

Data for LLM pretraining

Marco Kuhlmann

Department of Computer and Information Science

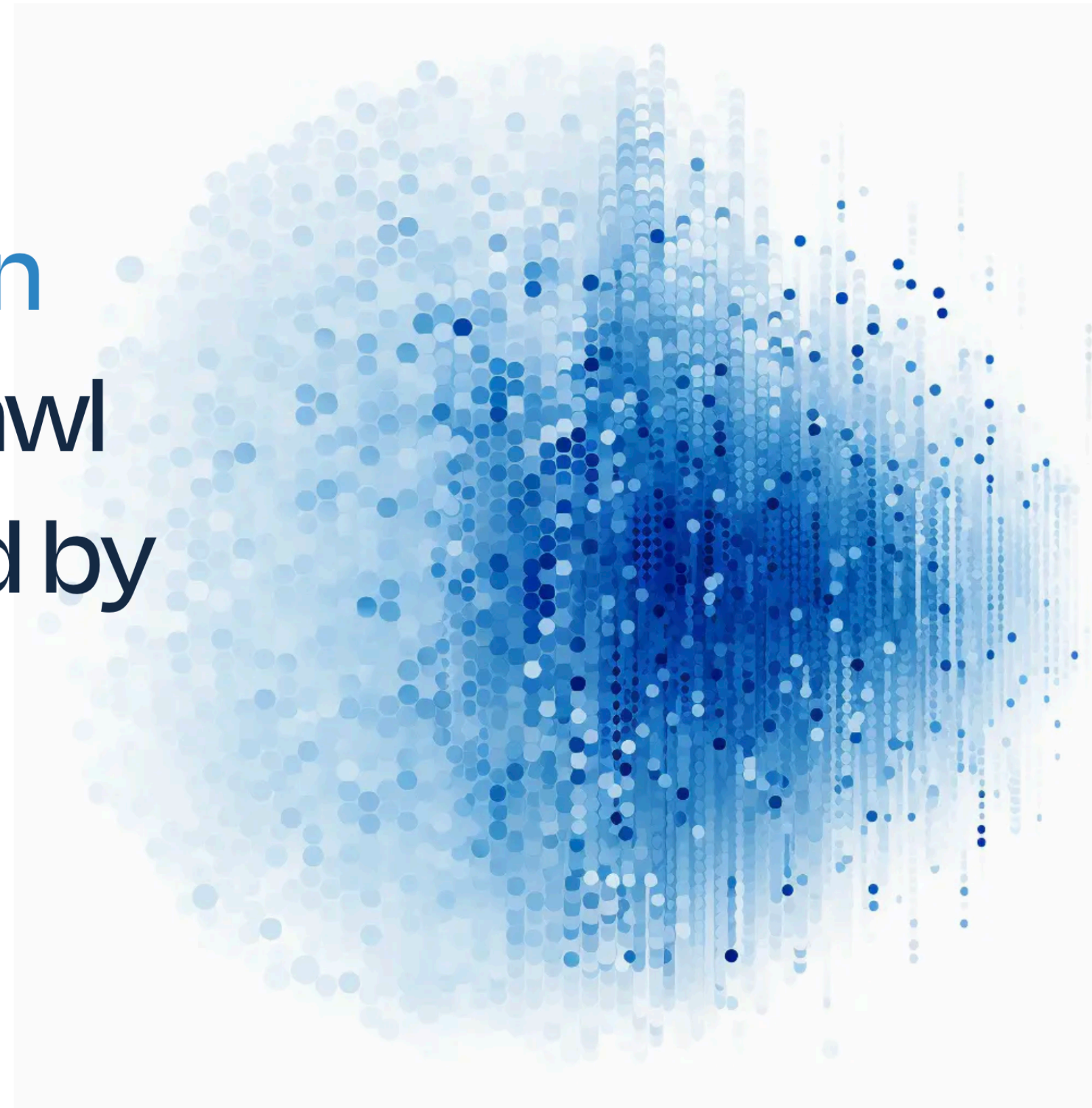
Data for LLM pretraining

- Training modern LLMs demands vast amounts of data. This data is often sourced from the Internet.
- While abundant, Internet data is unstructured, noisy, and biased, making it an imperfect representation of language.
- Internet text data requires extensive postprocessing and quality filtering to enhance relevance and diversity.

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

Common Crawl is a 501(c)(3) non-profit founded in 2007.

We make wholesale extraction, transformation and analysis of open web data accessible to researchers.

[Overview](#)

Common Crawl

Crawl ID (Year–Week)	Number of Pages	Total Size WARC	Total Size WET
2024-51	2.64 B	80.92 TiB	7.37 TiB
2023-50	3.35 B	99.25 TiB	9.30 TiB
2022-49	3.35 B	92.59 TiB	9.58 TiB
2021-49	2.50 B	68.66 TiB	7.18 TiB

Source: [Common Crawl](#)

F^one^uweb

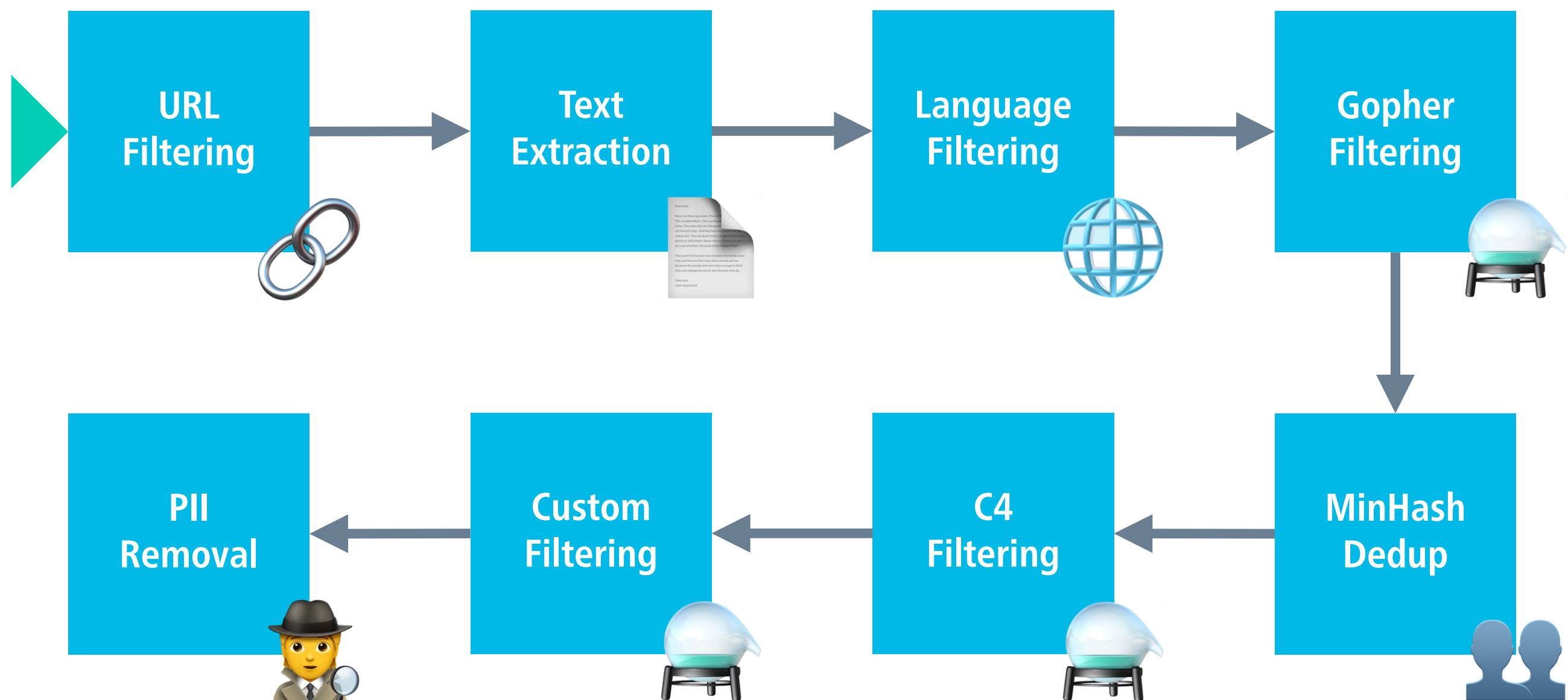
The finest collection of data the web has to offer



“15 trillion tokens of the finest data the web has to offer”

[Source](#)

The FineWeb pipeline



[Penedo et al. \(2024\)](#)

Basic filtering

- URL filtering using blocklists

Examples: adult content, malware, phishing sites

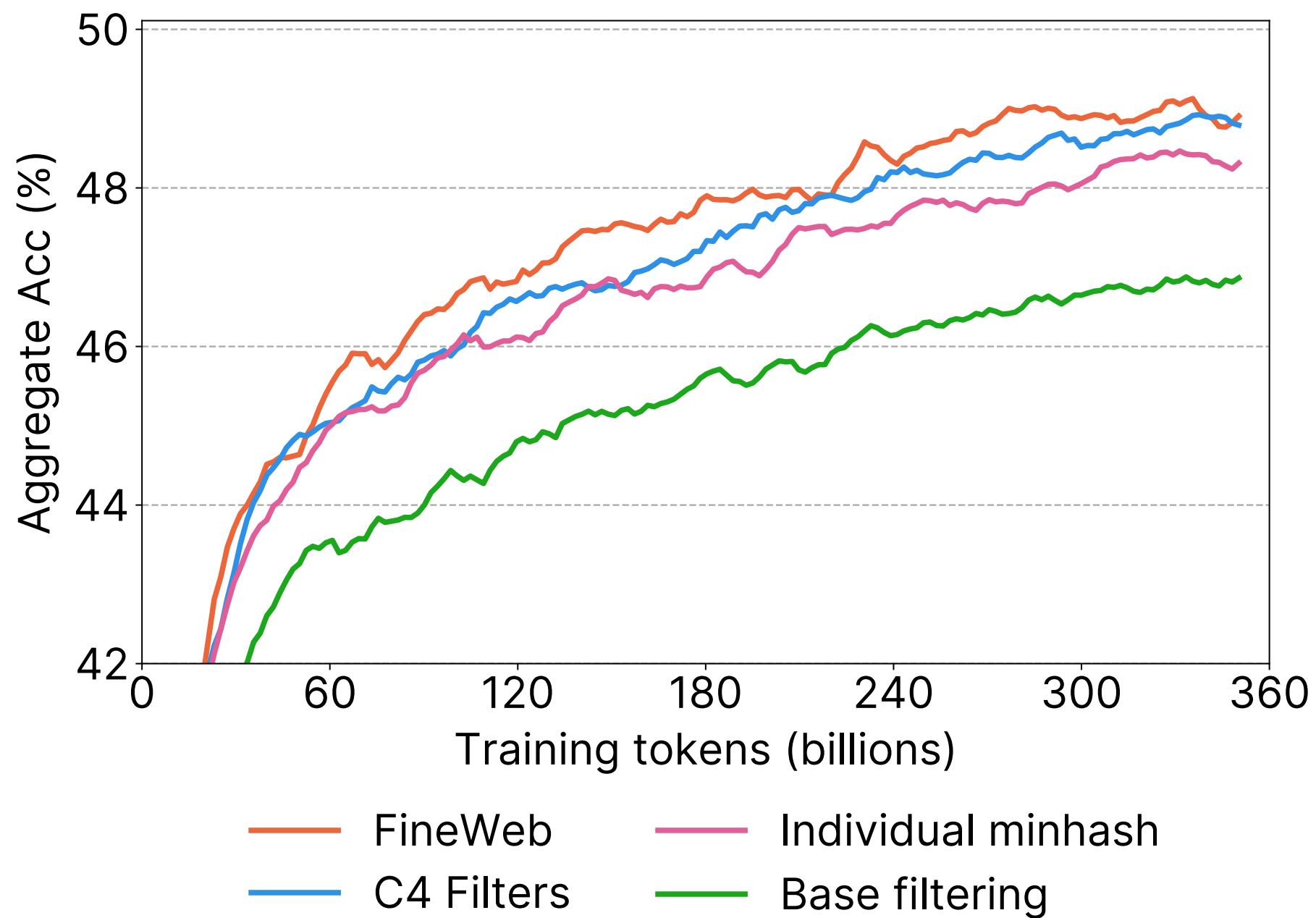
- Language filtering

Uses fastText; keep only English text with a score $\geq 65\%$

- Heuristic filtering

Filter for length, symbol-to-word ratio, common/uncommon words, etc.

Impact of filtering



Text extraction

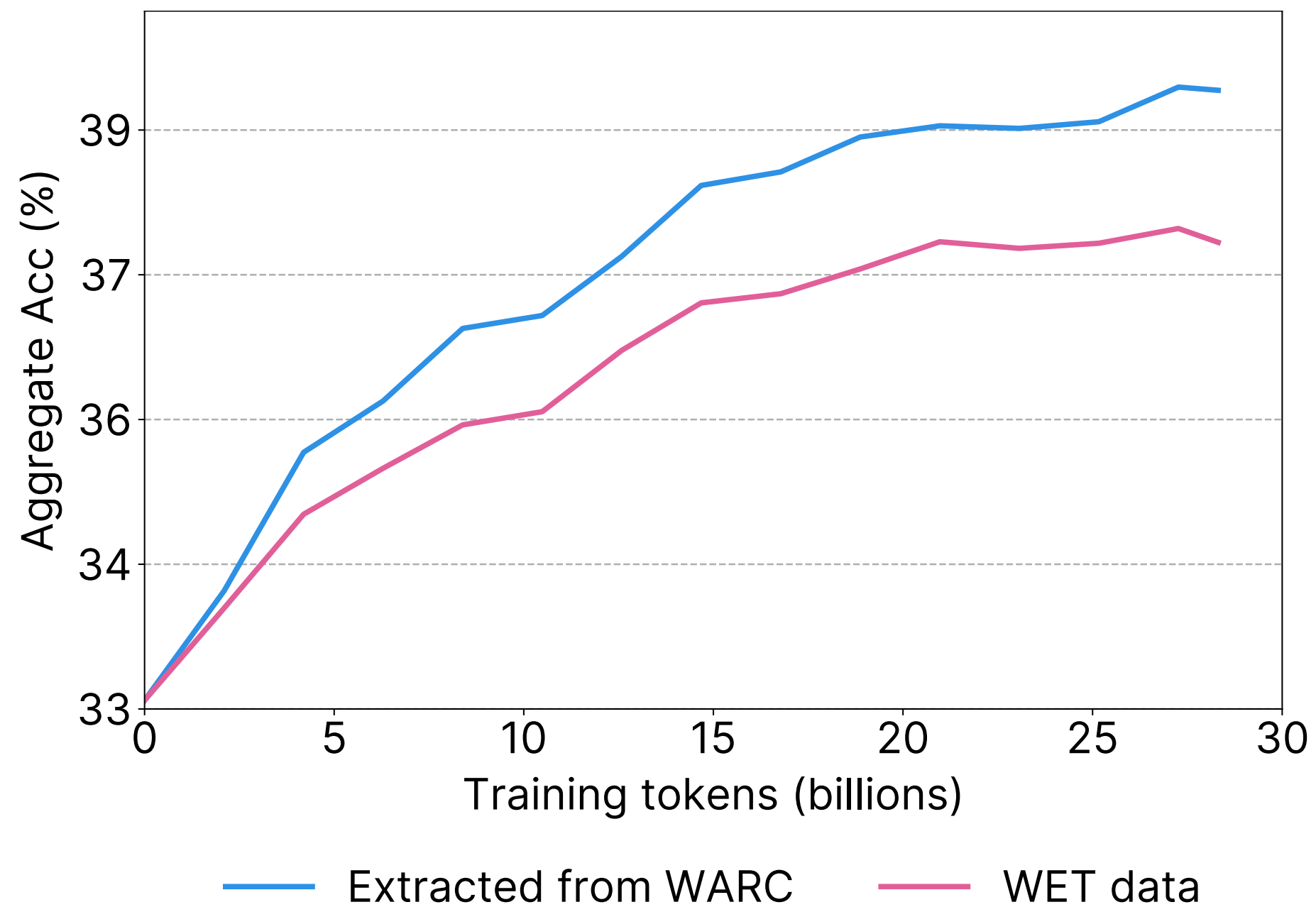
[illegible]

Linköping University (LiU; Swedish: Linköpings universitet) is a public research university based in Linköping, Sweden. Originally established in 1969, it was granted full university status in 1975 and is one of Sweden's largest academic institutions.[5]

Linköpings universitet | |
Type	Public research university

FineWeb uses Trafilatura.

Relevance of text extraction



Penedo et al. (2024)

Deduplication

- Large-scale web datasets contain significant amounts of duplicate content, which can lead to overfitting.
- Deduplication leads to a more diverse dataset and reduces computational cost.
- Deduplicating massive datasets requires efficient similarity detection techniques or other fuzzy approaches.

embedding-based similarity or MinHash

Name	Based on	Release year	Number of tokens
<u>C4</u>	Common Crawl	2019	156B
WebText	Own Crawl (OpenAI)	2019	300B
<u>CC-100</u>	Common Crawl	2020	532B
MassiveText	Own Crawl (Google)	2022	2.3T
<u>OSCAR</u>	Common Crawl	2023	523B
<u>RedPajama</u>	Common Crawl	2023	30.4T
<u>RefinedWeb</u>	Common Crawl	2023	500B
<u>Dolma</u>	Common Crawl	2024	3T
<u>FineWeb</u>	Common Crawl	2024	15T

FineWeb and FineWeb-EDU

- Design choices were validated through training data ablation studies and evaluated on downstream task benchmarks.
- The authors released a 1.3T-token filtered subset of FineWeb, focusing on high-quality educational web pages.
- FineWeb-EDU was built using an educational quality classifier, trained on labels generated by Llama (1.71B).

linear regression on top of an embedding model

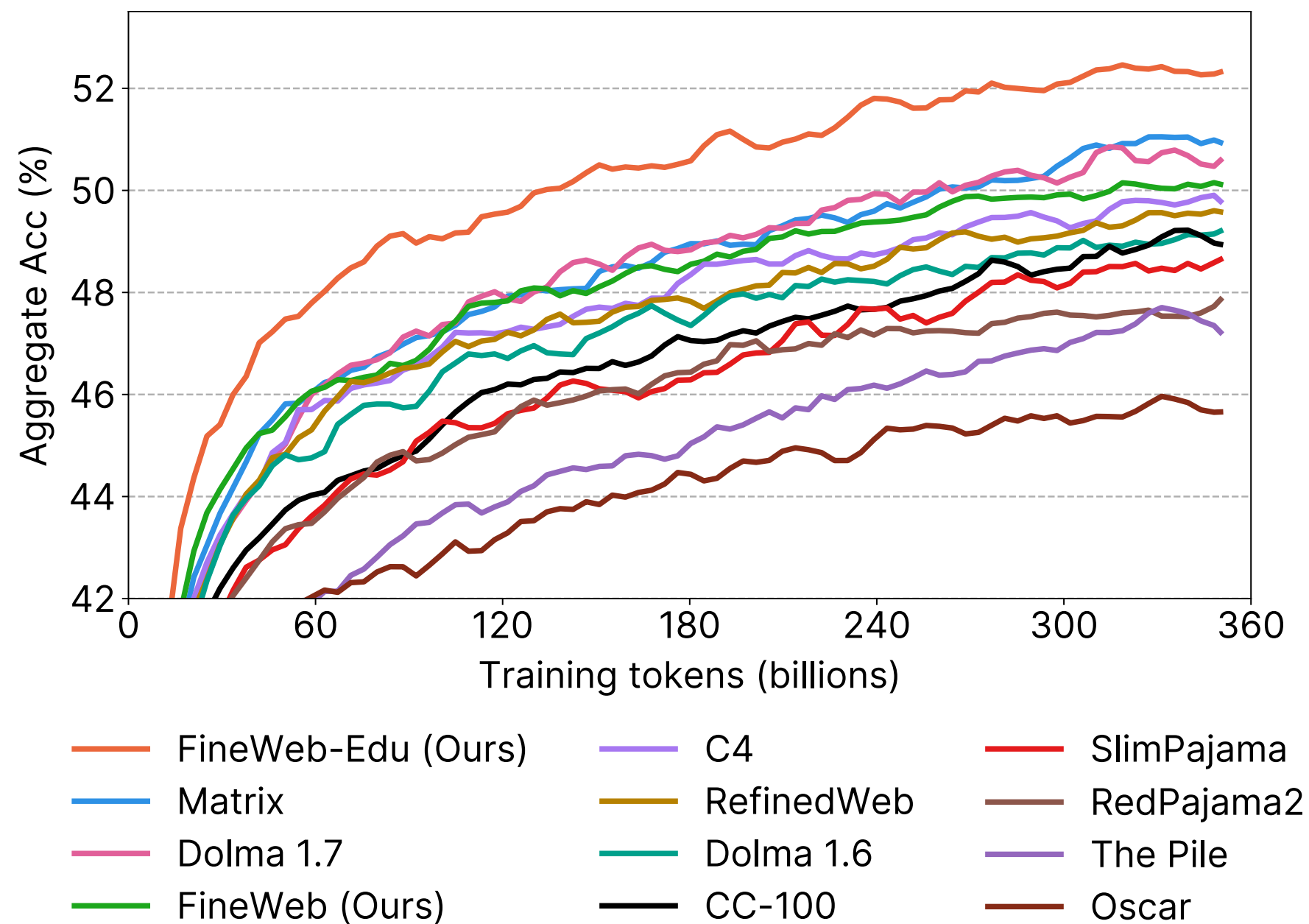
Below is an extract from a web page. Evaluate whether the page has a high educational value and could be useful in an educational setting for teaching from primary school to grade school levels using the additive 5-point scoring system described below. Points are accumulated based on the satisfaction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

Impact of high-quality data



Penedo et al. (2024)