

Training Data Curation

Instructor

Sourab Mangrulkar

Machine Learning Engineer at

Creator of PEFT



Training Data Curation

Process of collecting and aggregating the training data

Training Data Curation

Mixture of filtered web data and curated high quality corpora

- Wikipedia
- Stackoverflow
- Github
- Social media conversations
- Books
- Technical papers

Principles

- Massive Scale

Principle: Massive Scale



GPT-3

300 Billion Tokens

Principle: Massive Scale



2 Trillion Tokens

Principle: Massive Scale



3500 Billion Tokens

Falcon LLM

Principles

- Massive Scale
- Diversity
- High Quality

Open Source Datasets

Common Crawl

Link: <https://commoncrawl.org/>

The screenshot shows the homepage of the Common Crawl website. At the top left is the Common Crawl logo, which consists of a hexagonal icon followed by the text "COMMON CRAWL". To the right are navigation links: "The Data", "Resources", "Community", "About", "Search", and a highlighted "Contact Us" button. The main content area features a large, stylized graphic of blue dots forming a central peak, with more dots fading out towards the edges. To the left of the graphic, the text reads: "Common Crawl maintains a free, open repository of web crawl data that can be used by anyone." Below this text is a paragraph stating: "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." At the bottom left is a "Overview" button.

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

Refined Web Dataset

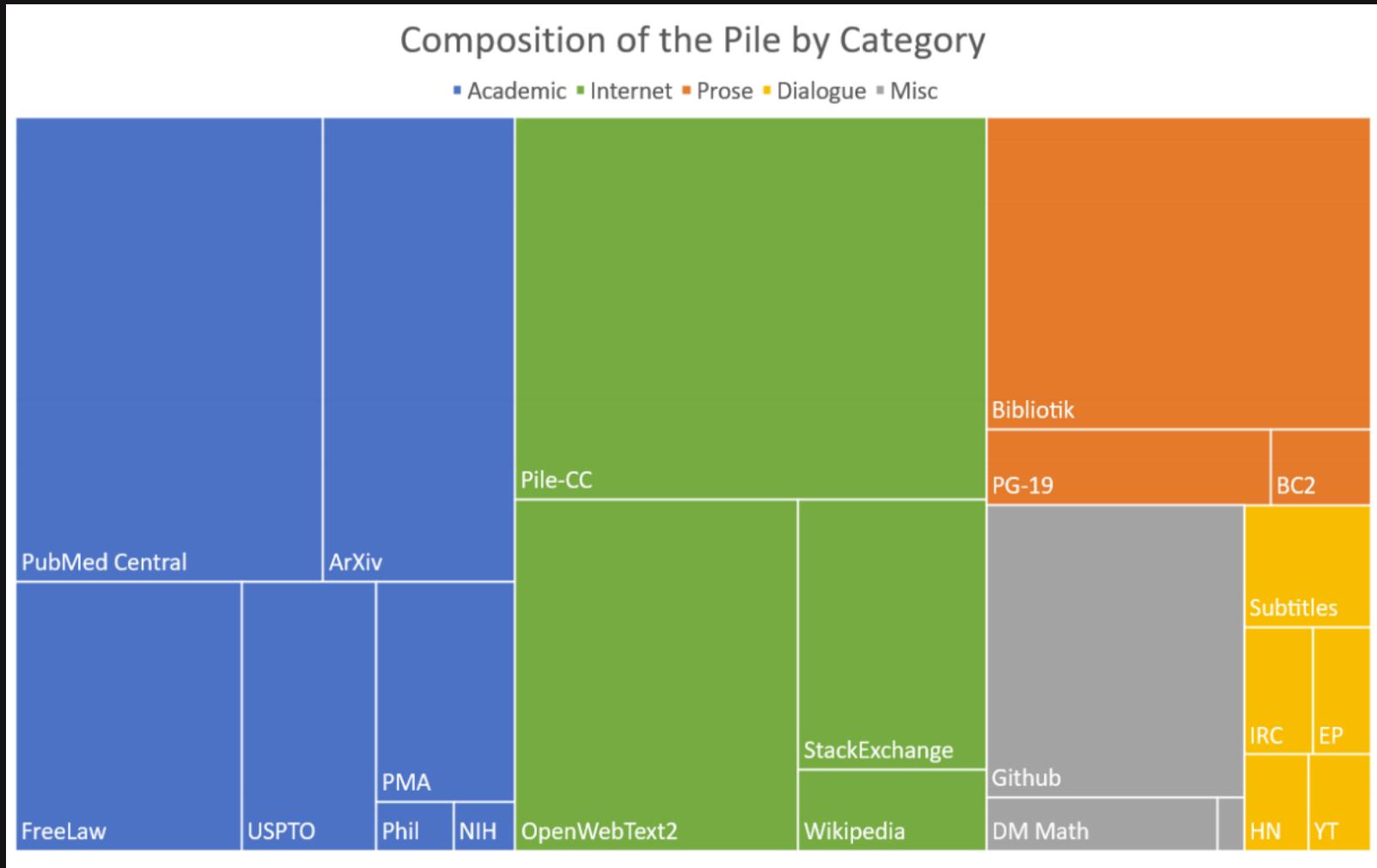
Link: <https://huggingface.co/datasets/tiiuae/falcon-refinedweb>

The screenshot shows the Hugging Face dataset page for 'falcon-refinedweb'. At the top, there's a navigation bar with links for Models, Datasets, Spaces, Posts, Docs, Solutions, Pricing, Log In, and Sign Up. Below the navigation is a search bar and a breadcrumb trail showing 'Datasets: tiiuae/falcon-refinedweb'. There are also like and 641 follower counts.

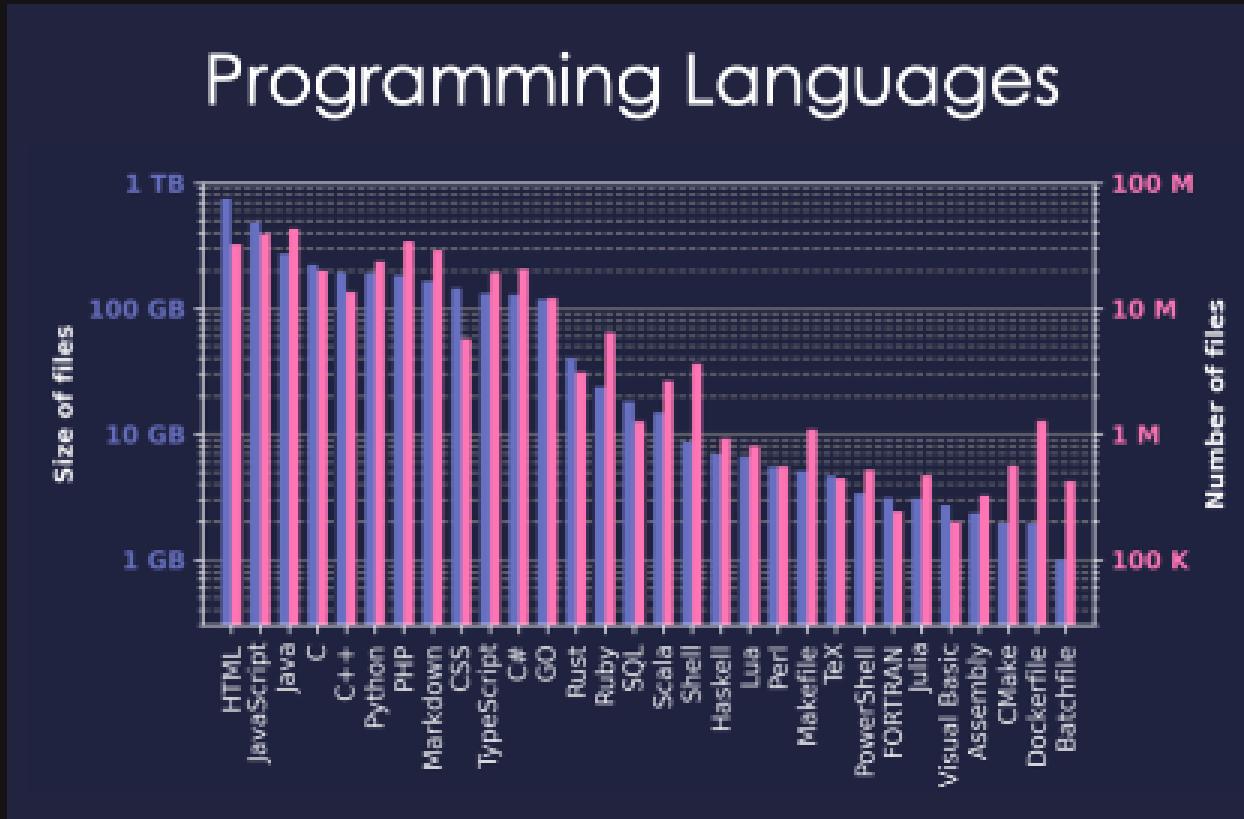
Below the header, there are filters for Tasks (Text Generation), Languages (English), Size Categories (100B < n < 1T), ArXiv links (arxiv:2306.01116, arxiv:2203.15556, arxiv:2107.06499), DOI (doi:10.57967/hf/0737), and License (ODC-BY).

The main content area has tabs for Dataset card (selected), Files and versions, and Community (16). A red warning box indicates that the dataset has 7 unsafe files. The Dataset Viewer section shows a preview of the 'train' split (968M rows) with columns: content, url, timestamp, dump, and segm. The content column shows text snippets from various web pages. To the right, there are statistics: Downloads last month (2,922), a button to Use in dataset library, and edit options. Below that are links to the homepage (falconlm.tii.ae), paper (arxiv.org), and point of contact (falconlm@tii.ae). It also shows the size of downloaded files (1.68 TB) and auto-converted Parquet files (1.68 TB), along with the number of rows (968,000,015). At the bottom, it lists models trained or fine-tuned on this dataset, including 'tiiuae/falcon-40b-instruct'.

Pile



The Stack



Math Pile

Link: <https://huggingface.co/datasets/GAIR/MathPile>

The screenshot shows the Hugging Face dataset page for "GAIR/MathPile". The top navigation bar includes "Datasets", the project name "GAIR/MathPile", a "like" button (140), and a search icon. Below the header, filters are shown for "Languages: English", "Size Categories: 1B< n <10B", "ArXiv: arxiv:2312.17120", and "License: cc-by-nc-sa-4.0". The main content area has three tabs: "Dataset card" (selected), "Files and versions", and "Community" (with 2 notifications). A prominent message box states: "You need to agree to share your contact information to access this dataset. This repository is publicly accessible, but you have to accept the conditions to access its files and content. By using this data, you agree to comply with the original usage licenses of all sources contributing to MathPile. If the source data of this dataset is subject to a more restrictive license than CC BY-NC-SA 4.0, then this dataset conforms to that more stringent licensing. In all other scenarios, it is governed by the CC BY-NC-SA 4.0 license. Access to this dataset is granted automatically once you accept the license terms and complete all the required fields below." It also says "Log in or Sign Up to review the conditions and access this dataset content." Below this, there's an "Update:" section with a list of recent changes:

- [2023/01/06] We release the commercial-use version of MathPile, namely [MathPile_Commercial](#).
- [2023/01/06] We release the new version (v0.2, cleaner version) of MathPile. It has been updated to the `main` branch (also the v0.2 branch). The main updates are as follows:
 - fixed a problem with the display of mathematical formulas in the Wikipedia subset, which was caused by the HTML conversion

On the right side, there's a sidebar with "Downloads last month: 183", buttons for "Use in dataset library" and "Edit dataset card", and a "More" options menu. Below these are sections for "Models trained or fine-tuned on GAIR/MathPile" and a list of five models:

- fblgit/UNA-POLAR-10.7B-InstructMath-v2 (Text Generation, Updated 16 days ago, 844 downloads, 3 stars)
- fblgit/UNA-POLAR-10.7B-InstructMath-v1 (Text Generation, Updated 16 days ago)
- Master2032/AL1 (Updated 5 days ago)
- Juanfco/Juanpancho (Updated 4 days ago)
- vahid625/test (Token Classification, Updated 2 days ago)

Steps involved in Training Data Curation

- Estimate training data size using scaling laws
- Focus on high-quality training data

Thank You
