Natural Language Processing

Data for LLM pretraining

Marco Kuhlmann
Department of Computer and Information Science

Data for LLM pretraining

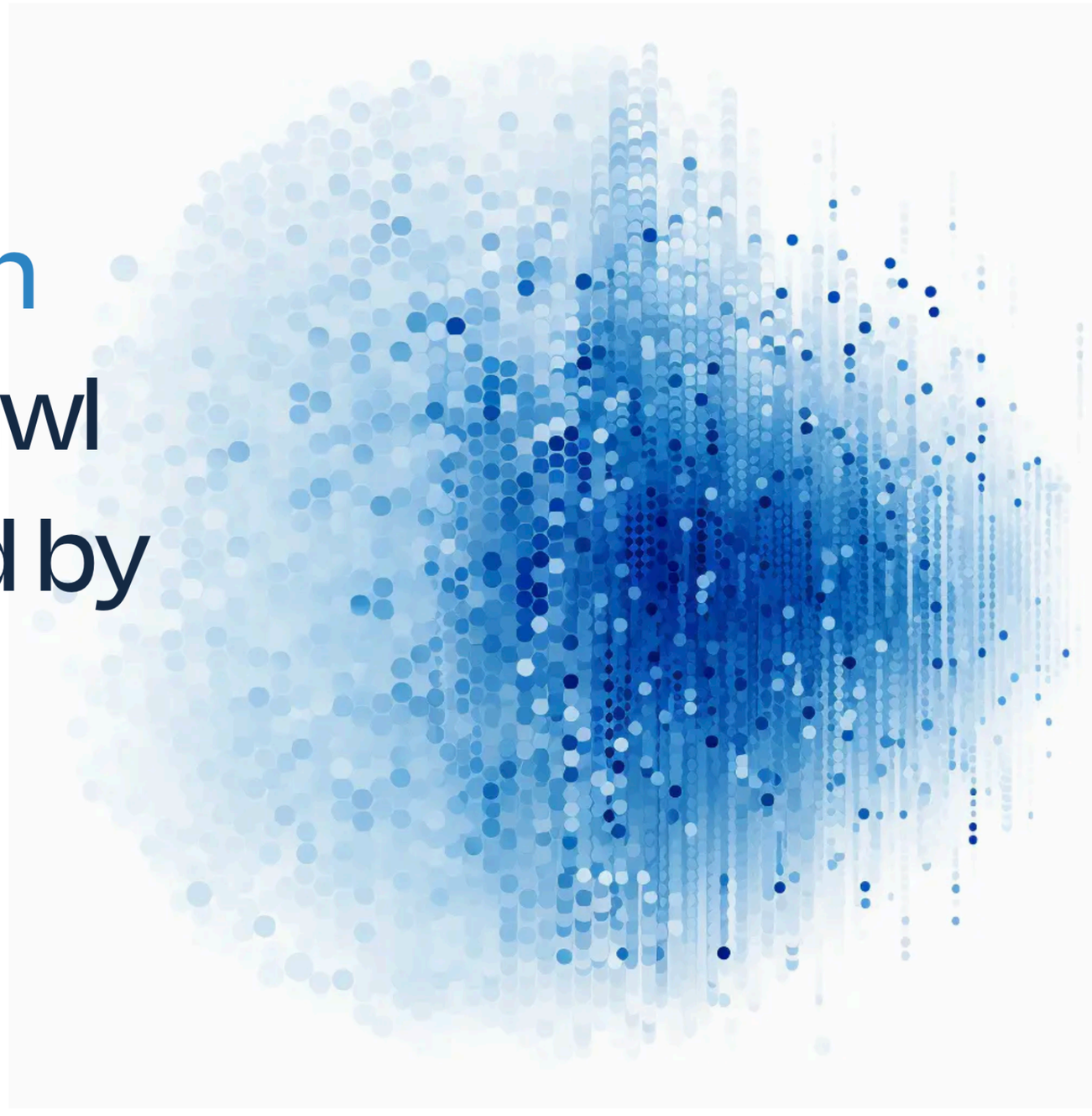
- Training modern LLMs demands vast amounts of data. This data is often sourced from the Internet.
- While abundant, Internet data is unstructed, noisy, and biased, making it an imperfect representation of language.
- Internet text data requires extensive postprocessing and quality filtering to enhance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

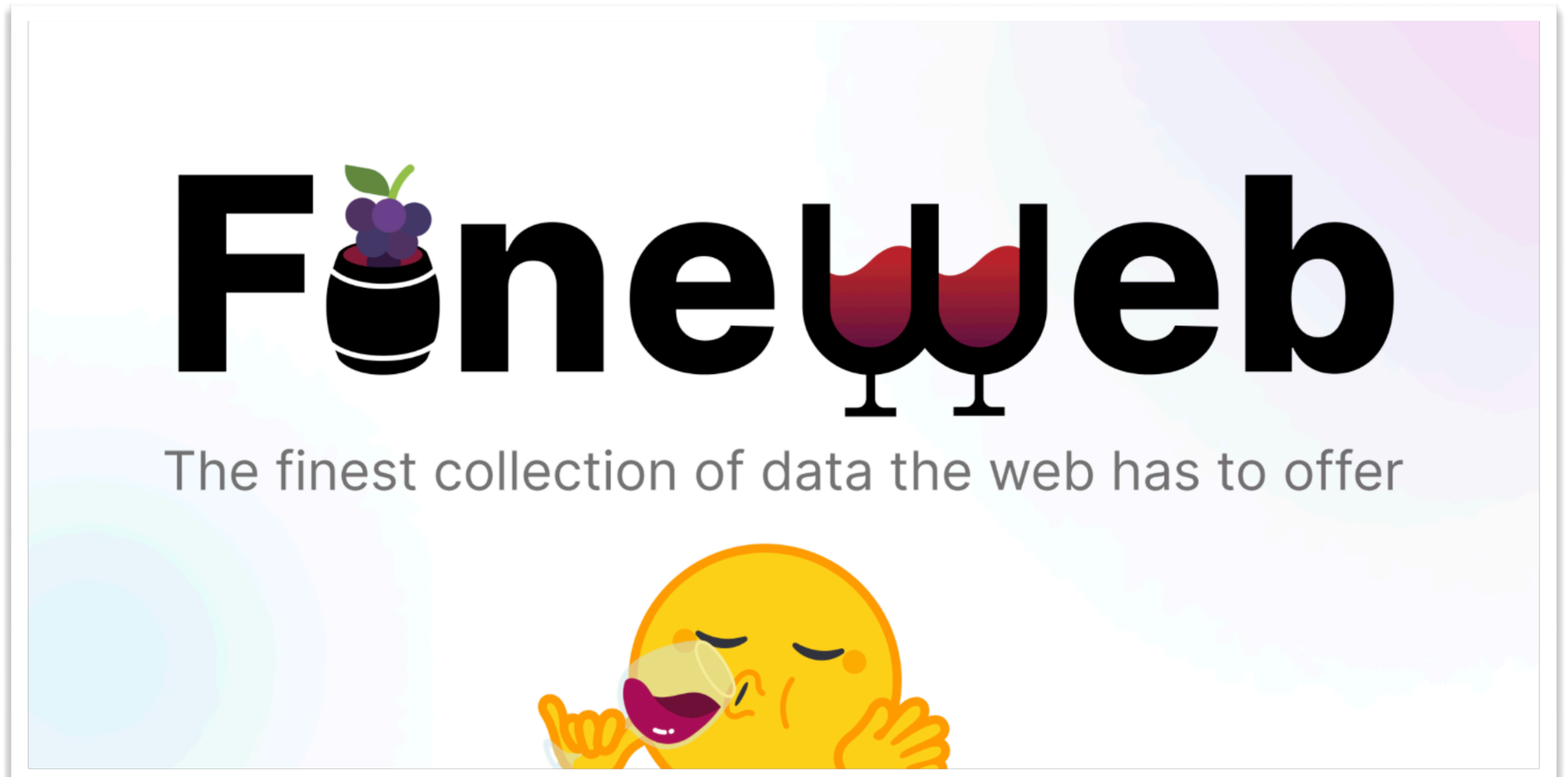
Overview



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

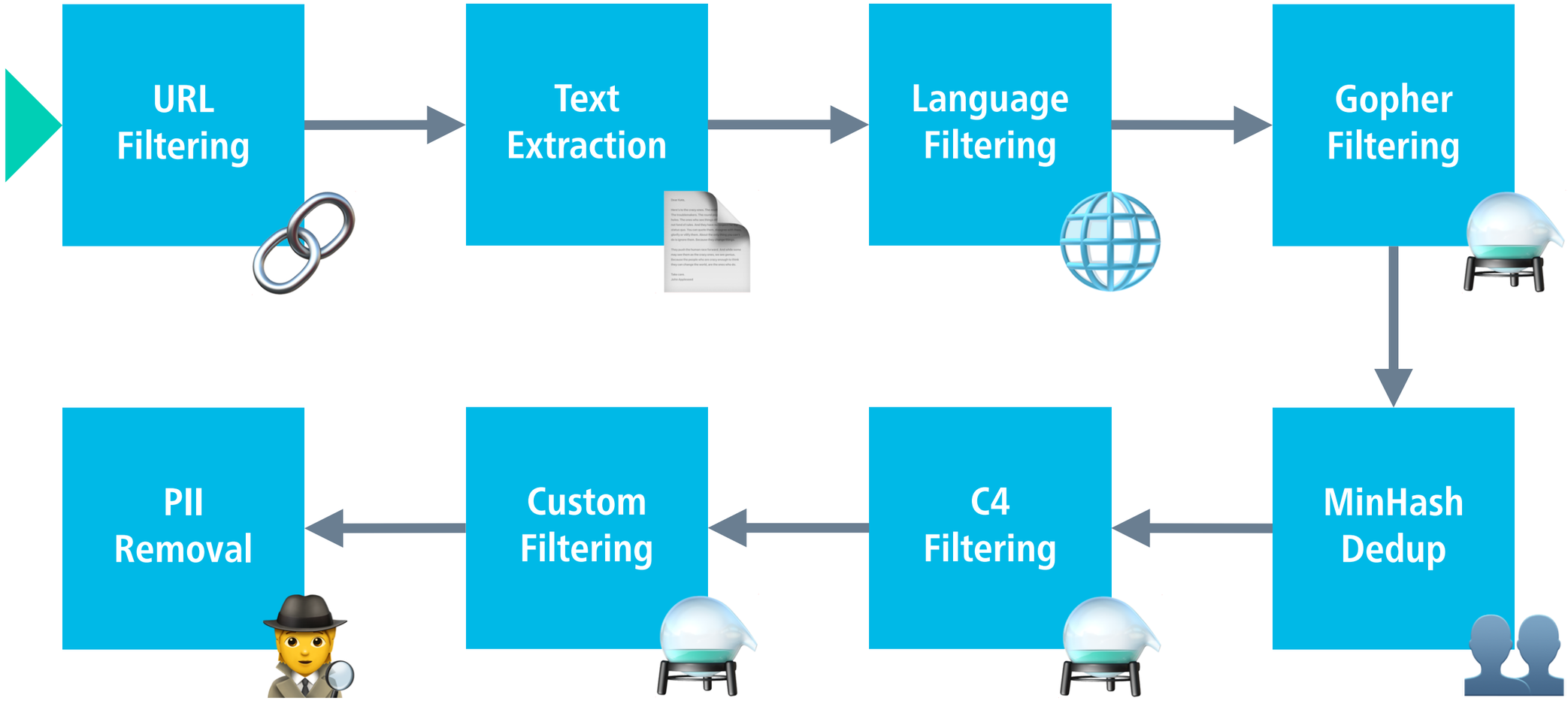
Source: Common Crawl



“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

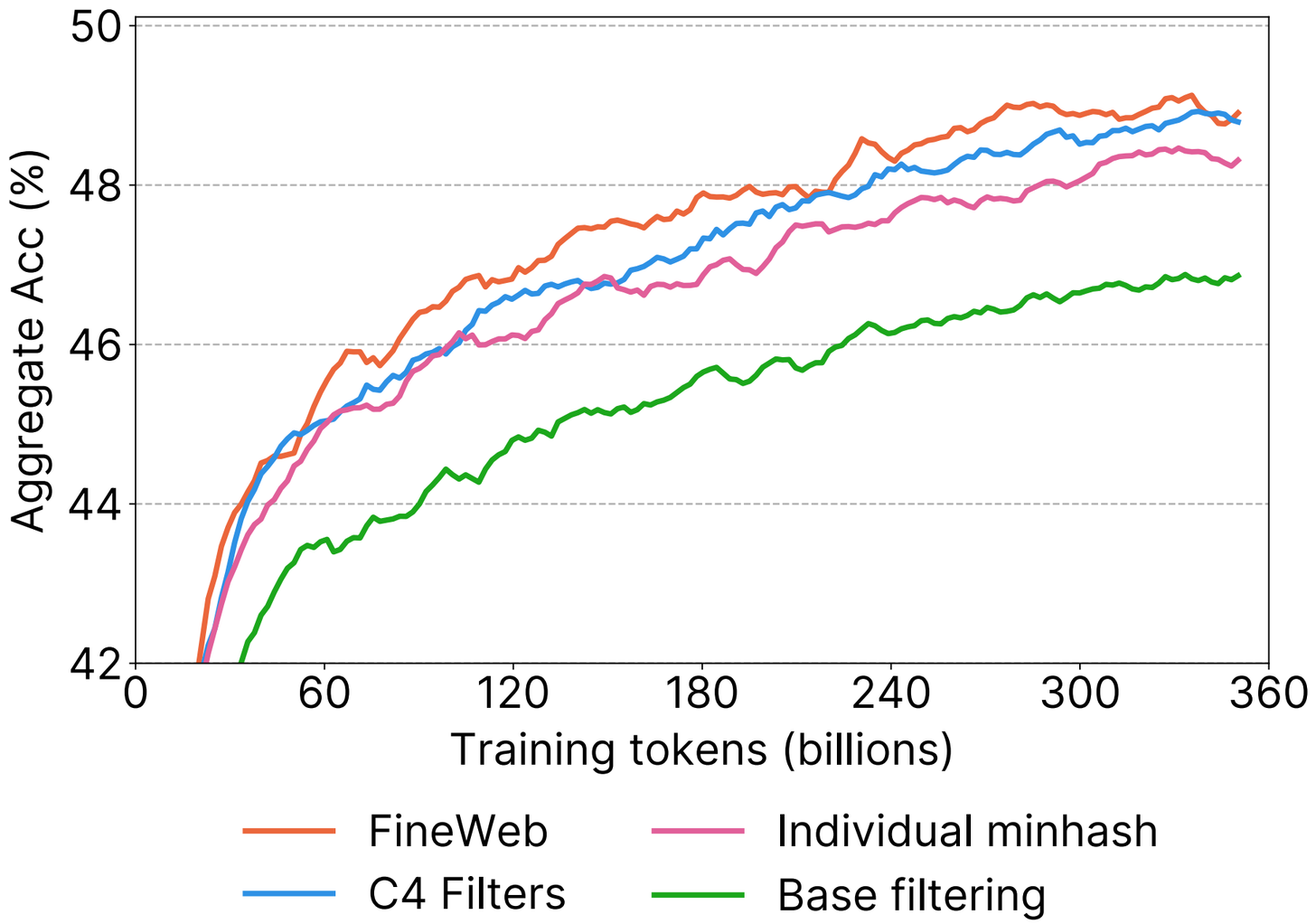
- Language filtering

  Uses fastText; keep only English text with a score ≥ 65%

- Heuristic filtering

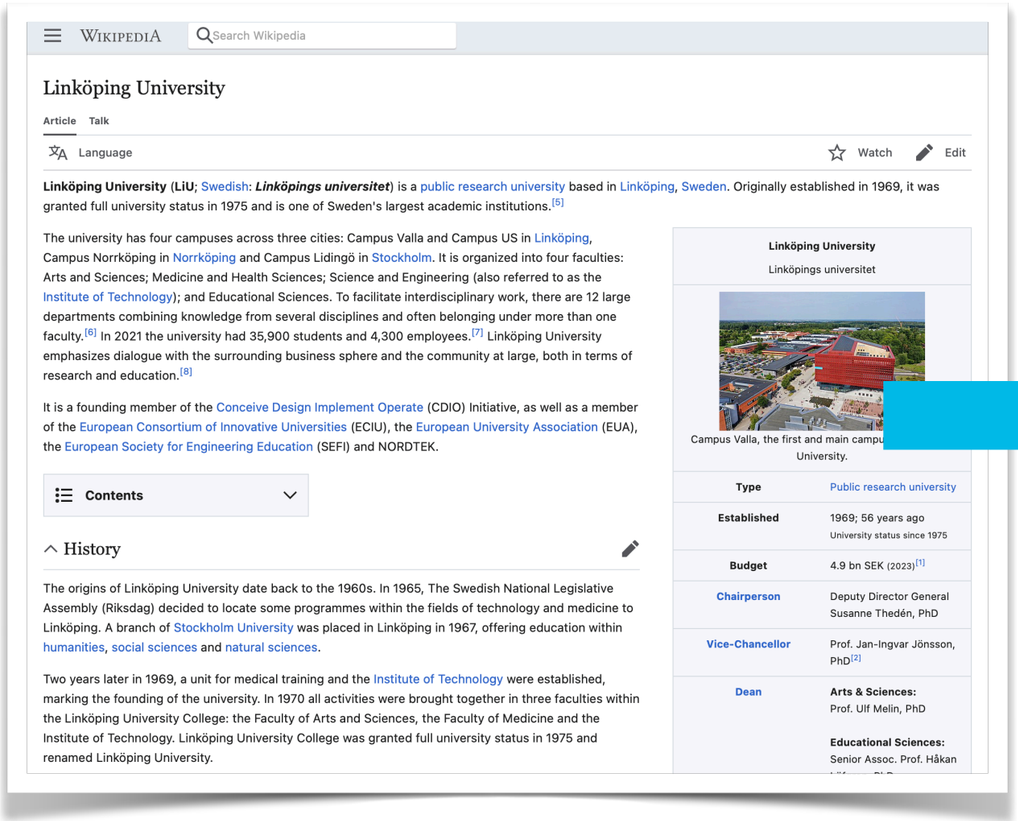
  Filter for length, symbol-to-word ratio, common/uncommon words, etc.

Impact of filtering



Penedo et al. (2024)

Text extraction



Wikipedia

Search Wikipedia

Linköping University

Article Talk

Language Watch Edit

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]

The university has four campuses across three cities: Campus Valla and Campus US in Linköping, Campus Norrköping in Norrköping and Campus Lidingö in Stockholm. It is organized into four faculties: Arts and Sciences; Medicine and Health Sciences; Science and Engineering (also referred to as the Institute of Technology); and Educational Sciences. To facilitate interdisciplinary work, there are 12 large departments combining knowledge from several disciplines and often belonging under more than one faculty.[6] In 2021 the university had 35,900 students and 4,300 employees.[7] Linköping University emphasizes dialogue with the surrounding business sphere and the community at large, both in terms of research and education.[8]

It is a founding member of the Conceive Design Implement Operate (CDIO) Initiative, as well as a member of the European Consortium of Innovative Universities (ECIU), the European University Association (EUA), the European Society for Engineering Education (SEFI) and NORDTEK.

Contents

History

The origins of Linköping University date back to the 1960s. In 1965, The Swedish National Legislative Assembly (Riksdag) decided to locate some programmes within the fields of technology and medicine to Linköping. A branch of Stockholm University was placed in Linköping in 1967, offering education within humanities, social sciences and natural sciences.

Two years later in 1969, a unit for medical training and the Institute of Technology were established, marking the founding of the university. In 1970 all activities were brought together in three faculties within the Linköping University College: the Faculty of Arts and Sciences, the Faculty of Medicine and the Institute of Technology. Linköping University College was granted full university status in 1975 and renamed Linköping University.

Linköping University

Linköpings universitet

Campus Valla, the first and main campu University.

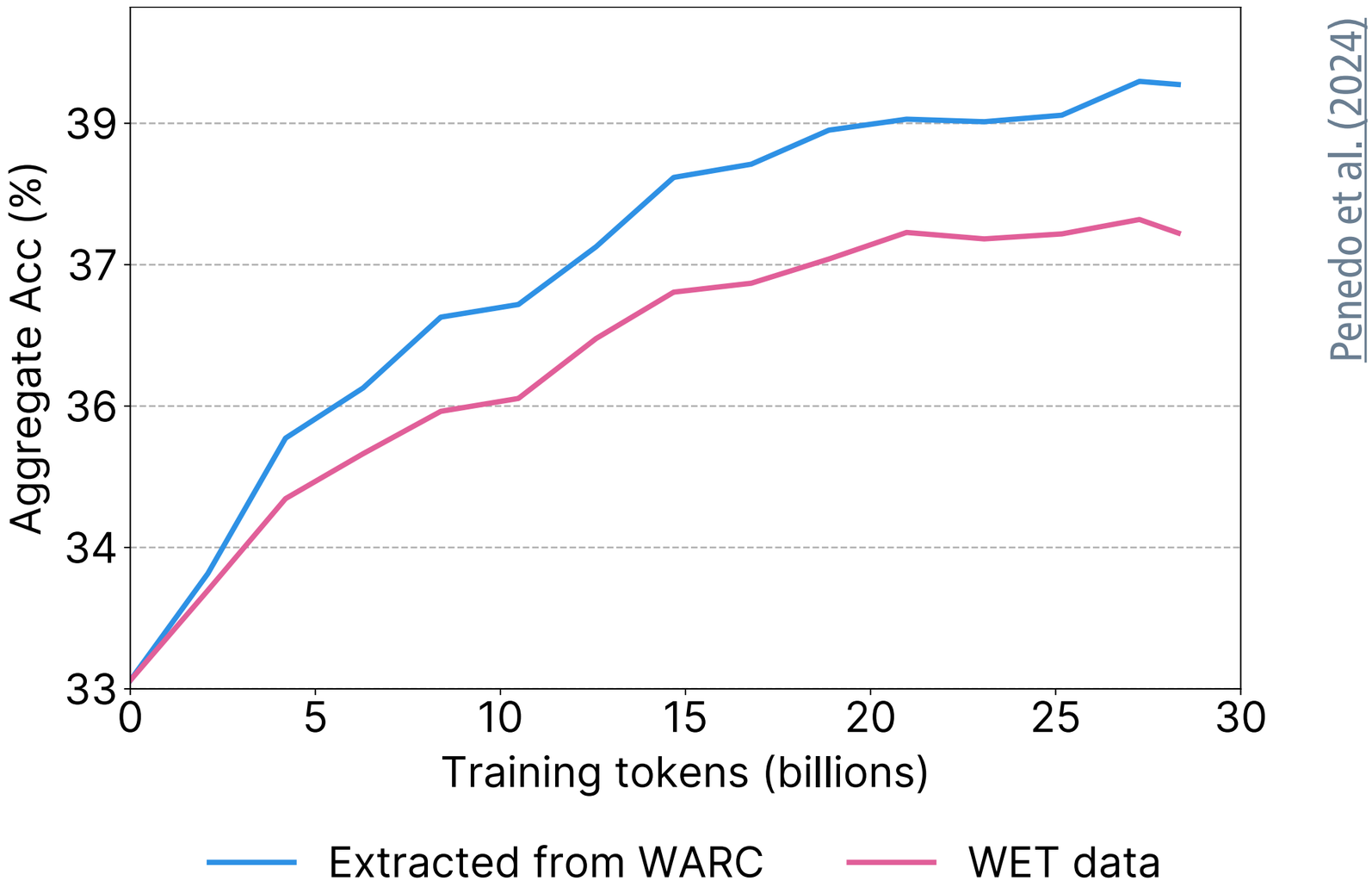
Type	Public research university
Established	1969; 56 years ago University status since 1975
Budget	4.9 bn SEK (2023)[1]
Chairperson	Deputy Director General Susanne Thedén, PhD
Vice-Chancellor	Prof. Jan-Ingvar Jönsson, PhD[2]
Dean	**Arts & Sciences:** Prof. Ulf Melin, PhD **Educational Sciences:** Senior Assoc. Prof. Håkan

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 e of Sweden's largest academic institutions.[5]

Linköpings universitet | |
Type	Public research university

FineWeb uses Trafilatura.

Relevance of text extraction



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
C4	Common Crawl	2019	156B
WebText	Own Crawl (OpenAI)	2019	300B
CC-100	Common Crawl	2020	532B
MassiveText	Own Crawl (Google)	2022	2.3T
OSCAR	Common Crawl	2023	523B
RedPajama	Common Crawl	2023	30.4T
RefinedWeb	Common Crawl	2023	500B
Dolma	Common Crawl	2024	3T
FineWeb	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

Penedo et al. (2024)

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.

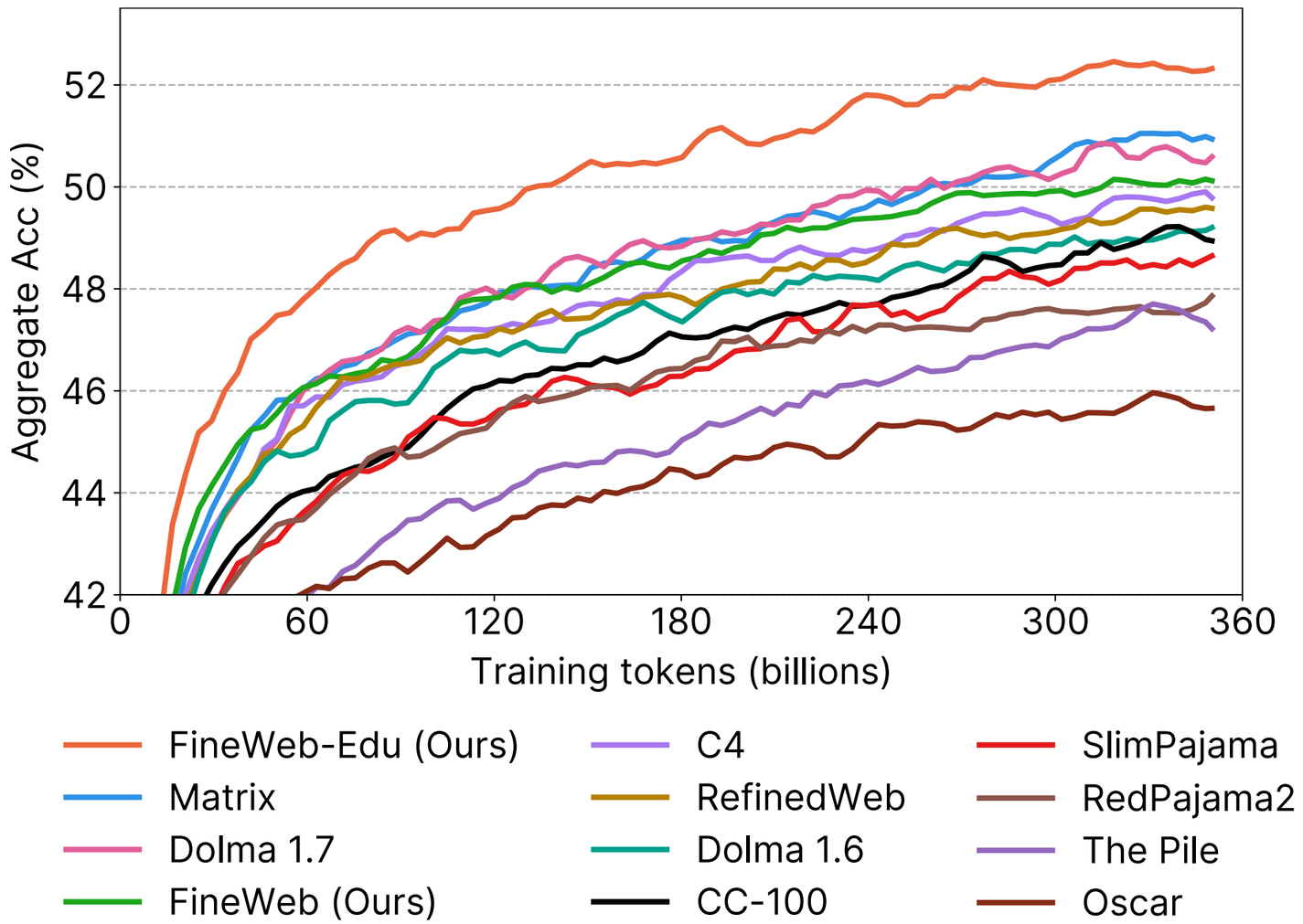
- Grant a fourth point if the extract highly relevant and beneficial for educational purposes for a level not higher than grade school, exhibiting a clear and consistent writing style. It could be similar to a chapter from a textbook or a tutorial, offering substantial educational content, including 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)