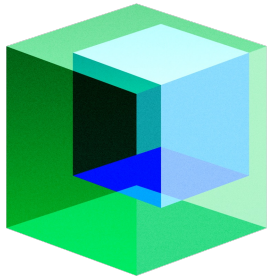
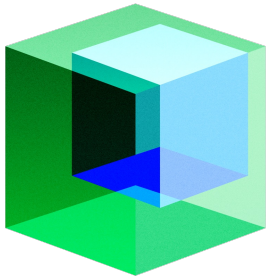
Technology Innovation Institute

OPEN SOURCE SUMMIT ABU DHABI 2025



Open-Source Data & Tools for the IBM Granite AI Models



Petros Zerfos, Ph.D.

pzerfos@us.ibm.com

Principal Research Scientist & Manager
Data & Tools for AI Models
IBM Research

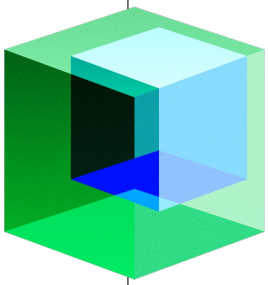
The Generative AI landscape is rapidly evolving

Small, Cost Effective — *Large, Expensive*

Open Source Small Language Models (SLMs)	Open Source Large Language Models	Closed Frontier Models
e.g. **Granite 4**, Phi-4, Gemma-3, Olmo 3, Qwen 3, etc.	e.g. Llama-4, DeepSeek-R1, Kimi-K2, Nemotron-340B	e.g. Gemini 3 Pro, GPT-5.1, Claude-3.7
Lowest power, but also most cost-effective.	Large, less powerful but catching up, less expensive but still impractical for every use case.	Large, powerful, overkill to use for every task. Trained with the pursuit of AGI

Granite

A family of enterprise-focused SLMs designed to help companies put GenAI to work.



Language

Granite-3.3-8B-Instruct
Granite-3.3-2B-Instruct

Granite 4.0 Family (NEW)

Safety Guardrails

Granite-Guardian-3.3-8B

Embeddings

Granite-Embeddings-30M-English
Granite-Embeddings-125M-English
Granite-Embeddings-107M-Multilingual
Granite-Embeddings-278M-Multilingual
Granite-Vision-3.3-2B-Embedding

Time Series

Granite-TimeSeries-Flowstate-r1.0

Vision

Granite-Vision-3.3-2B

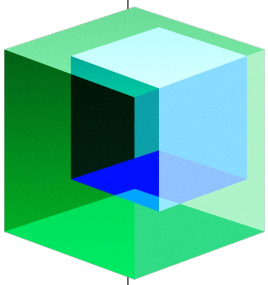
Granite-Docling-258M (NEW)

Speech

Granite-Speech-3.3-8B
Granite-Speech-3.3-2B

Granite 4

A family of enterprise-focused SLMs designed to help companies put generative AI to work.



Open

- Everything Granite available under the no-nonsense Apache 2.0 license
- Both instruct and base models shared for easy customization

Efficient

- Hybrid Mamba2 architecture designed for small GPU footprint and hardware-constrained deployments
- Range of model sizes provided to enable optimizing model size for a use case

Trusted

- First Open-Source model that is ISO 42001 certified
- Checkpoints cryptographically-signed to verify source of origin
- Launched with new white-hat hacker bug bounty program

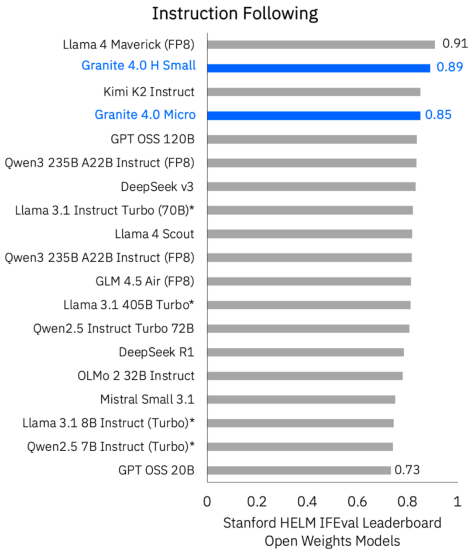
Granite 4

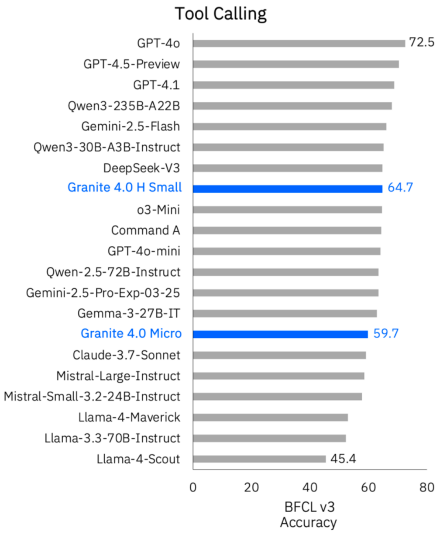
A family of enterprise-focused SLMs designed to help companies put generative AI to work.

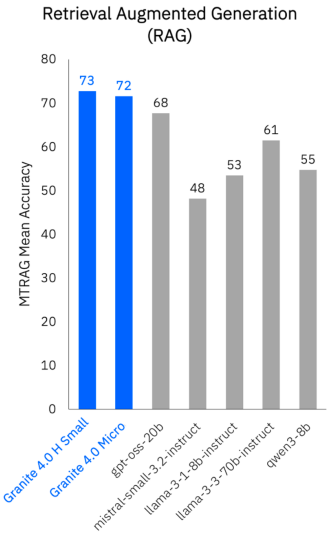
	Granite-4.0-H-Small	Granite-4.0-H-Tiny	Granite-4.0-H-Micro	Granite-4.0-Micro
Architecture Type	Hybrid, Mixture of Experts	Hybrid, Mixture of Experts	Hybrid, Dense	Traditional, Dense
Model Size	32B total parameters 9B activated parameters	7B total parameters 1B activated parameters	3B total parameters	3B total parameters
Intended Use	Workhorse model for key enterprise tasks like RAG and agents	Designed for low latency, edge, and local applications, and as a building block to perform key tasks (like function calling) quickly within agentic workflows*		Alternative option for users when Mamba2 support is not yet optimized (e.g. llama.cpp, PEFT, etc)
Est Memory Reqs (8-bit, 128K context length, batch = 1)	33 GB	8 GB	4 GB	9 GB
Example Hardware** (8-bit, 128K context length, batch = 1)	NVIDIA L40S (<$10K)	RTX 3060 12GB (<$1K)	Raspberry-Pi 8GB (<$100)	RTX 3060 12GB (<$1K)

*Further differentiating points between Tiny and Micro will be identified after evaluations are finalized
**Theoretical estimates based on provided specs, not tested

Granite 4.0 Optimizes Performance for Enterprise-Relevant Tasks



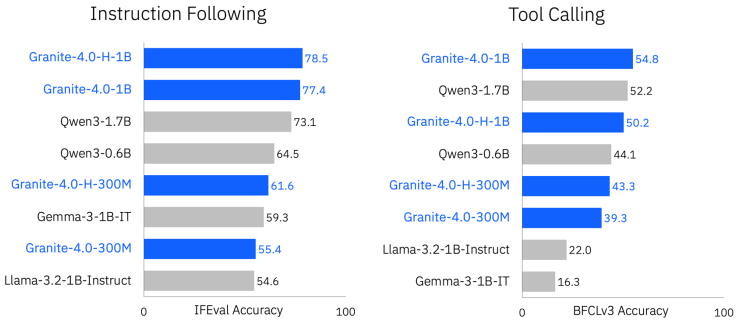


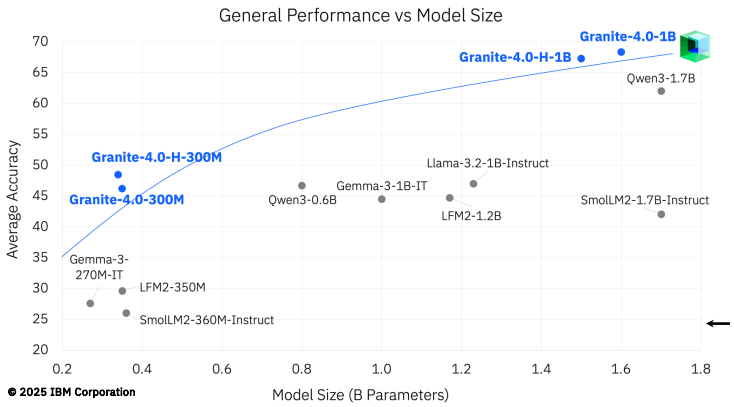


*Original model is open weights, but exact Turbo version could not be found on HF

Granite 4.0 Nano

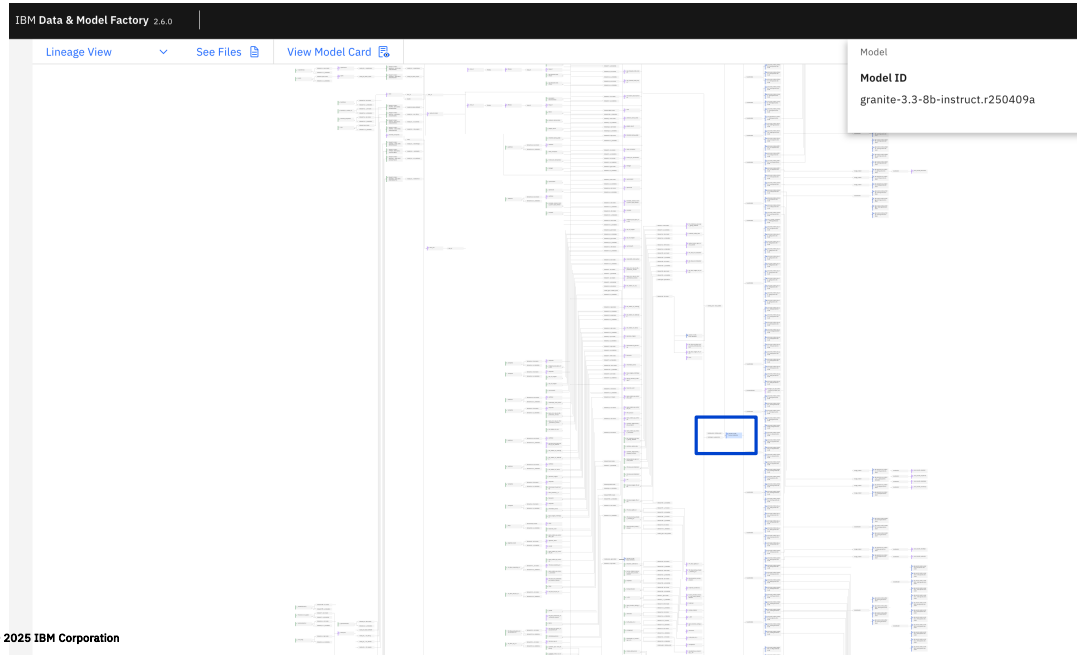
- **Granite 4.0 H 1B** – A ~1.5B parameter, dense LLM featuring a hybrid-SSM based architecture.
- **Granite 4.0 H 350M** – A ~350M parameter, dense LLM featuring a hybrid-SSM based architecture.
- **Granite 4.0 1B and Granite 4.0 350M** – Alternative traditional transformer versions of our 1B and 350M Nano models





Benchmarks across **General Knowledge** (MMLU, BBH), **Math** (GSM8K, GSM8K-Symbolic), **Code** (EvalPlus, CruxEval-O) **and Safety** (AttaQ, SALAD-Bench)

Trust depends on data and model governance

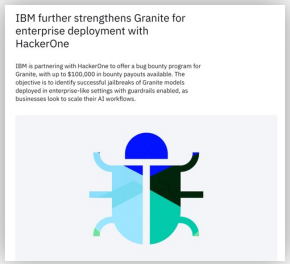


Technology Innovation Institute

IBM doubles down on trust & safety partnerships for Granite

- Ranked 1st in Stanford AI Transparency Index 2025 (from #4 in 2024)
- New $100k bug bounty program in partnership with HackerOne
- New red teaming partnership for Granite-vision with HiddenLayer
- New partnership with Schellman LLC. Granite obtained ISO 42001 compliance certification
- Renewed red teaming partnership with Robust Intelligence

schellman
Quality, above all.


IBM further strengthens Granite for enterprise deployment with HackerOne

IBM is partnering with HackerOne to offer a bug bounty program for Granite, with up to $100,000 in bounty payouts available. The objective is to identify successful jailbreaks of Granite models deployed in enterprise-like settings with guardrails enabled, as businesses look to scale their AI workflows.


schellman

CERTIFICATE OF REGISTRATION
Artificial Intelligence Management System – ISO/IEC 42001:2023

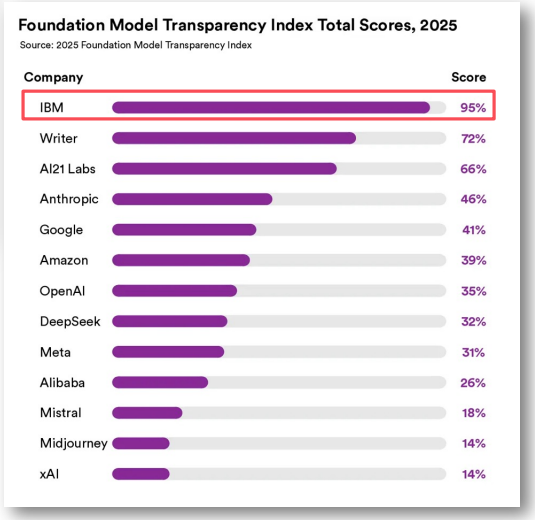
International Business Machines Corporation

hackerone

HIDDENLAYER
PROTECT YOUR ADVANTAGE

ROBUST INTELLIGENCE
now part of CISCO

OPEN SOURCE SUMMIT
ABU DHABI 2025


Foundation Model Transparency Index Total Scores, 2025

Source: 2025 Foundation Model Transparency Index

Company	Score
IBM	95%
Writer	72%
AI21 Labs	66%
Anthropic	46%
Google	41%
Amazon	39%
OpenAI	35%
DeepSeek	32%
Meta	31%
Alibaba	26%
Mistral	18%
Midjourney	14%
xAI	14%

Stanford AI Transparency Index 2025

https://crfm.stanford.edu/fmti/December-2025/index.html

Technology Innovation Institute

OPEN SOURCE SUMMIT ABU DHABI 2025

Every conversation in AI starts with Models and ends with Data

79% identify data preparation and generation* as the most common strategic task performed by AI teams.

30% view data volume and complexity* as one of the most challenging aspects of AI implementation.

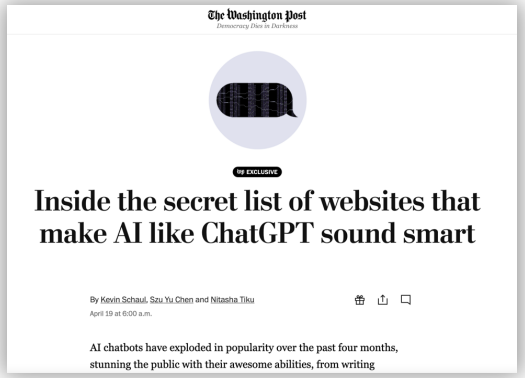
Quality of data affects the quality of the model

George Fuechsel,
IBM 305 RAMAC
technician

* Gartner, Explore Data-Centric AI Solutions to Streamline AI Development, 2023

IBM

Issues with GenAI training data curation



The Washington Post

Democracy Dies in Darkness

WP EXCLUSIVE

Inside the secret list of websites that make AI like ChatGPT sound smart

By Kevin Schaul, Szu Yu Chen and Nitasha Tiku

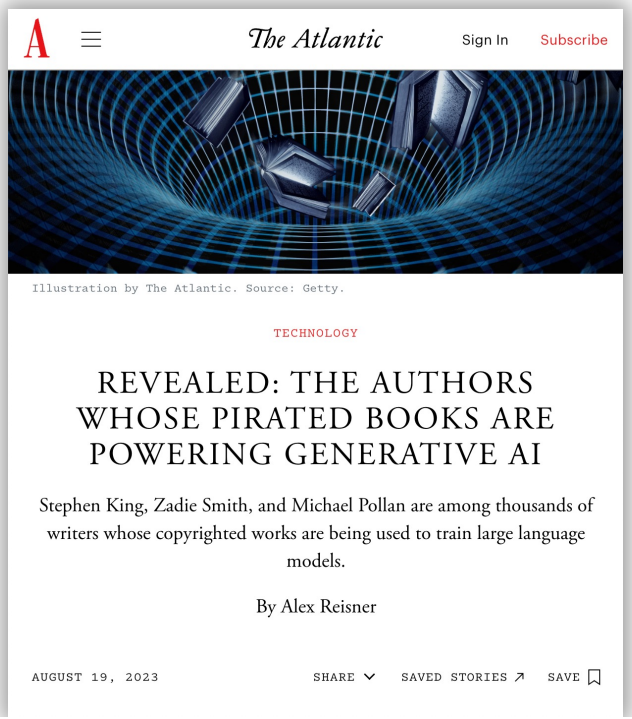
April 19 at 6:00 a.m.

AI chatbots have exploded in popularity over the past four months, stunning the public with their awesome abilities, from writing

Fresh concerns raised over sources of training material for AI systems

Investigations reveal limited efforts to 'clean' datasets of fascist, pirated and malicious material

The Guardian



The Atlantic Sign In Subscribe

Illustration by The Atlantic. Source: Getty.

TECHNOLOGY

REVEALED: THE AUTHORS WHOSE PIRATED BOOKS ARE POWERING GENERATIVE AI

Stephen King, Zadie Smith, and Michael Pollan are among thousands of writers whose copyrighted works are being used to train large language models.

By Alex Reisner

AUGUST 19, 2023 SHARE SAVED STORIES SAVE

Technology Innovation Institute

OPEN SOURCE SUMMIT ABU DHABI 2025

IBM

GneissWeb was used to train IBM Granite 4.0

- Trending in Top-5 at HuggingFace 🤗
  (as of Oct. 7th)
- Granite 4.0:
  - Granite-4.0-H-Small (MoE 32B/A9B)
  - Granite-4.0-H-Tiny (MoE 7B/A1B)
  - Granite-4.0-H-Micro (Dense 3B)

IBM Granite 4.0 Release:
https://www.ibm.com/new/announcements/ibm-granite-4-0-hyper-efficient-high-performance-hybrid-models

Good models need Good data!

Open datasets for pre-training LLMs

Public, non-profit
Internet crawl

~310B web pages

→Text extraction from HTML

→Language identification

→Heuristic rule-based filters

Number of words, character repetition, filler words, URL ban list, ...

→Deduplication

→Model-based filtering

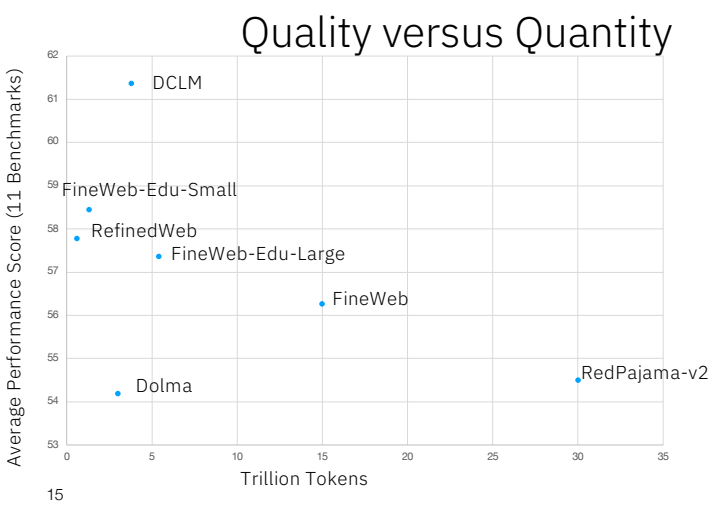
Optionally mix other data sources
(e.g., Wikipedia, ArXiv, SSRN, FIDC, SEC, Edgar, FreeLaw,...)

- C4 [Raffel et al., 2020]
- Pile [Gao et al., 2020]
- Redpajama [2023]
- RefinedWeb [Penedo et al., 2024]
- Dolma [Soldaini et al., 2024]
- FineWeb [Penedo et al., 2024]
- FineWeb-Edu [Penedo et al., 2024]
- DCLM-Baseline [Li et al., 2024]

Limitations of current pre-training datasets

State-of-the-Art Datasets

- Large datasets (5T+ tokens) – limited performance
- Aggressive filtering to achieve quality
- Rely on model-based filtering for quality, often requiring GPUs



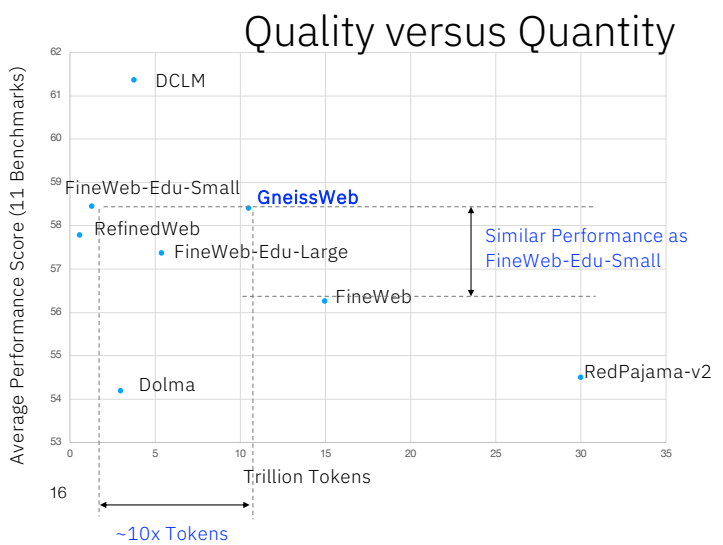
GneissWeb: advancing open innovation in large-scale training data

State-of-the-Art Datasets

- Large datasets (5T+ tokens) – limited performance
- Aggressive filtering to achieve quality
- Rely on model-based filtering for quality, often requiring GPUs

GneissWeb

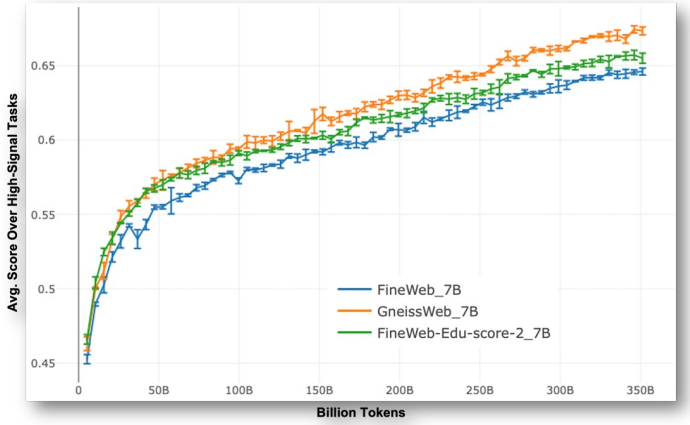
- 10T tokens / 12.6B docs / 35TB size with quality higher than large (5T+ tokens) SOTA datasets
- CPU-friendly recipes (model+heuristics)
- Novel quality annotators, judiciously designed ensemble filter
- Finance data: ~5TB



GneissWeb: SOTA Common Crawl for pre-training

- 10T token dataset derived from FineWeb V1.1.0
- Ablation models *outperform* those trained on FineWeb V1.1.0 by **>1.5%** over 20 benchmarks
- Open-source tools and recipes for reproduction

https://huggingface.co/datasets/ibm-granite/GneissWeb



Combining GneissWeb components into a winning recipe

Exact Substring Deduplication

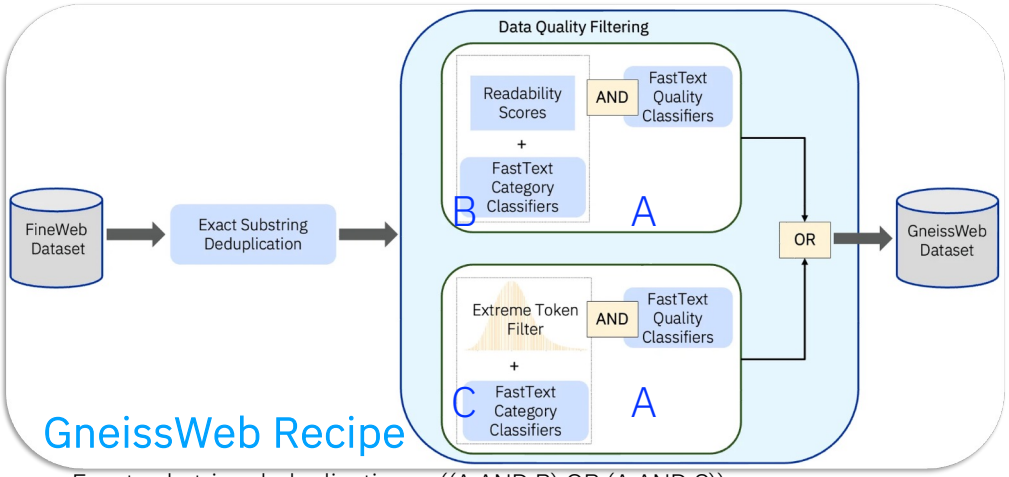
Custom Data Quality Classifiers

Filtering based on Readability Scores

Filtering Extreme-Tokenized Documents

Document Category Classifiers

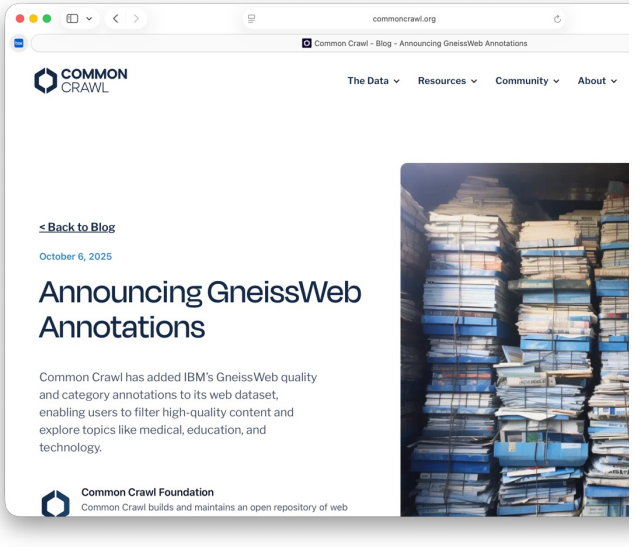
Novel Components highlighted in



- Exact substring deduplication → ((A AND B) OR (A AND C))
- GneissWeb ensemble filtering: A document is retained if either the fastText combination and category-aware, readability score filter agree to retain,
- OR the fastText combination and category-aware, extreme-tokenized filter agree to retain

Common Crawl Foundation is bringing GneissWeb to everyone!

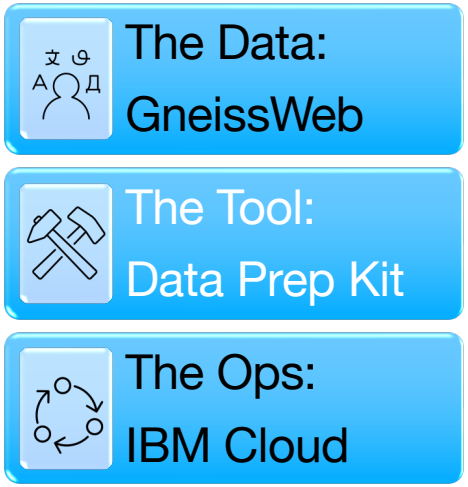
- Announced at IBM TechXchange 2025
- Using IBM's Bloom filter + Data Prep Kit (LF AI & Data Project)
- Applied it to every URL in the Crawls
- Created annotations
  - Quality score: passes GneissWeb's standard or not
  - Category: science, medical, education, etc.
- Published annotations at both URL- and host-level



Common Crawl Foundation Announcement:
https://commoncrawl.org/blog/announcing-gneissweb-annotations

Technology Innovation Institute

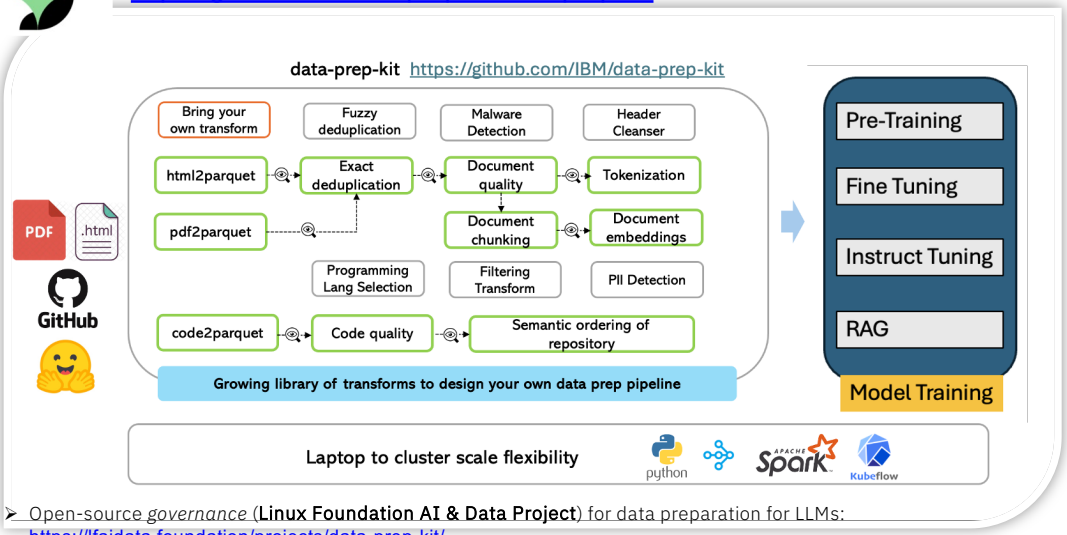
OPEN SOURCE SUMMIT ABU DHABI 2025



IBM

Data Prep Kit

https://github.com/data-prep-kit/data-prep-kit



- Open-source *governance* (**Linux Foundation AI & Data Project**) for data preparation for LLMs: https://lfaidata.foundation/projects/data-prep-kit/
- Spearheaded by IBM Research
- Used in the data preparation for IBM's Granite AI model training: https://huggingface.co/ibm-granite

Technology Innovation Institute

OPEN SOURCE SUMMIT ABU DHABI 2025

Large collection of transforms

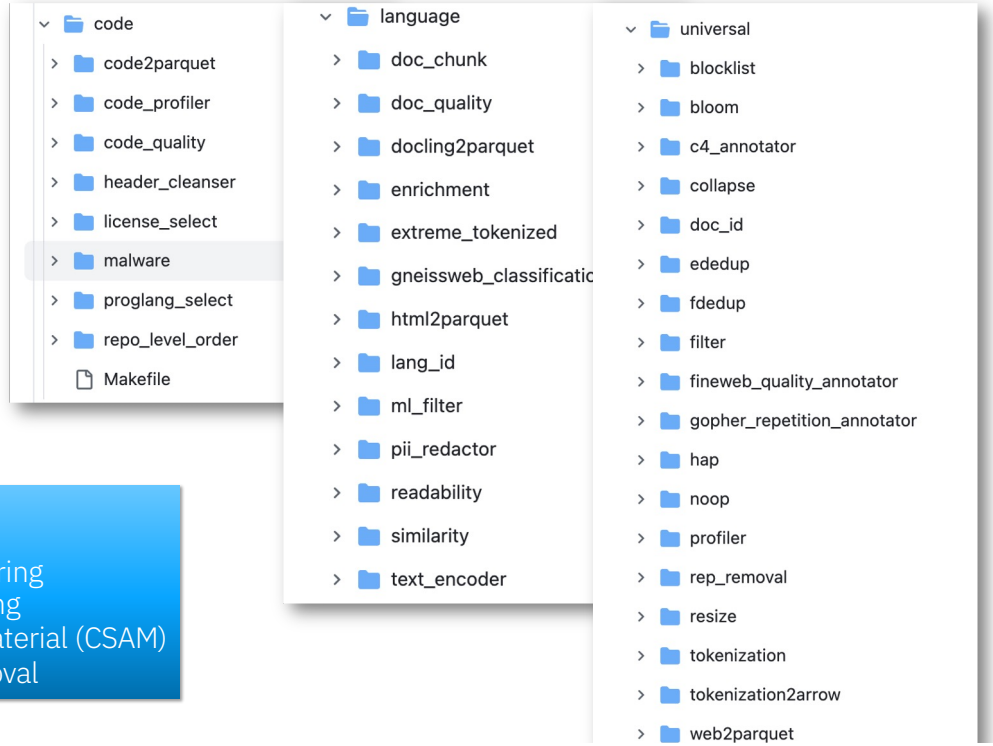
40+ built in Transforms

Bring Your Own Transform

- Content Extraction (including Granite-Docling)
- Document and Code quality
- Data Enrichment
- Annotations and filtering
- Language Identification
- Quality Annotation
- Deduplication
- Data Safety and Security
- Optimization & ready for training
- Bring your own Transform

IBM

Data Prep Kit Transforms

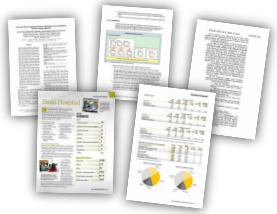


Granite Vision (VLM)

- Face blurring
- License-based filtering
- Non-US data filtering
- Child Sex Abuse Material (CSAM) detection and removal

Introducing Docling

1. **Open source** and permissively licensed (**Apache 2.0**)
2. **Cost performant**
3. Packaged as a **python library (no API's)**



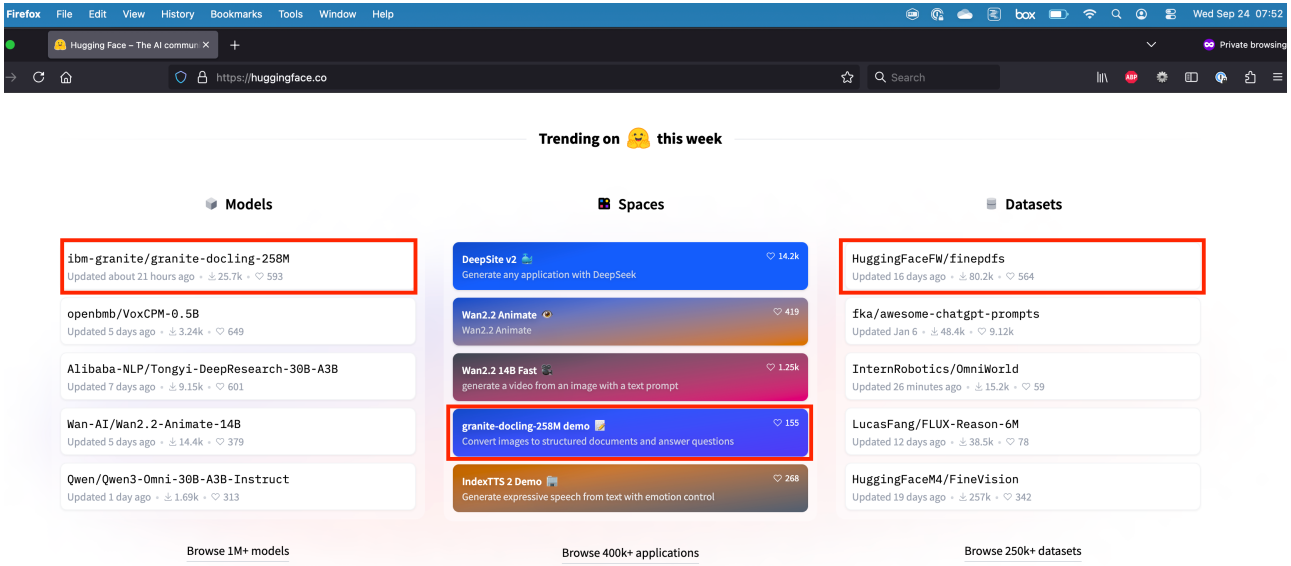
```
pip install docling

# a single document to markdown
docling https://arxiv.org/pdf/2408.09869.pdf

# a folder of documents to markdown and json
docling --to json --to md ./inputs/
```

- Parsing of multiple document formats incl. PDF, DOCX, XLSX, HTML, images, and more
- Unified, expressive DoclingDocument representation format
- Various export formats (Markdown, HTML, JSON)
- Many plug-and-play ecosystem integrations
- Support of OCR and Visual Language Models
- Support for Audio with Automatic Speech Recognition models
- Simple and convenient CLI

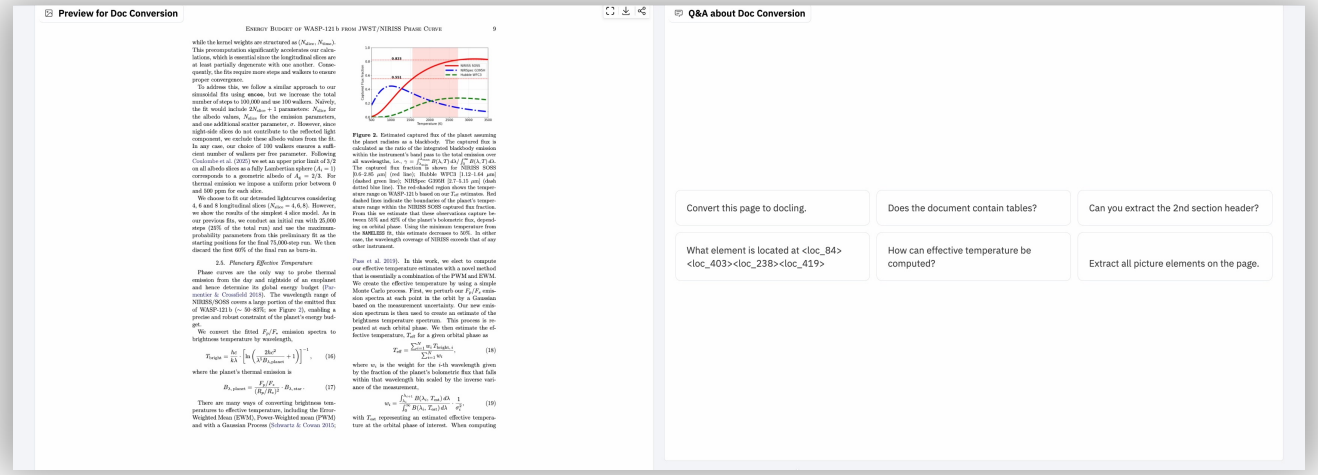
Granite-Docling Adoption



Technology Innovation Institute

OPEN SOURCE SUMMIT ABU DHABI 2025

http://ibm.biz/granite-docling-demo

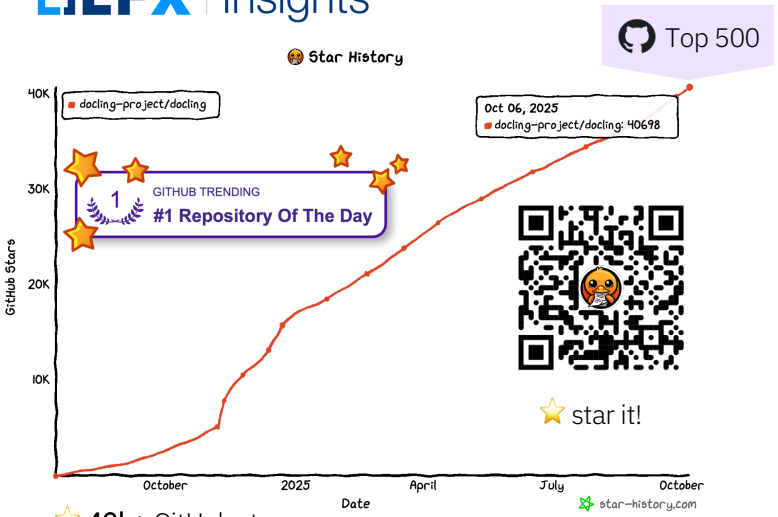


Granite-Docling content extraction supports Arabic and right-to-left reading flow

IBM

Docling Community Adoption

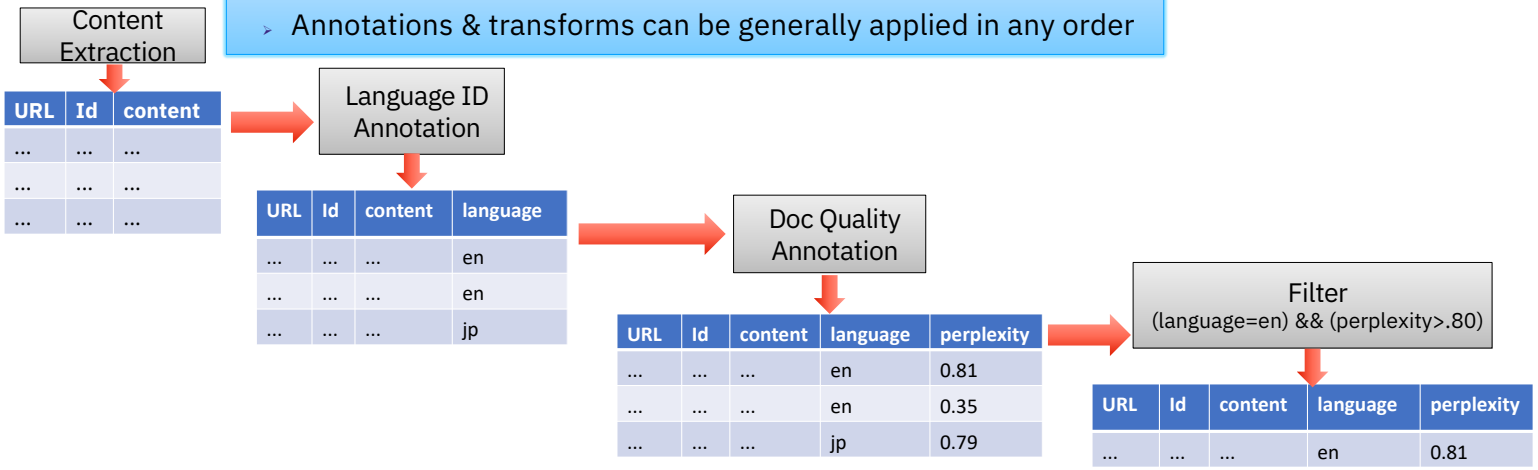
LFX | Insights



- 42k+ GitHub stars
- 1.5M+ downloads/month from PyPi

Unstructured Data Representation in Data Prep Kit

- Documents (language & code) are represented in a CSV-like form, in PyArrow tables (Parquet files)
- Annotations & transforms can be generally applied in any order



Data Prep Kit: from laptop- to cluster- scale

Laptop

Server

Datacenter

- Ray and Spark wrappers for scale-out with no code changes*
- Docker/Podman desktop, Kind cluster, local (file system) I/O, Cloud Object Storage, etc.
- Production-ready use: checkpointing, metadata, auditing
- Low-code pipeline orchestration via (KFP)

*for embarrassingly parallel transforms

Technology Innovation Institute

OPEN SOURCE SUMMIT ABU DHABI 2025

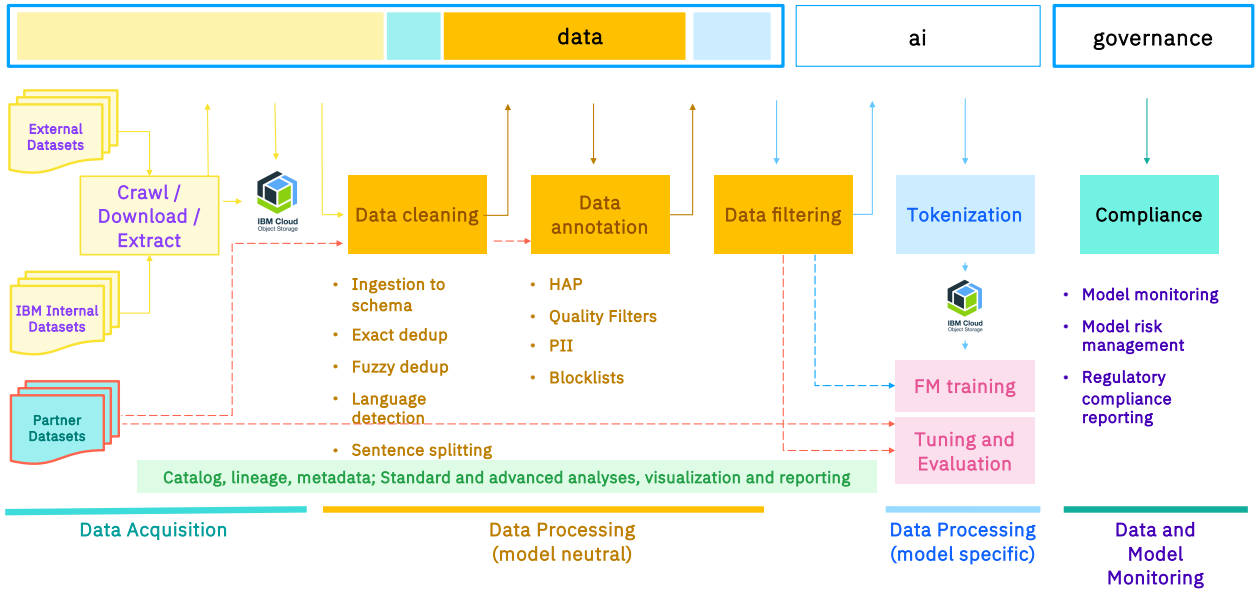
The Data:
GneissWeb

The Tool:
Data Prep Kit

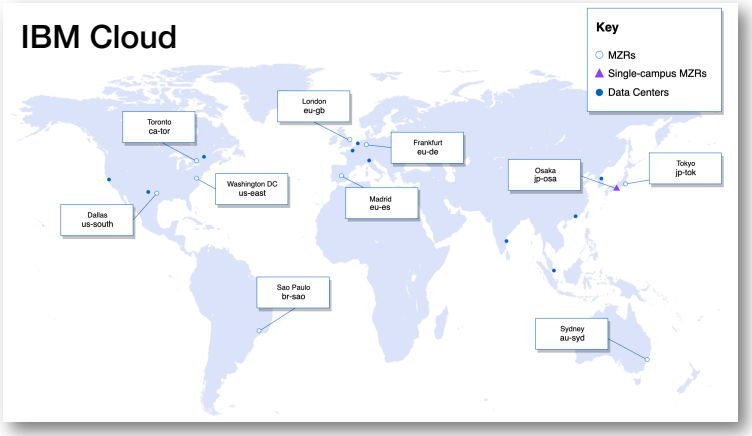
The Ops:
IBM Cloud

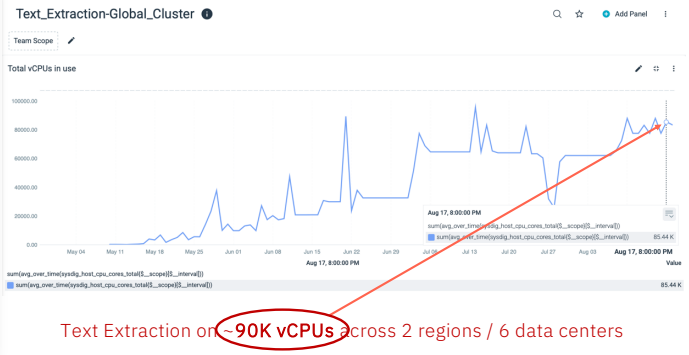
IBM

Data engineering journey: from raw data to models



Data Engineering for LLMs @ IBM Cloud





Large-scale data processing with the IBM Cloud as “Supercomputer”

- text extraction from the HTML pages of the 12PBs of Common-Crawl dataset (~309B docs)


Technology
Innovation
Institute


OPEN SOURCE
SUMMIT
ABU DHABI
2025


IBM