \text{Uniform}(1, \ldots, m)$ // Choose a group g at random
  $x, y \sim P_g$ // Sample x, y from group g
  $q' \leftarrow q^{(t-1)}$; $q'_g \leftarrow q'_g \exp(\eta_q \ell(\theta^{(t-1)}; (x,y)))$ // Update weights for group g
  $q^{(t)} \leftarrow q' / \sum_{g'} q'_{g'}$ // Renormalize q
  $\theta^{(t)} \leftarrow \theta^{(t-1)} - \eta_\theta q_g^{(t)} \nabla \ell(\theta^{(t-1)}; (x,y))$ // Use q to update θ
end

Group DRO, the DoReMi way

1. Train a small reference model $p_\theta(x)$ with default domain weights α_{ref}
2. Train a small proxy model $p_{\text{ref}}(x)$ with **<u>G</u>roup <u>D</u>istributionally <u>R</u>obust <u>O</u>ptimization** to obtain new domain weights $\bar{\alpha}$
3. Train large model with the new domain weights $\bar{\alpha}$

Our purpose!

NLL of proxy　NLL of ref

$$\min_{\theta} \max_{\alpha \in \Delta^k} L(\theta, \alpha) := \sum_{i=1}^{k} \boxed{\alpha_i} \cdot \left[\frac{1}{\sum_{x \in D_i} |x|} \sum_{x \in D_i} (\ell_\theta(x) - \ell_{\text{ref}}(x)) \right]$$

D_i: domain i　x: text example　|x|: example length

Group DRO, the DoReMi way

Require: Domain data $D_1, \ldots, D_k$, number of training steps T, batch size b, step size η, smoothing parameter $c \in [0,1]$ (e.g., $c = 10^{-3}$ in our implementation).

Initialize proxy weights θ_0

Initialize $\alpha_0 = \frac{1}{k}\mathbb{1}$

Sample minibatch $B = \{x_1, \ldots, x_j\}$ of size b from P_u, where $u = \frac{1}{k}\mathbb{1}$

for t from 1 to T **do**

Let $|x|$ be the token length of example x ($|x| \le L$)

Compute per-domain excess losses for each domain $i \in \{1, 2, \ldots, k\}$ ($\ell_{\theta,j}(x)$ is j-th token-level loss):

$$\lambda_t[i] \leftarrow \frac{1}{\sum_{x \in B \cap D_i} |x|} \sum_{x \in B \cap D_i} \sum_{j=1}^{|x|} \max\{\ell_{\theta_{t-1},j}(x) - \ell_{\text{ref},j}(x), 0\}$$

Update (exp is entrywise): $\alpha'_t \leftarrow \alpha_{t-1} \exp(\eta \lambda_t)$

Renormalize and smooth : $\alpha_t \leftarrow (1-c)\frac{\alpha'_t}{\sum_{i=1}^{k} \alpha'_t[i]} + cu$

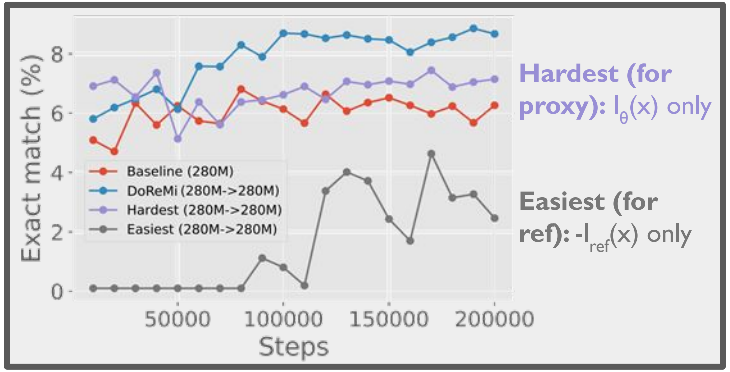
Update proxy model weights θ_t for the objective $L(\theta_{t-1}, \alpha_t)$ (using Adam, Adafactor, etc.)

end for

return $\frac{1}{T}\sum_{t=1}^{T} \alpha_t$

alternating optimization **excess loss**

$$\min_{\theta} \max_{\alpha \in \Delta^k} L(\theta, \alpha) := \sum_{i=1}^{k} \alpha_i \cdot \left[\frac{1}{\sum_{x \in D_i} |x|} \sum_{x \in D_i} \left(\ell_\theta(x) - \ell_{\text{ref}}(x)\right) \right]$$



Hardest (for proxy): $l_\theta(x)$ only

Easiest (for ref): $-l_{\text{ref}}(x)$ only

DoReMi, continued

1. Train a small reference model $p_\theta(x)$ with default domain weights α_{ref}

2. Train a small proxy model $p_{\text{ref}}(x)$ with **<u>G</u>roup <u>D</u>istributionally <u>R</u>obust <u>O</u>ptimization** to obtain new domain weights $\bar{\alpha}$

3. Train large model with the new domain weights $\bar{\alpha}$

end for
return $\boxed{\frac{1}{T}\sum_{t=1}^{T}\alpha_t}$

average weights $\bar{\alpha}$ over the training trajectory

Iterated DoReMi

$\alpha_{\text{ref}} \rightarrow \bar{\alpha}$

$\downarrow$

$\alpha_{\text{ref}} \rightarrow \bar{\alpha}$

$\downarrow$

$\alpha_{\text{ref}} \rightarrow \bar{\alpha}$

until $\|\bar{\alpha} - \alpha_{\text{ref}}\|_\infty$ is less than 1e-3

Performance insights

- Perplexity over every domain ⬇️
- Weights ~ downstream-tuned
  - "The results obtained on The Pile reproduce the observations recently made by the RedPjamas & RefinedWeb datasets: some components of The Pile should ideally be downsampled, and increased web data may be beneficial." https://openreview.net/forum?id=lXuByUeHhd&noteId=vHaLbQbUBw
- Proxy model underperforms main model
  - Use main model instead of proxy model, even if the sizes are the same!
- Choose a relatively small proxy model size (280M) to save compute

The Llama 3 Herd of Models

Post-training Data

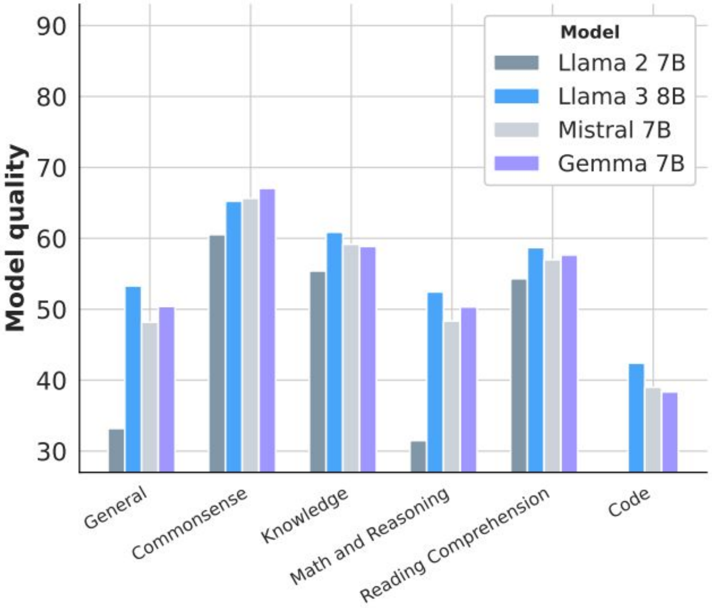
- Preference Data
    - multiple models: sample two responses from two different models for each user prompt
    - rate the strength of preference: significantly better, better, slightly better, or marginally better
    - edit the chosen response directly or prompt the model with feedback to refine its own response

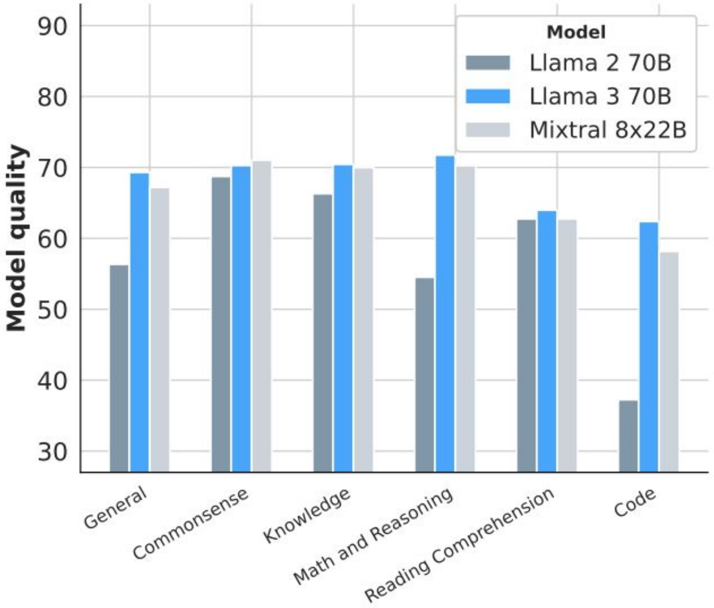
Post-training Data

- SFT Data
  - Rejection sampling: sample K (10~30) outputs from the latest chat model policy for each prompt, then select the best candidate
  - Synthetic data targeting specific capabilities
  - Small amounts of human-curated data

Evaluations

- Performance of pre-trained Llama 3 8B and 70B models on pre-training benchmarks





OpenAI o1 - Evaluations

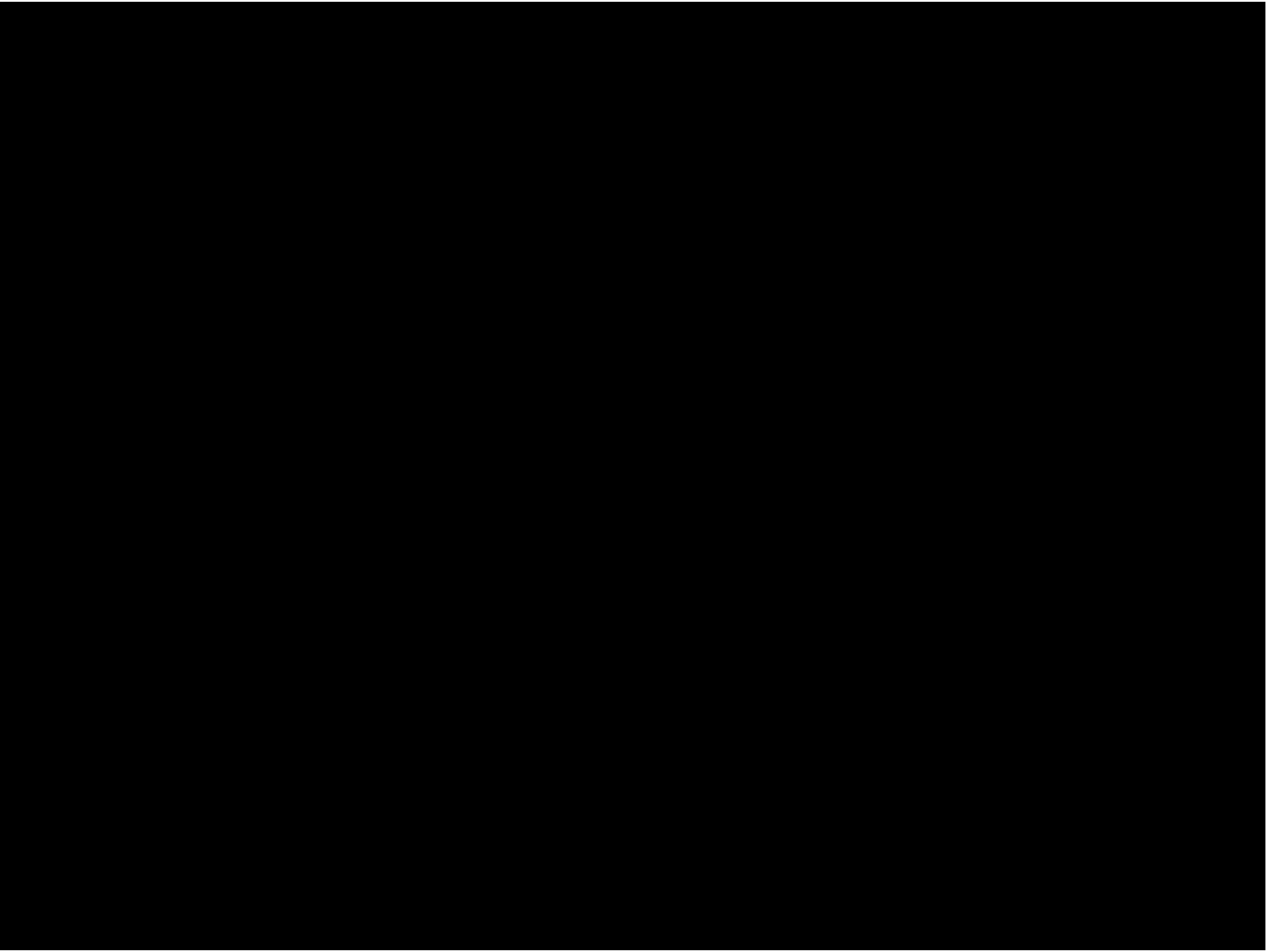
- Results for the disallowed content evaluations on GPT-4o, o1-preview, and o1-mini

Dataset	Metric	GPT-4o	o1-preview	o1-mini
Standard Refusal Evaluation	not_unsafe	0.99	0.995	0.99
	not_overrefuse	0.91	0.93	0.90
Challenging Refusal Evaluation	not_unsafe	0.713	0.934	0.932
WildChat [16]	not_unsafe	0.945	0.971	0.957
XSTest [17]	not_overrefuse	0.924	0.976	0.948

Chatbot Arena: An Open Platform for Evaluating LLMs by Human Preference

Live questions. Judged by humans.

Evaluation Metric \ Question Source	Static	Live
Ground Truth	MMLU, HellaSwag, GSM-8K	Codeforces Weekly Contests
Human Preference	MT-Bench, AlpacaEval	**Chatbot Arena**



Ask and vote!

Website: lmarena.ai

Arena | NEW: Overview | Arena (Vision) | Arena-Hard-Auto | Full Leaderboard

Total #models: **149**. Total #votes: **1,951,660**. Last updated: 2024-09-26.

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 Deprecate

Overall Questions

#models: 149 (100%) #votes: 1,951,660 (100%)

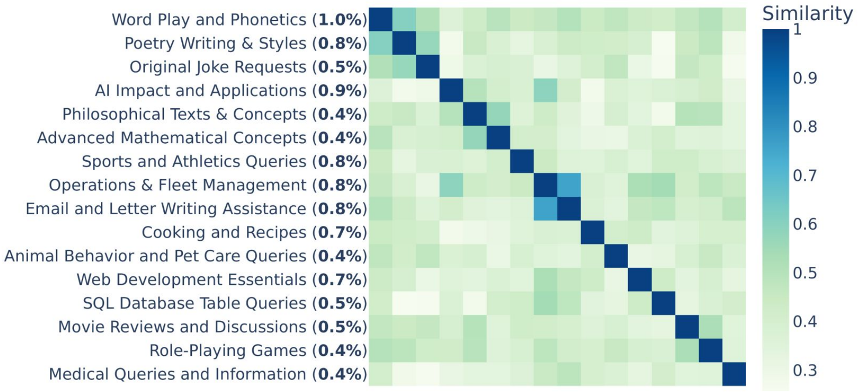
Rank* (UB)	Model	Arena Score	95% CI	Votes	Organization	License	Knowledge Cutoff
1	o1-preview	1339	+6/-7	9169	OpenAI	Proprietary	2023/10
1	ChatGPT-4o-latest (2024-09-03)	1337	+4/-4	16685	OpenAI	Proprietary	2023/10
3	o1-mini	1314	+6/-5	9136	OpenAI	Proprietary	2023/10
4	Gemini-1.5-Pro-Exp-0827	1299	+4/-3	31928	Google	Proprietary	2023/11
4	Grok-2-08-13	1293	+4/-3	27731	xAI	Proprietary	2024/3
6	GPT-4o-2024-05-13	1285	+3/-3	93428	OpenAI	Proprietary	2023/10
7	GPT-4o-mini-2024-07-18	1272	+3/-3	33166	OpenAI	Proprietary	2023/10
7	Claude 3.5 Sonnet	1269	+3/-3	67165	Anthropic	Proprietary	2024/4
7	Gemini-1.5-Flash-Exp-0827	1269	+3/-4	25027	Google	Proprietary	2023/11
7	Grok-2-Mini-08-13	1268	+4/-4	24956	xAI	Proprietary	2024/3
7	Gemini Advanced App (2024-05-14)	1266	+3/-3	52218	Google	Proprietary	Online
7	Meta-Llama-3.1-405b-Instruct-bf16	1266	+6/-7	8787	Meta	Llama 3.1 Community	2023/12
7	Meta-Llama-3.1-405b-Instruct-fp8	1266	+4/-4	33654	Meta	Llama 3.1 Community	2023/12
8	GPT-4o-2024-08-06	1264	+4/-3	25215	OpenAI	Proprietary	2023/10
10	Qwen2.5-72b-Instruct	1257	+8/-7	6017	Alibaba	Qwen	2024/9

The leaderboard

- Pairwise comparisons ➡️ BT scores & rankings
- Active sampling – Which model pair to choose for this round?

Behind the scenes…

- Detecting anomalous users – This user's ratings ↔ historical distribution
- Topic modeling – UMAP ➡️ HDBSCAN ➡️ GPT-4
- Quality validation
  - Prompts – Challenging enough; GPT-4 can also make the judge's job
  - Voting – Agreement with experts' choice



Discussion

- Other possible ways to improve pre-training data quality?
- Is there a better way to get high and low quality datasets?
- What are the problems of Chatbot Arena?