Data

Alexandra Birch

School of informatics

Based on slides by Pasquale Minervini

Outline

- Overview
- Pre-training Data Sources
- Pre-training Data Selection
- Post-training Data
- Data Ethics and Law
- Summary

Readings: Albalak et al. (2024) A Survey on Data Selection for Language Models, TML. (Section 2 and 3)
Penedo et al. (2024) The FineWeb Datasets: Decanting the Web for the Finest Text Data at Scale, NeurIPS

Data is the key

Data determines capabilities

- Model learns only what's in the training data (knowledge, reasoning patterns, biases)

Data quality requires human effort

- Most resource intensive and valuable component of modern AI
- Requires sustained investment into human expertise and careful curation
- Long tail of problems: domains, edge cases, quality issues

Competitive Differentiator

- Proprietary datasets (GPT-4, Claude) provide performance edge
- Architectural innovations shared in papers, system engineering can be reverse-engineered – datasets give sustained advantage
- Reasons for secrecy: (i) competitive dynamics and (ii) copyright liability

Types of Data

Pre-training data – on raw text
Annealing or mid-training – more focus on particular capabilities, higher quality data
Post-training – focus on evaluation tasks, desired behaviour
consist of: instruction training, RL, alignment, task fine-tuning

Terminology:

Base model: after pre-training + annealing
Instruct model: after post-training

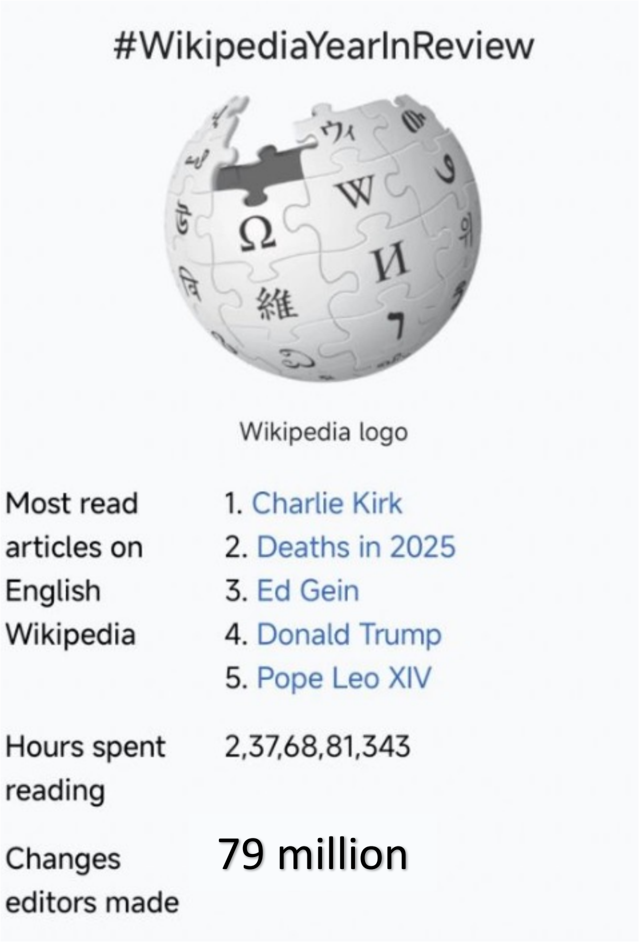
Each task has own data: translation, long context, reasoning etc. which we will look at in future lectures

Pre-training Data Sources

Wikipedia

- High quality, structured data
- Large community of **15 million writers**
  - 13 editors: more than 1 million edits!
- Sophisticated mechanism for quality standards even across very controversial pages
- Multilingual coverage – over 300 languages
- Free and accessible – creative commons licence
- Rich Metadata
- Limitations: geographic and demographic biases – only 8-15% are women

Used in training **almost all LLMs** alongside other larger, more diverse corpora

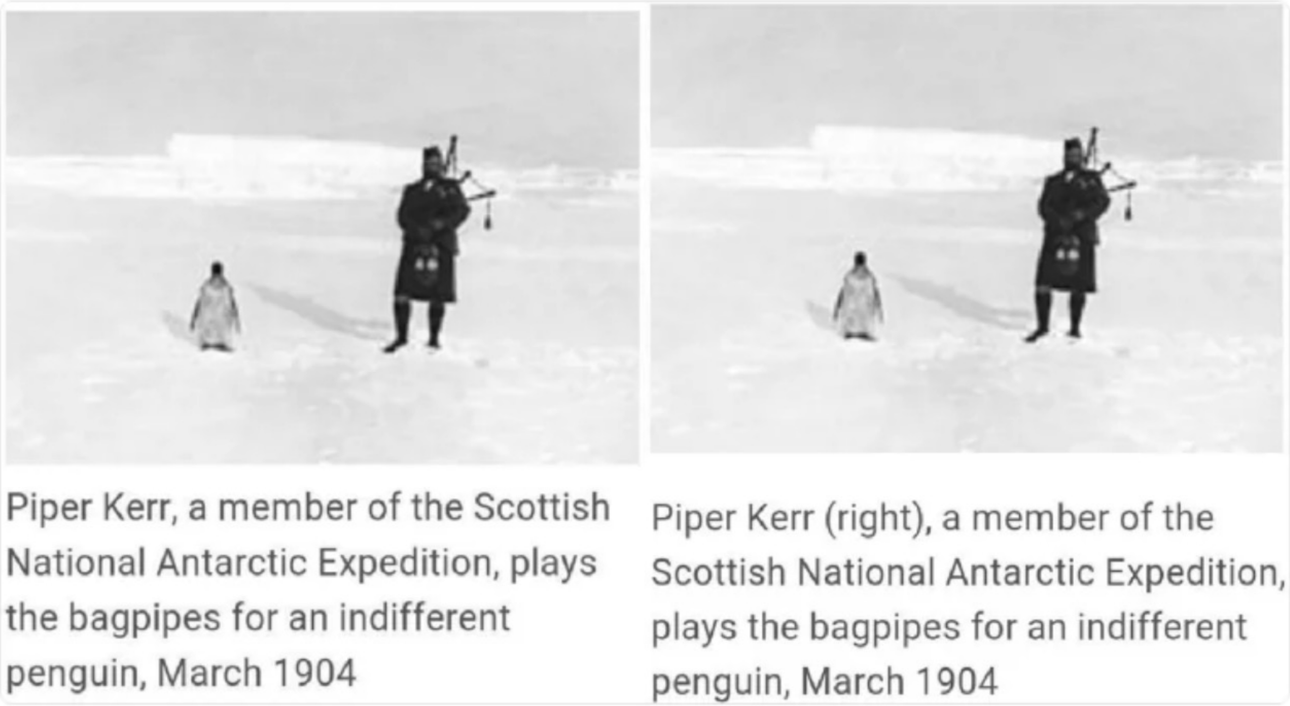


#WikipediaYearInReview

Wikipedia logo

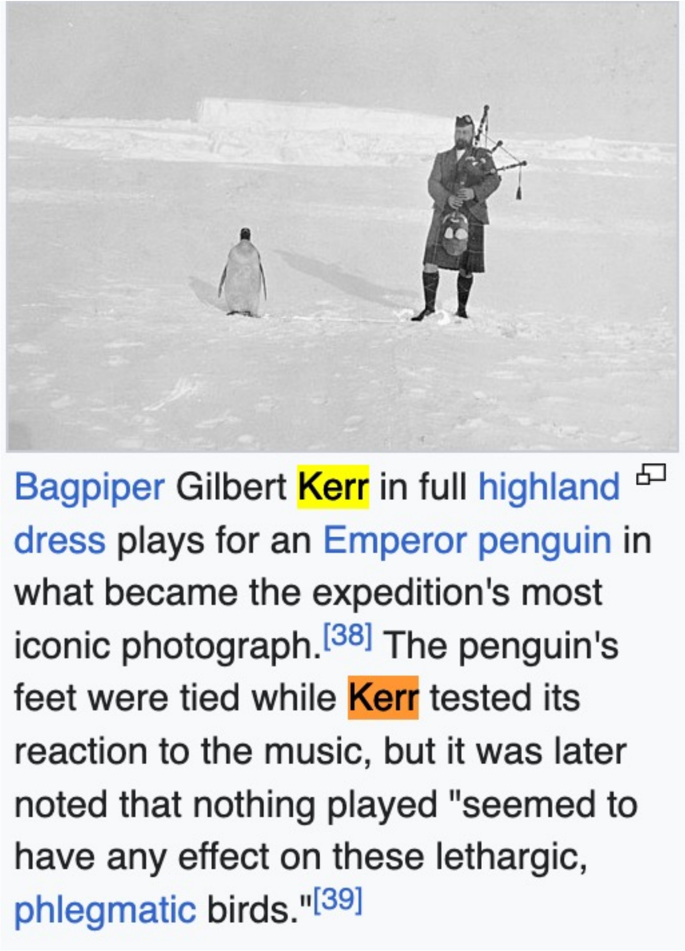
Most read articles on English Wikipedia	1. Charlie Kirk 2. Deaths in 2025 3. Ed Gein 4. Donald Trump 5. Pope Leo XIV
Hours spent reading	2,37,68,81,343
Changes editors made	79 million

Wikipedia



Piper Kerr, a member of the Scottish National Antarctic Expedition, plays the bagpipes for an indifferent penguin, March 1904

Piper Kerr (right), a member of the Scottish National Antarctic Expedition, plays the bagpipes for an indifferent penguin, March 1904



Bagpiper Gilbert Kerr in full highland dress plays for an Emperor penguin in what became the expedition's most iconic photograph.[38] The penguin's feet were tied while Kerr tested its reaction to the music, but it was later noted that nothing played "seemed to have any effect on these lethargic, phlegmatic birds."[39]

Web Data

- The biggest and most freely available data source is the web
- How big is it? **HUGE**

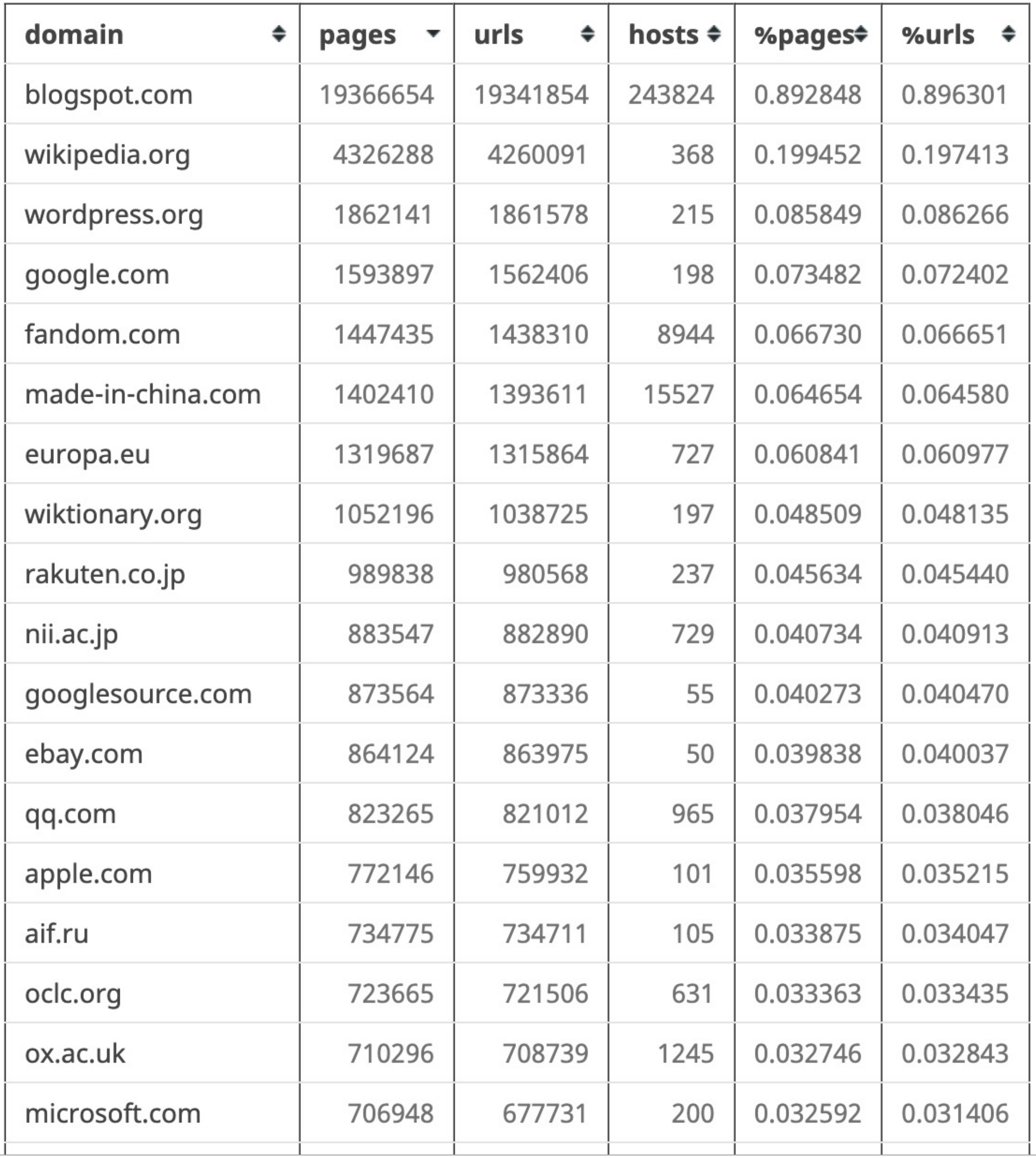
  Google search indexes about **400 billion web pages**
- Actual web far larger in the deep web – can be just institutional data that is behind a login eg. Amazon, but also dark web with deliberately encrypted data
- Spans a broad range of **domains, genres, languages** – world knowledge and linguistic knowledge that LLMs need
- Broad but not universal – overrepresents younger users and developed countries especially USA

https://zyppy.com/seo/google-index-size/

Common Crawl

- Non-profit organization started in 2007
- **Mission:** To provide open and free access to web data, a resource that was once primarily available only to large corporations like Google.
- **Data Size:** The corpus contains petabytes of data, with **monthly** crawls adding billions of new pages.
- **Accessibility:** The data is hosted on Amazon Web Services and is freely available for anyone to use and analyze.
- **Coverage strategy**: Crawls contain some overlap but deliberately diversify content capture
- **Scale of crawls:** in 2016 crawling took 12 days on a cluster of 100 machines

From groups.google.com



domain	pages	urls	hosts	%pages	%urls
blogspot.com	19366654	19341854	243824	0.892848	0.896301
wikipedia.org	4326288	4260091	368	0.199452	0.197413
wordpress.org	1862141	1861578	215	0.085849	0.086266
google.com	1593897	1562406	198	0.073482	0.072402
fandom.com	1447435	1438310	8944	0.066730	0.066651
made-in-china.com	1402410	1393611	15527	0.064654	0.064580
europa.eu	1319687	1315864	727	0.060841	0.060977
wiktionary.org	1052196	1038725	197	0.048509	0.048135
rakuten.co.jp	989838	980568	237	0.045634	0.045440
nii.ac.jp	883547	882890	729	0.040734	0.040913
googlesource.com	873564	873336	55	0.040273	0.040470
ebay.com	864124	863975	50	0.039838	0.040037
qq.com	823265	821012	965	0.037954	0.038046
apple.com	772146	759932	101	0.035598	0.035215
aif.ru	734775	734711	105	0.033875	0.034047
oclc.org	723665	721506	631	0.033363	0.033435
ox.ac.uk	710296	708739	1245	0.032746	0.032843
microsoft.com	706948	677731	200	0.032592	0.031406

Top registered domains in terms of page captures of last monthly crawl:
CC-MAIN-2025-51

From commoncrawl.github.io

GitHub

- Code is essential for programming tasks, but also helps develop reasoning behaviour
- GitHub started in 2008, acquired by Microsoft in 2018
- It contains a massive number of repositories: 630M projects
- It has valuable metadata: stars, project maturity and activity, issues and pull requests
- License information - enables filtering for permissively-licensed code only

The Stack dataset **31TB of data** across **30 programming languages** of permissively licensed code

- A tool called "Am I in The Stack" for developers to search for copies of their code
- MIT, Apache, BSD

Kocetkov et al. (2023) The stack: 3 TB of permissively licensed source code

Pre-training Data Selection

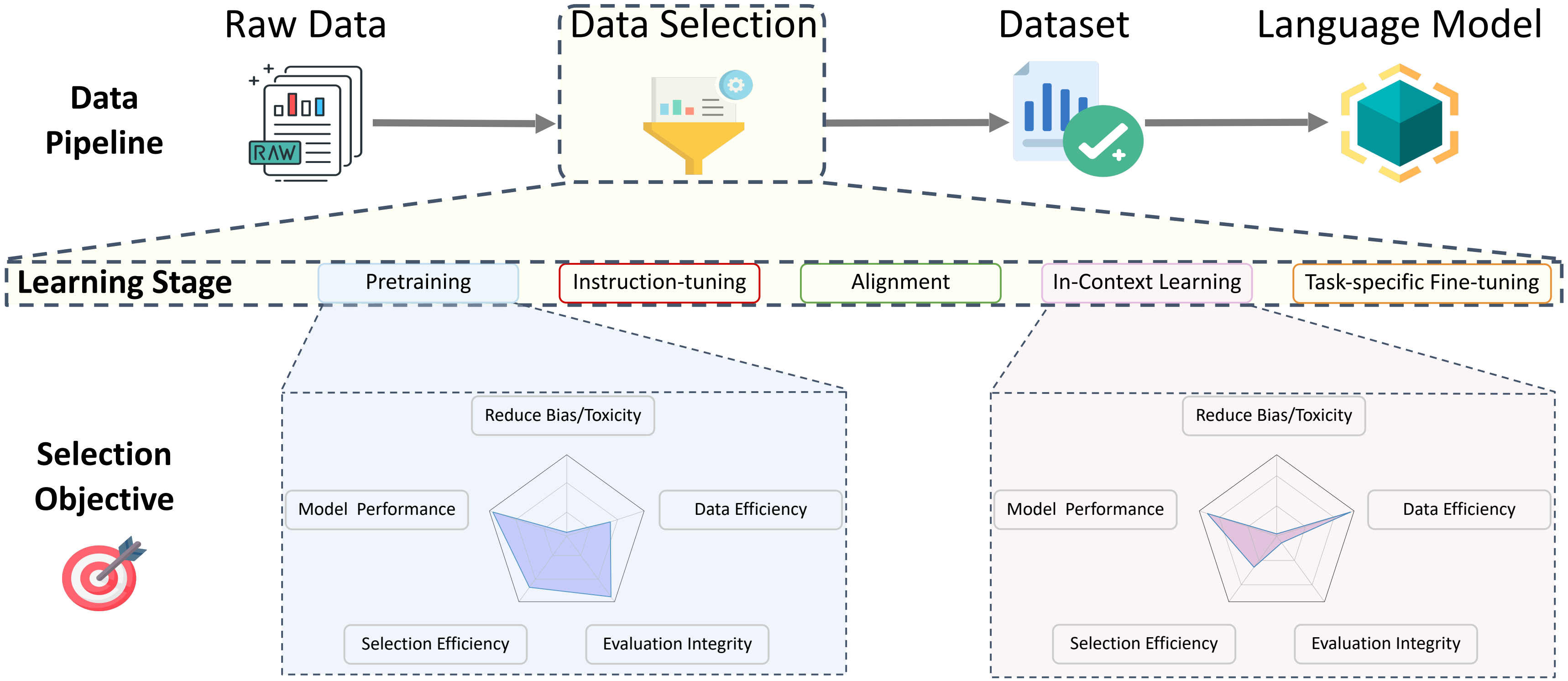
Data Selection

Machine learning models are a method for modeling statistical patterns in data

Goal of data selection:

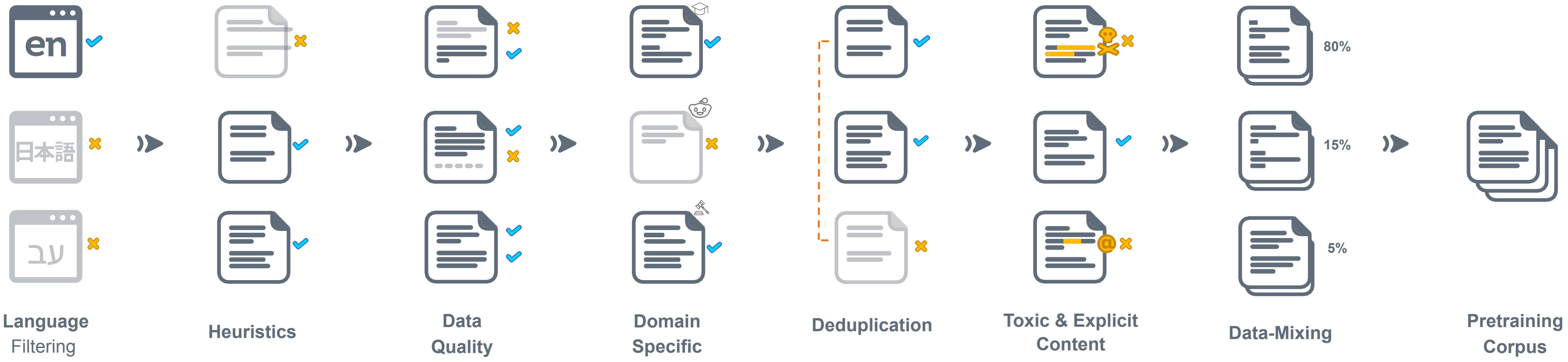
- Optimal Dataset: one that matches as closely as possible the distribution under which it is tested
- But also reduce costs
- Ensure integrity of evaluation metrics
- Reduce undesirable behaviours

Data Selection Criteria



[Albalak et al. (2024) A Survey on Data Selection for Language Models,](https://) TML

Pipeline



Albalak et al. (2024) A Survey on Data Selection for Language Models, TML

Language Filtering

- Why not go **multilingual**?
  - More difficult and expensive to curate quality data in many languages
  - In practice most LLMs are heavily multilingual – use existing parallel datasets
  - Still want to control amount of each language in the mix and focus on English

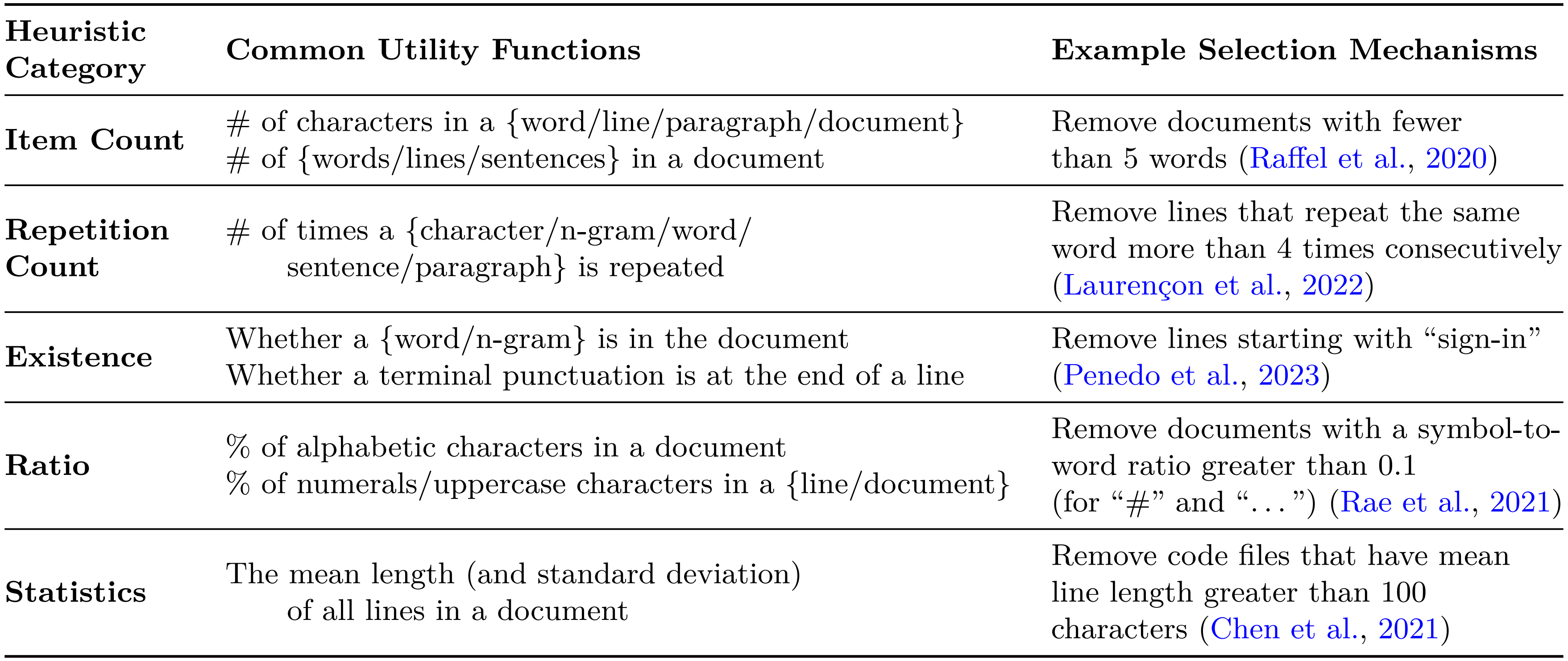
- For some lower-resourced languages, **URL-based methods** are useful eg. .fr or .es – not so useful for very low resource eg. Uyghur
- **Classifier based methods: fastText** is the current standard language identifier:
  - Pretrained to identify 157 languages
  - Fast implementation in C++ process 1,000 documents per second on a single CPU
  - Trained on multilingual sites eg. Wikipedia, Tatoeba

Challenges Remain:

- short sequences
- low-resource languages
- Setting thresholds: accidentally filtering out dialects of English
- Hard for similar languages (Bulgarian and Croatian)
- Ill-defined for code-switching (e.g., Spanish + English)

Joulin et al. (2016) Bag of Tricks for Efficient Text Classification

Heuristic Filtering



Heuristic Category	Common Utility Functions	Example Selection Mechanisms
Item Count	# of characters in a {word/line/paragraph/document} # of {words/lines/sentences} in a document	Remove documents with fewer than 5 words (Raffel et al., 2020)
Repetition Count	# of times a {character/n-gram/word/ sentence/paragraph} is repeated	Remove lines that repeat the same word more than 4 times consecutively (Laurençon et al., 2022)
Existence	Whether a {word/n-gram} is in the document Whether a terminal punctuation is at the end of a line	Remove lines starting with "sign-in" (Penedo et al., 2023)
Ratio	% of alphabetic characters in a document % of numerals/uppercase characters in a {line/document}	Remove documents with a symbol-to-word ratio greater than 0.1 (for "#" and "...") (Rae et al., 2021)
Statistics	The mean length (and standard deviation) of all lines in a document	Remove code files that have mean line length greater than 100 characters (Chen et al., 2021)

Albalak et al. (2024) A Survey on Data Selection for Language Models, TML

Data Quality

- High quality is not defined
- Examples are: Wikipedia, books, patents, and peer-reviewed journal articles
- Quality filtering requires fuzzy matching which generally has higher computational requirements
- data quality often use relatively cheap distributional representation methods, such as n-grams, to allow for more ambiguous filtering criteria
- classifier-based quality filtering, where the goal is to identify data points that are likely from the same (or similar) distribution as a known “high-quality” corpus of data points (reference corpus)

Deduplication

Why Deduplicate Training Data?

- Improved Generalization
- Efficient Resource Utilization
- Fair Representation
- Reduced Overfitting

Methods:

- URLs, hashing, string metrics and model representations
- **Exact matching** eg. URL deduplication
- **Approximate matching** eg. MinHash – very good at finding templated text

The Core Problem

$n \times {(n-1)}/{2}$ or $O(n^2)$ where n = number of documents

Eg. n = 1 million means 500 billion comparisons

MinHash

1. Convert Documents to Sets

Break each document into tokens (words, character n-grams, etc.)

Example: "the cat sat" → {the cat, cat sat}

2. Create Hash Signatures

Apply multiple hash functions to each element in the set

For each hash function, keep only the *minimum* hash value (hence "MinHash")

This creates a compact "signature" (e.g., 128 numbers) representing the document

3. Estimate Similarity

Compare signatures instead of full documents

The fraction of matching values in the signatures approximates the **Jaccard similarity** (overlap between sets)

If 64 out of 128 signature values match, similarity = 50%

- Documents with similar content will likely produce similar minimum hash values.
- The more hash functions you use, the more accurate the similarity estimate.
- Works on billions of documents (used by Common Crawl, Google, etc.)

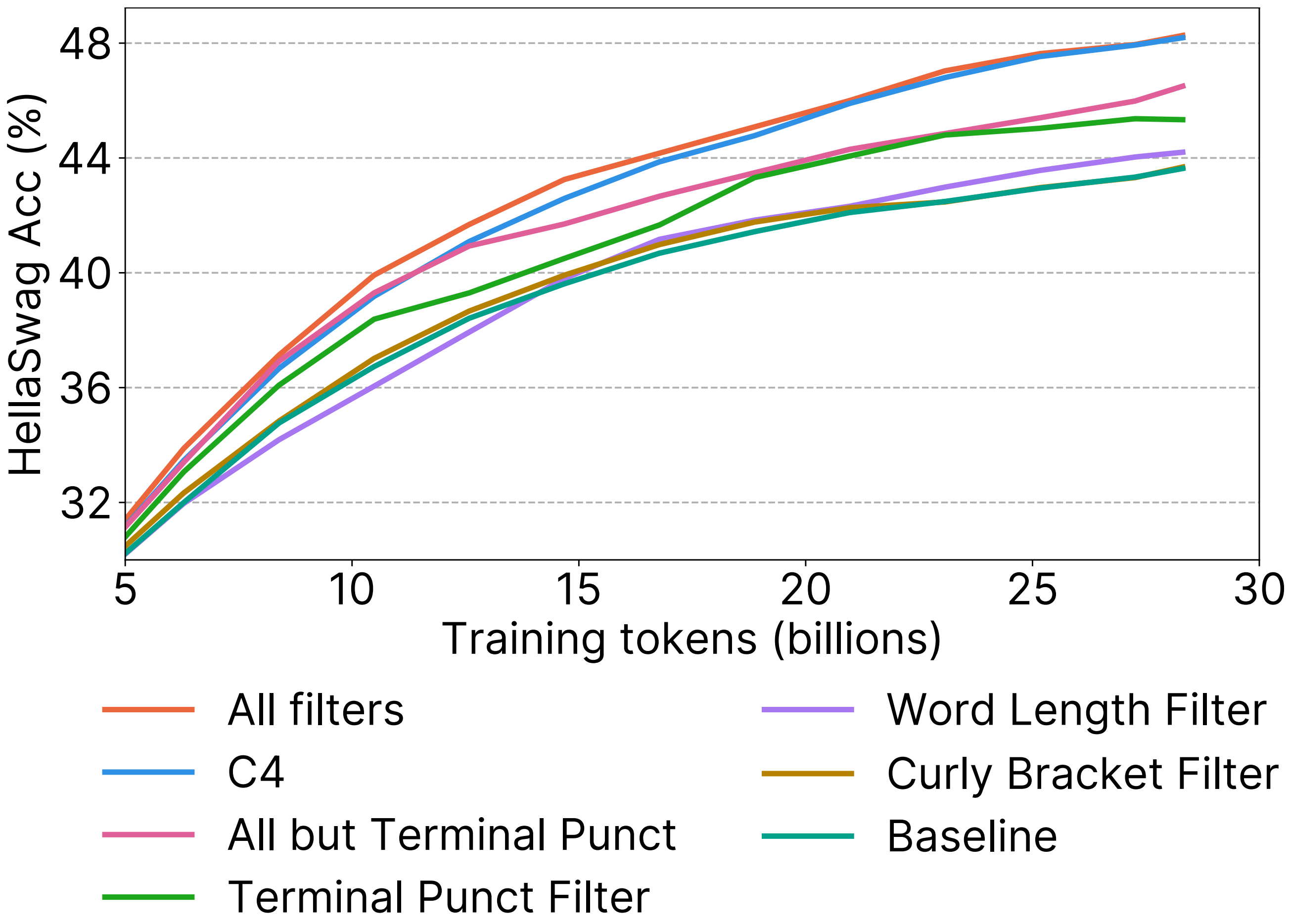
For details: https://medium.com/@omkarsoak

Put in Practice: FineWeb

- Quality trumps quantity!
- FineWeb, a 15T token dataset of text sourced from 96 Common Crawl snapshots
- Principled strategy for choosing and tuning filtering heuristics that helped produce a small set of effective filters out of over fifty candidate filters
- How different deduplication strategies and granularities can impact performance
- Data quality using educational classifier

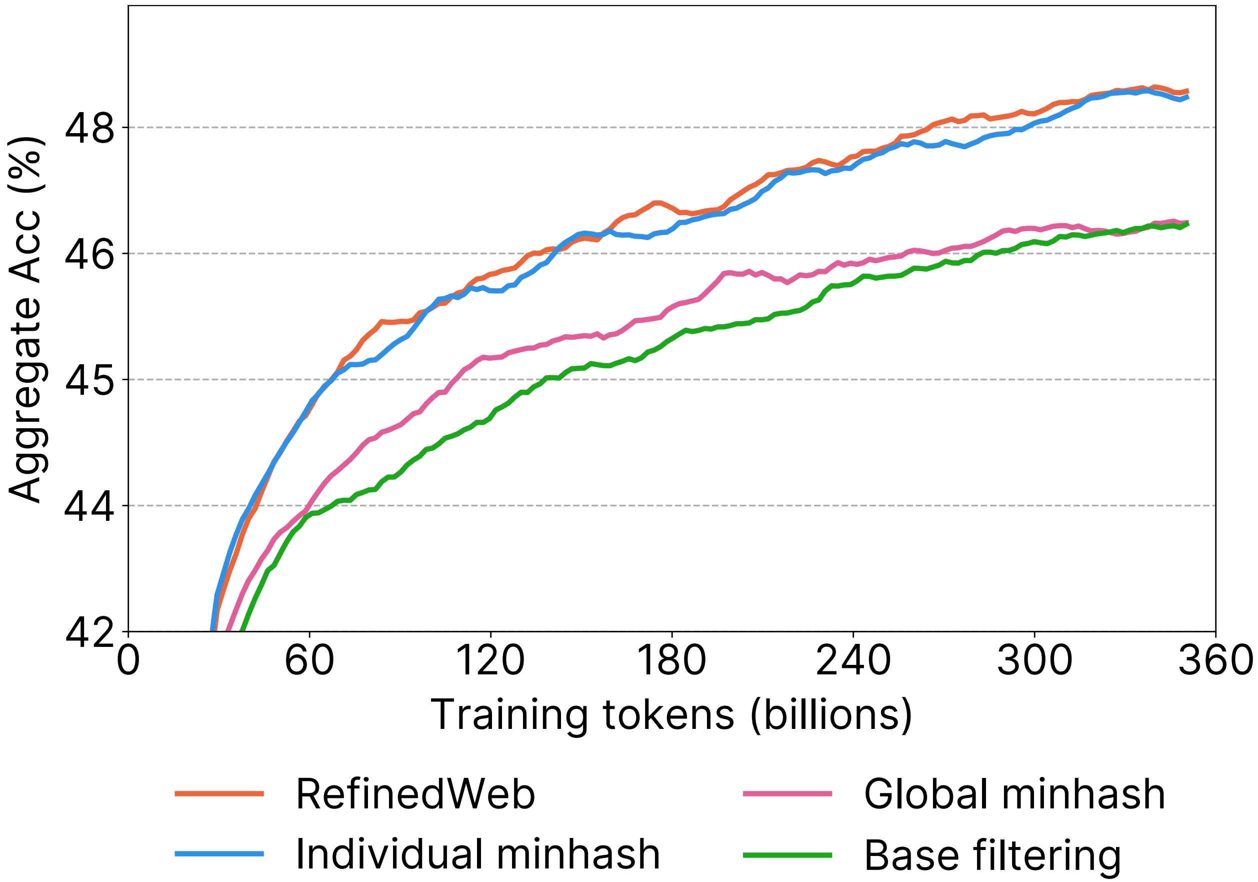
Penedo et al. (2024) The FineWeb Datasets: Decanting the Web for the Finest Text Data at Scale, NeurIPS

Heuristic Filtering



Penedo et al. (2024) The FineWeb Datasets: Decanting the Web for the Finest Text Data at Scale, NeurIPS

Deduplication



Global dedup: 4T tokens
Indiv. dedup: 20T tokens

[Penedo et al. (2024) The FineWeb Datasets: Decanting the Web for the Finest Text Data at Scale, NeurIPS]

FineWeb-Edu

Below is an extract from a web page. Evaluate whether the page has a high educational value ... Points are accumulated based on the satisfaction of each criterion:

- Add 1 point if the extract provides some basic information relevant to educational topics, …
- Add another point if the extract addresses certain elements pertinent to education …
- Award a third point if the extract is appropriate for educational use and introduces key concepts relevant to school curricula…
- 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…
- Bestow a fifth point if the extract is outstanding in its educational value, perfectly suited for teaching either at primary school or grade school…

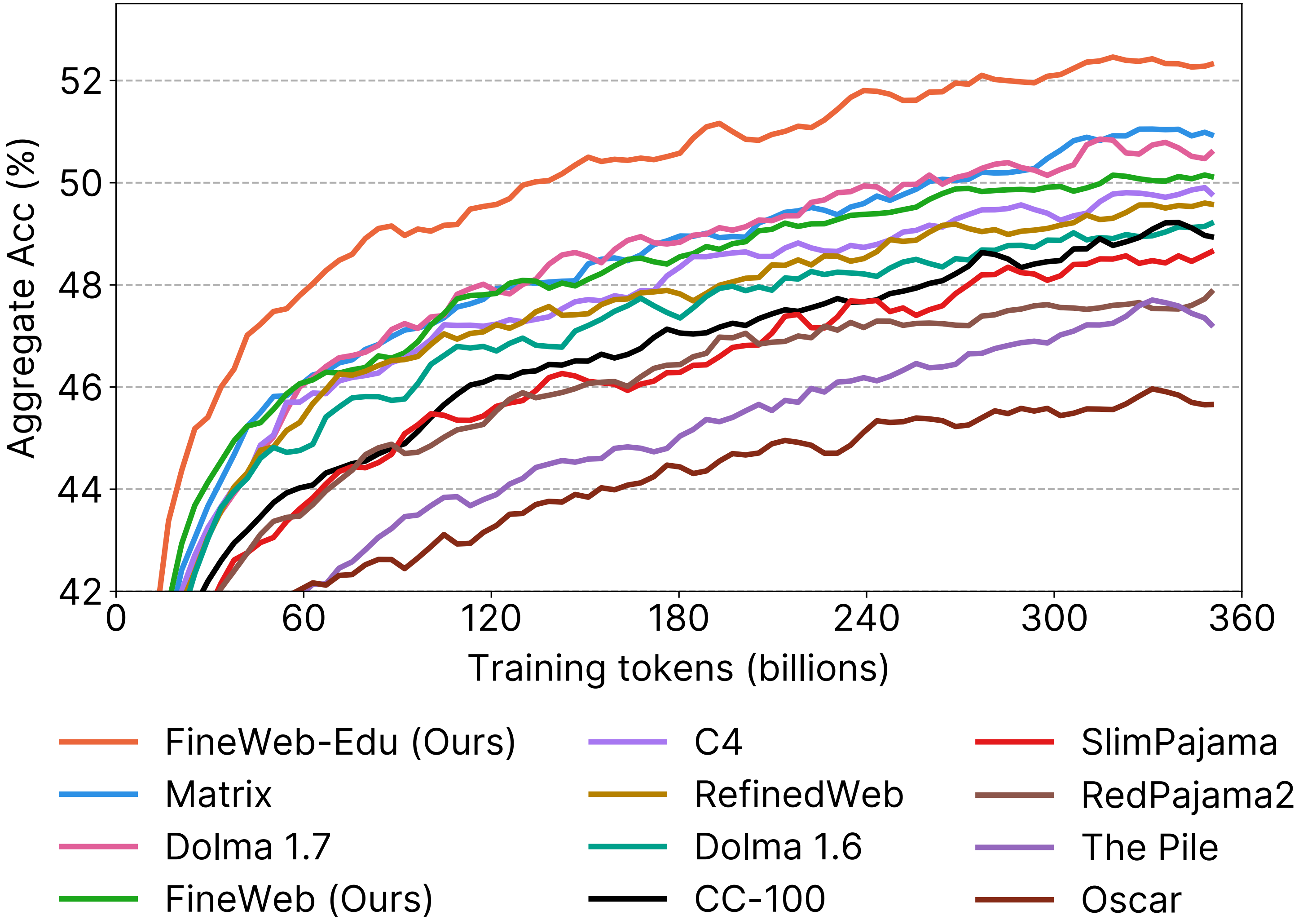
The extract: <EXAMPLE>.
After examining the extract:
- Briefly justify your total score, up to 100 words.
- Conclude with the score using the format: "Educational score: <total points>"

Used Llama-3-70B-Instruct to score 460,000 sample documents from FineWeb

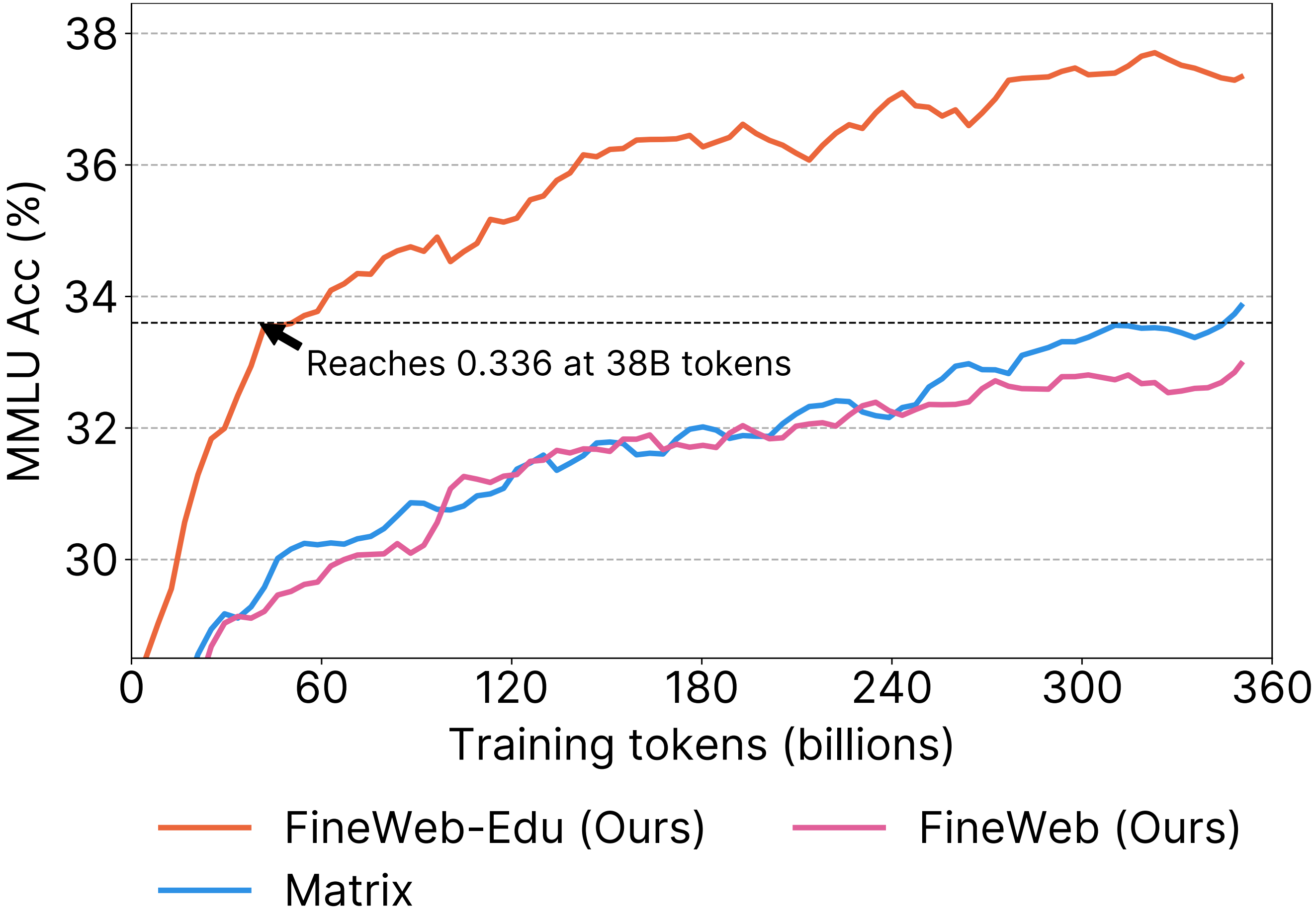
FineWeb-Edu

- Trained lightweight linear regression classifier on Snowflake-arctic-embed-m embeddings
- Enabled efficient filtering of 15 trillion tokens (vs. $1M+ cost using LLM directly)
- Applied minimum threshold of score ≥3 for inclusion in FineWeb-Edu
- Threshold chosen to optimize trade-off between reasoning benchmarks (MMLU, ARC) and general benchmarks (HellaSwag)
- Retained 1.3 trillion tokens of highest-quality educational content from 15 trillion token FineWeb corpus

FineWeb-Edu


Aggregate Acc (%)
52
50
48
46
44
42
0
60
120
180
240
300
360
Training tokens (billions)
FineWeb-Edu (Ours)
Matrix
Dolma 1.7
FineWeb (Ours)
C4
RefinedWeb
Dolma 1.6
CC-100
SlimPajama
RedPajama2
The Pile
Oscar

FineWeb-Edu


38
36
34
32
30
MMLU Acc (%)
0
60
120
180
240
300
360
Training tokens (billions)
Reaches 0.336 at 38B tokens
FineWeb-Edu (Ours)
FineWeb (Ours)
Matrix

Toxic and Personal Data

Toxic and Explict Content Removal:

Do not want model producing £$%^^&@£, but also means can’t identify this text

- Heuristic: lists of words to block in URLs or in text
- Classifier based methods: either trained or a commercial tool eg. Google’s Safe Search

Personally Identifiable Information Obfuscation:

Personal Names, addresses, telephone numbers, credit card numbers, emails etc.

- Want to remove or ideally replace with coherent synthetic information
- Want to do this for Wikipedia? No.

Post-Training Data

Instructions

Limitations of pre-trained models:
No explicit notion of *“what task am I being asked to do?”*
Does not learn how to follow goals, answer questions or behave helpfully

Instruction: Summarize the following text in one sentence.
Response: The text explains how photosynthesis converts sunlight into energy for plants.

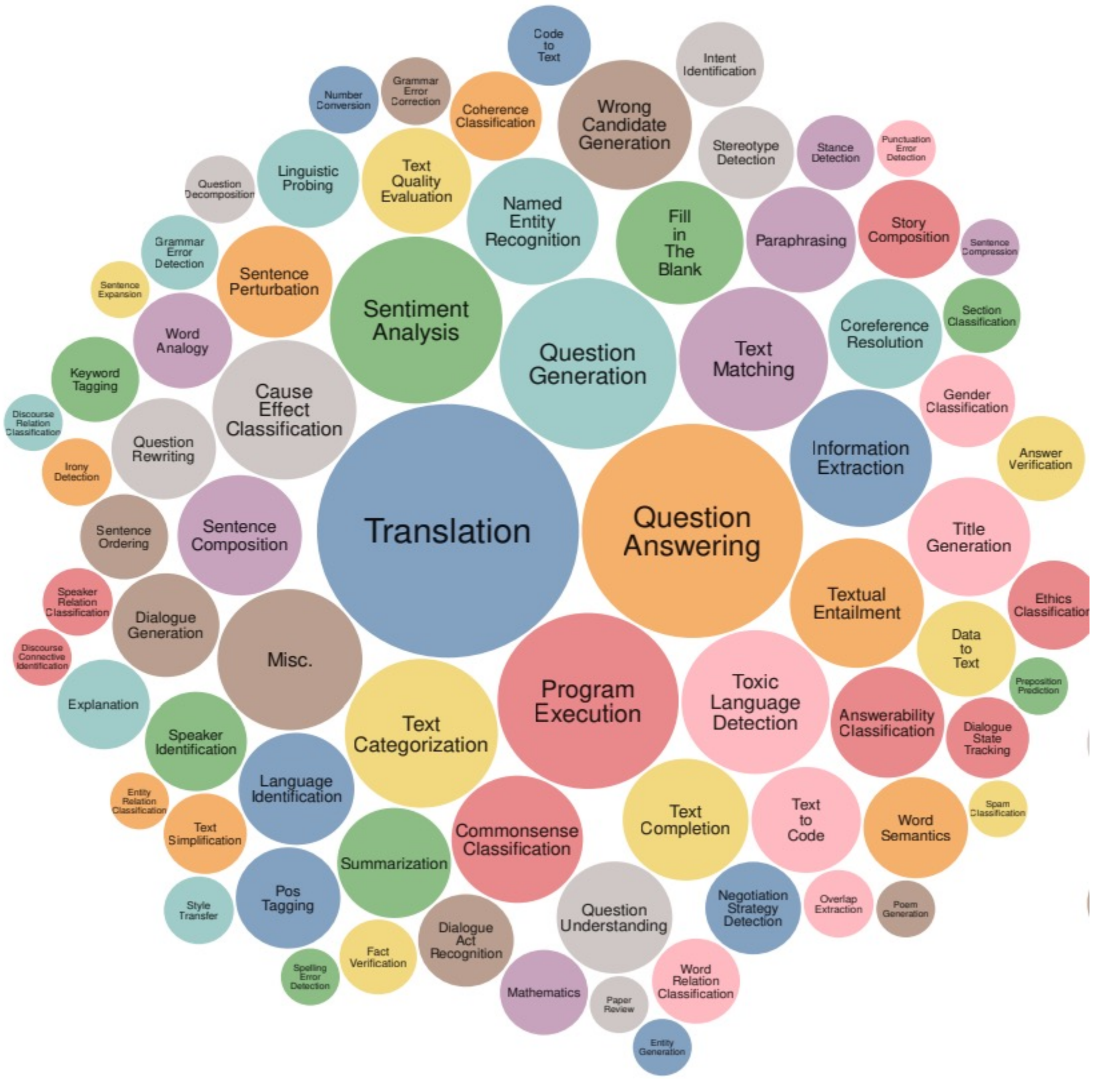
Instruction: Translate the sentence into French: “Good morning.”
Response: Bonjour.

Instruction: List two benefits of regular exercise.
Response: Regular exercise improves cardiovascular health and reduces stress.

What instruction tuning adds
A supervised signal for **following commands** – what kind of behaviour is expected
Exposure to **many task types** via shared instruction format turns word predictor into system can generalize
Alignment with how humans naturally interact with models, and how to be helpful, and truthful

Super Natural Instructions

- A large-scale instruction-tuning dataset covering **1,600+ diverse NLP tasks** - far more than previous work
- Tasks are written in **natural language**, resembling how humans give real instructions
- Enables strong **zero-shot and cross-task generalization**
- Key insight: *exposure to many well-written instructions improves a model's ability to handle unseen tasks*



Wang et al (2022) Super-Natural Instructions: Generalization via Declarative Instructions on 1600+ NLP Tasks

Synthetic Instruction Data

Main approaches to Synthetic Data:

- Self-generation + verification (STaR, math problems)
- Distillation from stronger models (Alpaca, Orca) ➡ see more in lecture 10 Distillation
- Evolutionary methods (Evol-Instruct, WizardLM)
- Procedural generation (code, structured tasks)

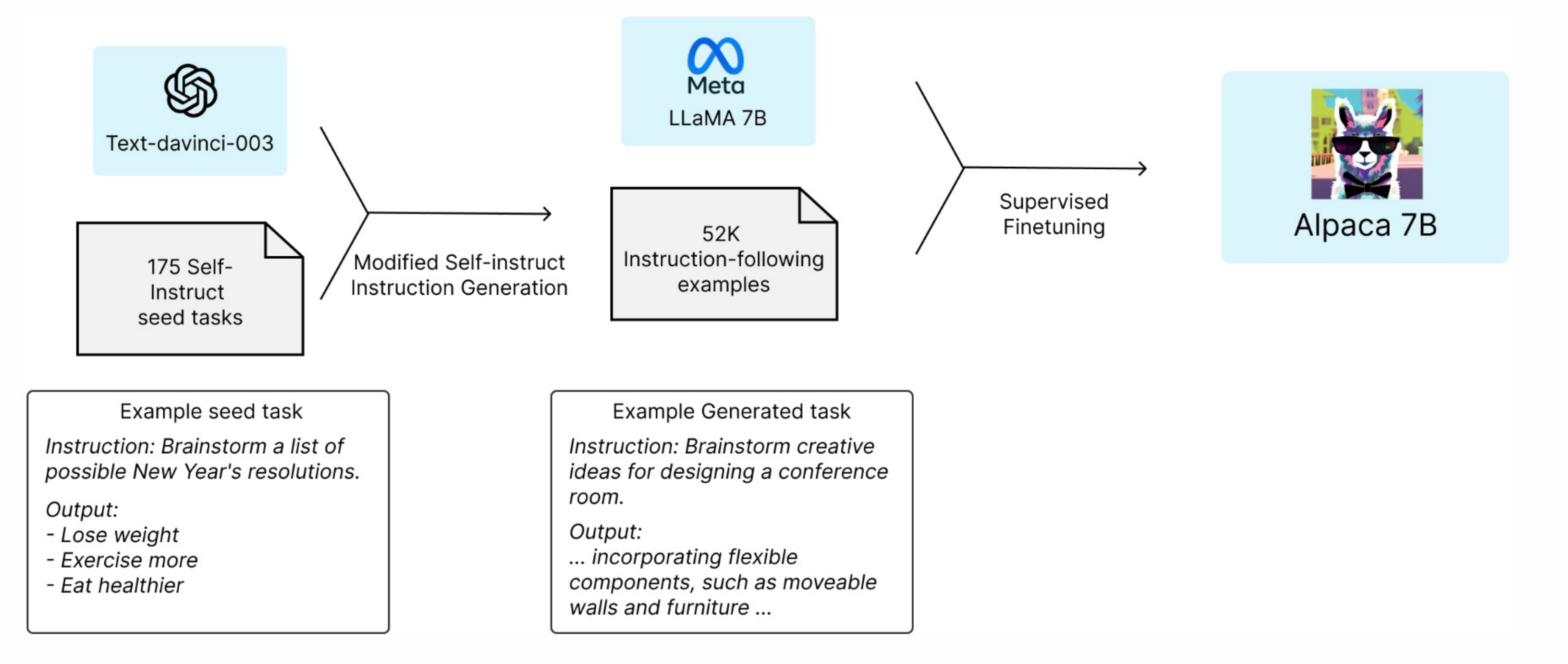
Synthetic Instruction Data

Challenges

- Quality control (how do you filter? Humans? LLM as judge?)
- Diversity maintenance (avoiding collapse – mix in human data)
- When is ground truth available? (verify correctness - eg math – makes synthetic data much more powerful)

Synthetic data is a core part of modern LLM development

Synthetic Instruction Data


Text-davinci-003
175 Self-Instruct seed tasks
Modified Self-instruct Instruction Generation
Meta
LLaMA 7B
52K Instruction-following examples
Supervised Finetuning
Alpaca 7B

Example seed task

Instruction: Brainstorm a list of possible New Year's resolutions.

Output:
- Lose weight
- Exercise more
- Eat healthier

Example Generated task

Instruction: Brainstorm creative ideas for designing a conference room.

Output:
... incorporating flexible components, such as moveable walls and furniture ...

- Alpaca: Synthetic instruction dataset inspired by Self-Instruct
- 52K instruction response pairs generated using a larger LLM cost less than $500 using the OpenAI API.
- Enabled instruction tuning of LLaMA with minimal compute cost
- Popularized open, reproducible instruction-following models
- Sparked widespread community adoption and follow-up datasets

Taori et al. (2023) Stanford Alpaca: An Instruction-following LLaMA model

Data Ethics and Law

Representation

- Training data shapes whose voices, cultures, and perspectives models learn to reflect
- Underrepresentation can lead to bias, stereotyping, and exclusion of marginalized groups
- Overrepresentation of dominant groups may reinforce existing power imbalances
- Cultural and linguistic gaps can reduce model fairness, accuracy, and usefulness

Example: African American dialect
Prompt: Is this correct English? “She been finished her homework.”
Response: No

Copyright

- Literary, dramatic, musical and artistic works must comply with the criterion of originality and be fixed in material form (eg. Written down) – not ideas or facts
- **Automatic Protection**: Copyright applies automatically in the UK without any registration required, unlike patents. The threshold for protection is extremely low—even simple websites are copyrighted by default.
- **Duration**: lasts for the life of the author plus 70 years, after which works enter the public domain (examples include Shakespeare's plays, works by authors who died before 1955, and most content on Project Gutenberg).
- Most internet content is copyrighted, even if not explicitly marked.

There are two legal pathways to using Copyrighted Material:

1. Obtain a license from the copyright holder
2. Invoke fair dealing (UK's narrower equivalent of US fair use, limited to specific purposes: research and private study, criticism and review, news reporting, quotation, parody/pastiche/caricature, text and data mining)

License

- A license is granted by a licensor to a licensee.
- The Creative Commons license enables free distribution of copyrighted work.
- Examples: Wikipedia, Khan Academy, Free Music Archive, 10 million videos from YouTube, etc.
- Terms of Use: terms of service might impose additional restrictions.

    Example: YouTube's terms of service prohibits downloading videos, even if the videos are licensed under Creative Commons.

Summary

- Data is key to success
- High quality data comes from curated crawls eg. Wikipedia
- Large pretraining datasets come from noisy crawls
- Data selection pipeline is critical: cleaning, quality, deduplication
- Synthetic data can help train classifiers or create instructions
- Ethics of data is very important to consider