Open Data Flows

Rethinking AI infrastructure after the synthetic turn

Pierre-Carl Langlais
Pleias

Commons AI
December 10th, 2025

About pleias

Pleias is a Paris-based startup that's on a mission to **solve the key AI scalability challenges** for sensitive industries — data quality, lack of efficiency, compliance and security risks.

We provide clients with **vertical AI solutions at a fraction of traditional AI costs** thanks to our **powerful yet frugal foundation models.**

Members of the AI alliance and CurrentAI, we believe in the necessity of **open, copyright-free and factual data for AI**.

That's why we've released **Common Corpus - the largest fully open corpus for pre-training:** 2 trillion tokens with document-level licensing, provenance and language information.

pleias

WIKIMEDIA FOUNDATION

AI Alliance

mozilla

1. “We don’t talk about the data” – state of LLM pretrain

Training data issues...

Language models come with **a large number of data issues:**

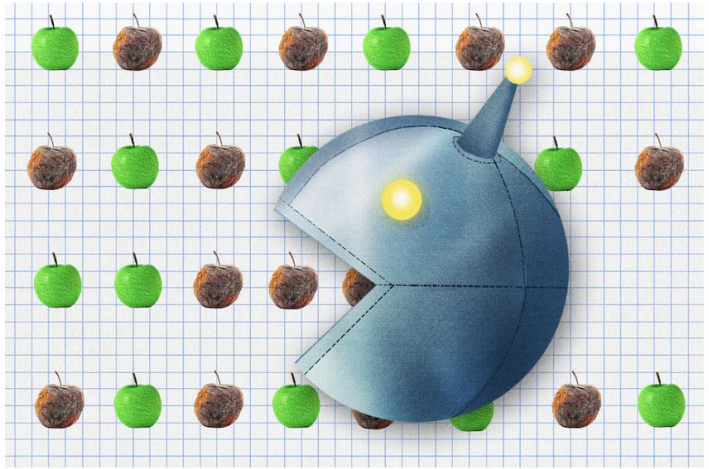
- Largest official source is **web archives** with no possible way to filter out problematic (or **poisoned**?) content at scale.
- In practice, big labs seem to routinely use shadow libraries and other sources of **pirated content**. This practice is at the center of the Meta trial.
- "**We don't talk about the data**": despite its centrality in training labs never communicate over the datasets.

Russia seeds chatbots with lies. Any bad actor could game AI the same way.

In their race to push out new versions with more capability, AI companies leave users vulnerable to "LLM grooming" efforts that promote bogus information.

April 17, 2025

11 min 86



(Washington Post illustration; iStock)

...are deployer liabilities.

In the current legislation, deployers of models are fully liable.

- You have no guarantees the model won't **output copyrighted content**.
- You can't be completely sure the alignment really fit your regulations and expected norms (the "DeepSeek" problem")
- You don't know whether the model is really able to process internal data which may be **widely different** from the internet data used for training: half of crawled archives are less than 300 words.

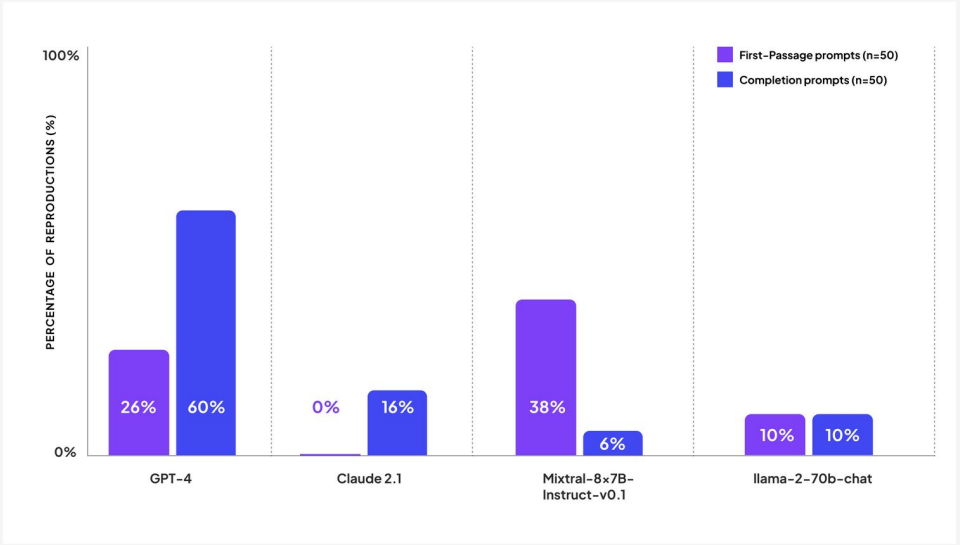
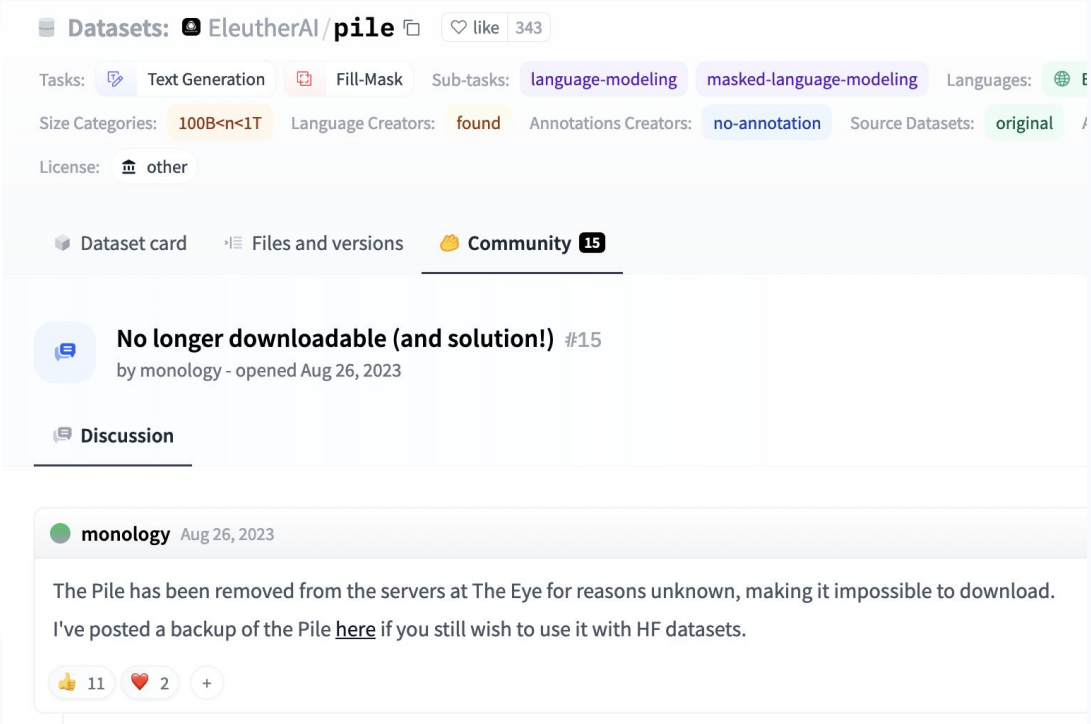


Figure 1: Percentage of prompts resulting in exact reproductions from copyrighted works

Tragedy of the commons

While closed labs are protected by obfuscation, open research efforts have been much more precarious. Datasets and models are routinely removed and sometimes this even leds to trial.

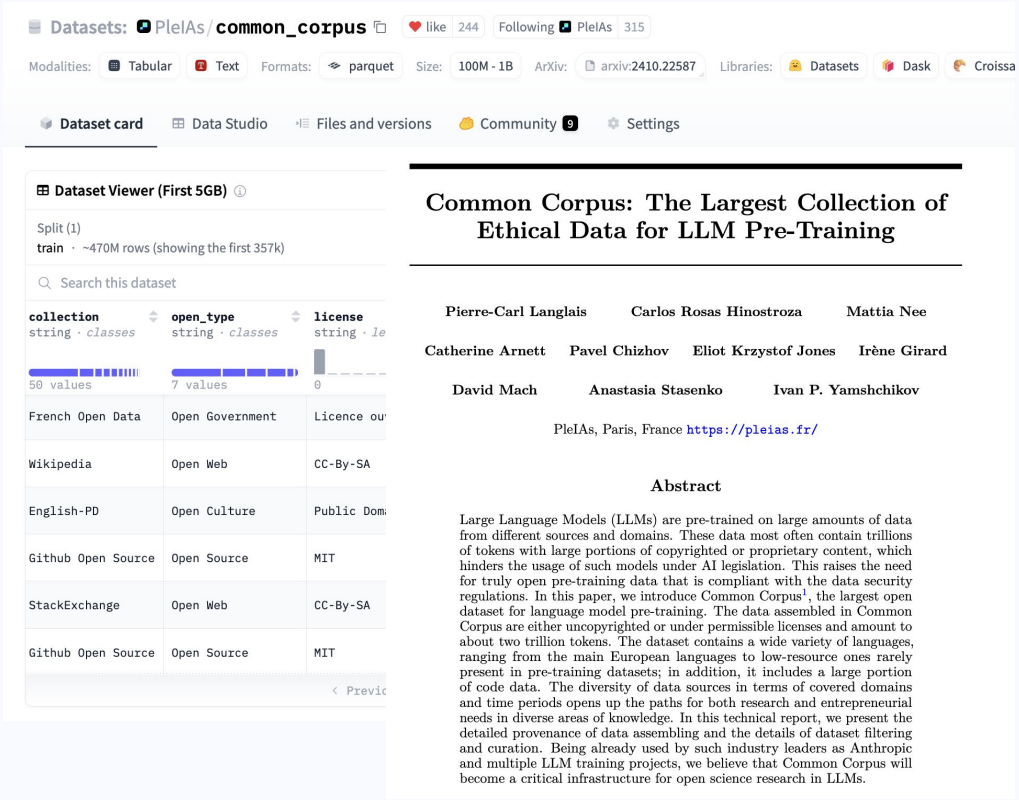
The issue is especially preeminent in Europe due to the absence of **fair use**. Text & data mining exception only cover fair use, not **releasability**. Right now, most “open everything” LLMs rely on HuggingFace being hosted in the US.



Open Pretraining Data

Common Corpus is the largest collection of open and releasable pretraining data. It's made of 500 millions documents (2 trillions tokens) all associated to an open license.

Common Corpus is an integral part of the Open Trusted Data Initiative and aims to embody the principles of data attribution.



Datasets: PleIAs / common_corpus like 244 Following PleIAs 315

Modalities: Tabular Text Formats: parquet Size: 100M - 1B ArXiv: arxiv:2410.22587 Libraries: Datasets Dask Croissa

Dataset card Data Studio Files and versions Community 9 Settings

Dataset Viewer (First 5GB)

Split (1)
train · ~470M rows (showing the first 357k)

Search this dataset

collection string · classes	open_type string · classes	license string · le
50 values	7 values	0
French Open Data	Open Government	Licence ou
Wikipedia	Open Web	CC-By-SA
English-PD	Open Culture	Public Dom
Github Open Source	Open Source	MIT
StackExchange	Open Web	CC-By-SA
Github Open Source	Open Source	MIT

Previc

Common Corpus: The Largest Collection of Ethical Data for LLM Pre-Training

Pierre-Carl Langlais Carlos Rosas Hinostroza Mattia Nee
Catherine Arnett Pavel Chizhov Eliot Krzystof Jones Irène Girard
David Mach Anastasia Stasenko Ivan P. Yamshchikov

PleIAs, Paris, France https://pleias.fr/

Abstract

Large Language Models (LLMs) are pre-trained on large amounts of data from different sources and domains. These data most often contain trillions of tokens with large portions of copyrighted or proprietary content, which hinders the usage of such models under AI legislation. This raises the need for truly open pre-training data that is compliant with the data security regulations. In this paper, we introduce Common Corpus[1], the largest open dataset for language model pre-training. The data assembled in Common Corpus are either uncopyrighted or under permissible licenses and amount to about two trillion tokens. The dataset contains a wide variety of languages, ranging from the main European languages to low-resource ones rarely present in pre-training datasets; in addition, it includes a large portion of code data. The diversity of data sources in terms of covered domains and time periods opens up the paths for both research and entrepreneurial needs in diverse areas of knowledge. In this technical report, we present the detailed provenance of data assembling and the details of dataset filtering and curation. Being already used by such industry leaders as Anthropic and multiple LLM training projects, we believe that Common Corpus will become a critical infrastructure for open science research in LLMs.

Open Pretraining Data

Since its release, Common Corpus has been downloaded more than 700,000 times, which make it one of the most popular pretraining dataset along with FineWeb and C4.

While we don't know the full extent of reused, its range goes way beyond models we pretrained at Pleias and include Nvidia (Parakeet), Anthropic (Circuit Transformers) Open tooling has been further instrumental for the training Harvard Initiative and the training of Apertus.

ELSEVIER
CrossMark
Available online at www.sciencedirect.com
ScienceDirect
Procedia Computer Science 267 (2025) 146–156
Procedia Computer Science
www.elsevier.com/locate/procedia

Proceedings of the Third EuroHPC user day

Pleias 1.0: the First Ever Family of Language Models Trained on Fully Open Data

Pierre-Carl Langlais[a,*], Pavel Chizhov[a,b], Mattia Nee[a], Carlos Rosas Hinostroza[a], Matthieu Delsart[a], Irène Girard[a], Anastasia Stasenko[a], Ivan P. Yamshchikov[a,b]

[a] PleIAs, Paris, France
[b] THWS, Würzburg, Germany

Abstract

Linguistic diversity and strong generalization in foundation language mode
tokens with very large model parameter counts. However, most such train
protected or private data that is not explicitly published under the licence
ethical concerns. We introduce **Pleias 1.0**, a family of comparatively sma
parameters) trained *exclusively on public domain or permissively licensed d*
that our results are fully auditable and reproducible. Furthermore, we fine-tu
(RAG) task and demonstrate that these models – despite their smaller si
magnitude more parameters on RAG evaluations. All models, data, and
standard for transparency and compliance.

Keywords: fully open source language models; multilingual small language model

Datasets 1,042 | Filter by nar | Full-text search | ↑↓ Sort: Most downloads

m-a-p/FineFineWeb
Viewer • Updated Dec 19, 2024 • 4.89B • 503k • 50

allenai/c4
Viewer • Updated Jan 9, 2024 • 10.4B • 432k • 415

HuggingFaceFW/fineweb
Viewer • Updated Jan 31 • 25B • 386k • 2.16k

jat-project/jat-dataset
Viewer • Updated Feb 16, 2024 • 258M • 365k • 40

HuggingFaceFW/fineweb-edu
Viewer • Updated Jan 31 • 3.3B • 163k • 679

PleIAs/common_corpus
Viewer • Updated Feb 11 • 470M • 149k • 259

Open Pretraining Data

Canary 1B

MOSEL

Parakeet V2
AUTOMATIC SPEECH RECOGNITION

FineVideo

Multimodal extension

Institutional Data Initiative
at Harvard Law School Library

APERTVS

Tooling

Common Corpus

Explai-nability

Transformer Circuits Thread

LLM Training

GPT-NL

pleias

Salamandra

The Common Corpus extended universe

Reality hit

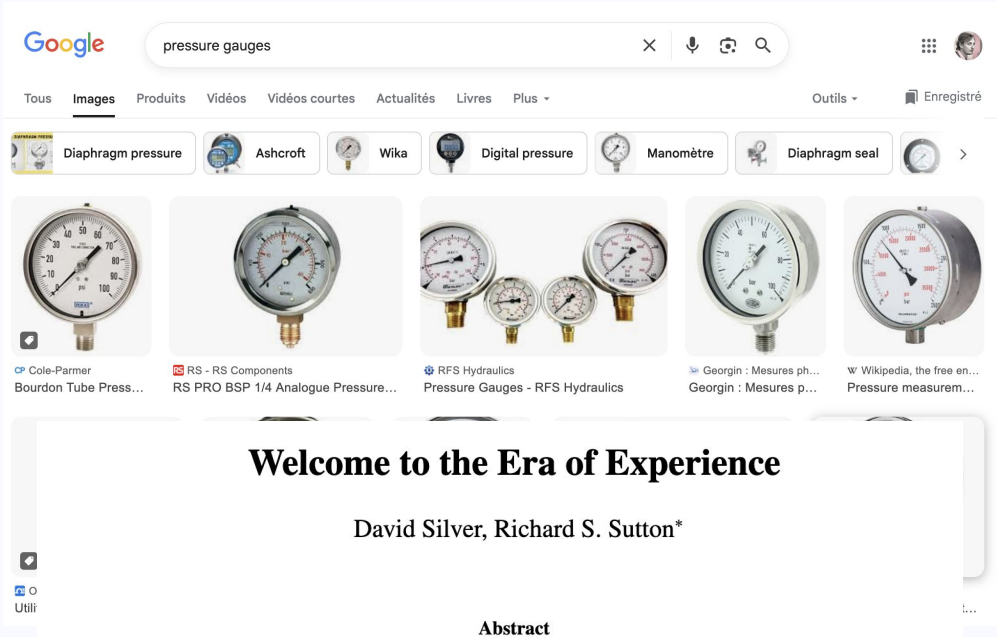
Most people don’t care (even regulators)

2. Paradigm shift? Switching to synthetic environments

Switching to environments

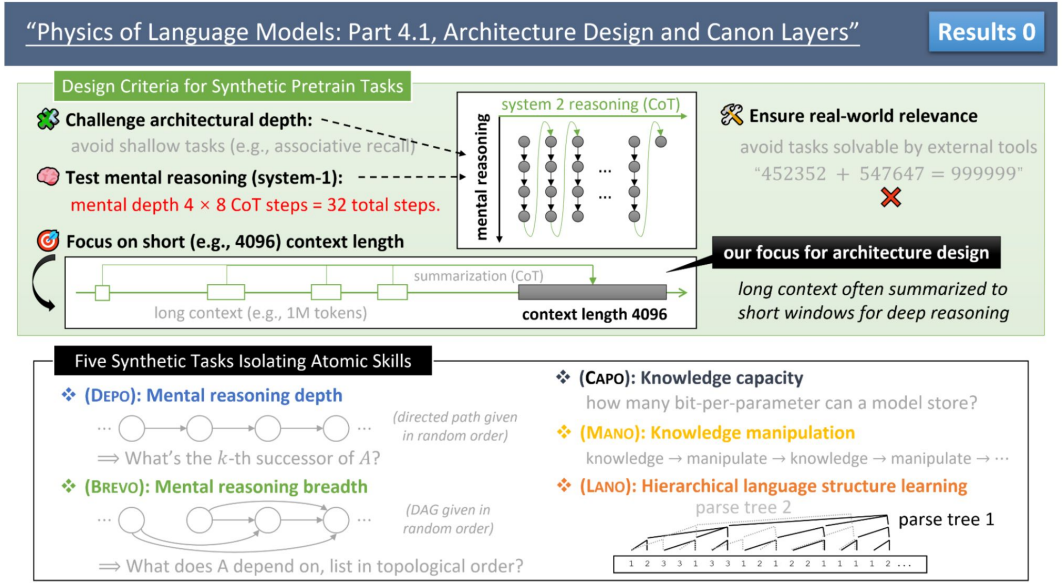
Over the last year, there has been a significant shift of paradigm of AI training with the development of reasoning models and the ascending role of synthetic and RL environments, to the point it is openly questioned if “pretraining as we know it will end”.

Beyond the concerns over the “data wall”, web data seems to hit a capability ceiling in many areas: vision languages models routinely fail to read clocks or gauges since most available images are product descriptions.



Switching to controlled environments

In Frontier labs and, increasingly openly documented research, large pretraining dataset are being completed if not replaced by *synthetic environment* or *synthetic playgrounds*. A primary motivation has been **increasing data efficiency** and focus training on the acquisition of targeted skills.



"*We design synthetic tasks to systematically evaluate specific capabilities of language model architectures under controlled conditions, minimizing confounds and enabling clean comparisons*" (Physics of Language Model, 4.1)

"*Future pipelines may need to unify pre-/mid-/post-training: injecting reasoning data earlier and more continuously.*" (Physics of Language Model, 4.2)

Switching to environments

Environments were initially pioneered in Math, as synthetic data generation can be controlled by *formal checks* or *logical compilers*. This requires the existence of a dynamic OSS ecosystem for formalization: this is the case in Math with Lean or Coq, but not in many other domains (physics, linguistics, even poetry). The lack of centralized open formulas comparable to Mathematica (Wikifunctions?) is a significant hurdle

4.6 Synthetic Datasets

Despite being among the largest formal mathematics libraries, the Metamath library remains scarce in the context of deep learning, especially in light of the advantages demonstrated on various NLP tasks by pre-training on large corpora. Also *set.mm* mostly focuses on well-known high-level theorems and does not include a large number of technical lemmas resembling the type of mathematics exercises used as curriculum for humans. Finally, Metamath lacking high level tactics such as HOL Light's `ARITH_RULE`[7], or Lean's `ring`[8], it is critical to ensure that our models are capable of proving at least basic technical theorems generally handled by high-level tactics in other systems (in domains such as arithmetic or ring equalities and inequalities)

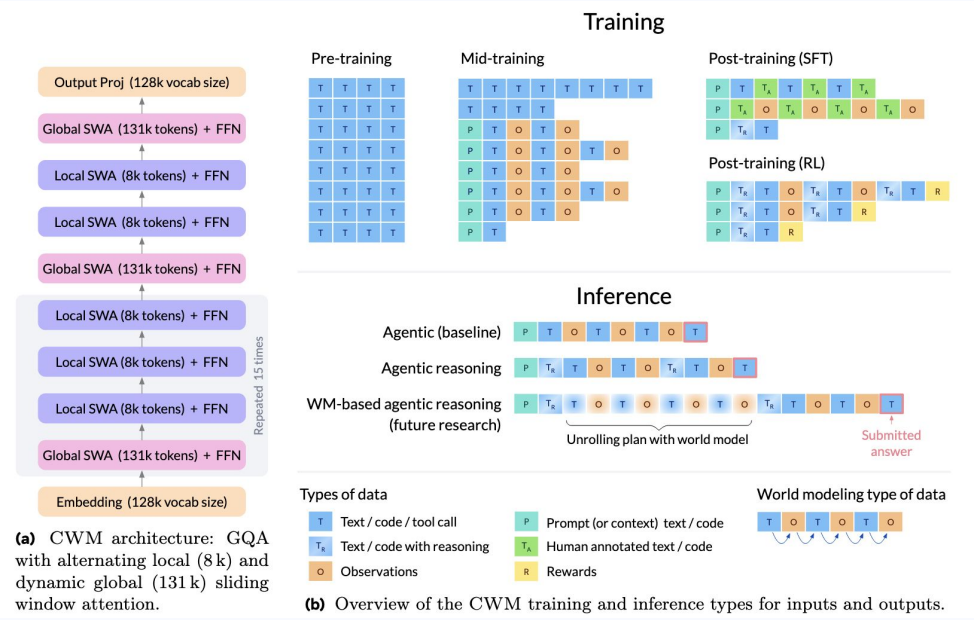
To achieve this goal we designed synthetic datasets allowing us to generate proofs for each of these domains at will while controlling precisely by how many proofs we augment our training set.

We describe below the synthetic datasets we designed and report in section 5 the sample complexity associated with these synthetic tasks.

As soon as 2020 (!) GPT-F from OpenAI is an early math prover exclusively trained on synthetic data with formal checks.

Switching to environments

Code is currently the main use case. Straight execution of code in controlled settings (dockers) with emulated data allow to uncover all kinds of poorly documented issues and features. Basically translating lived experiences into text that were lacking in training until now.



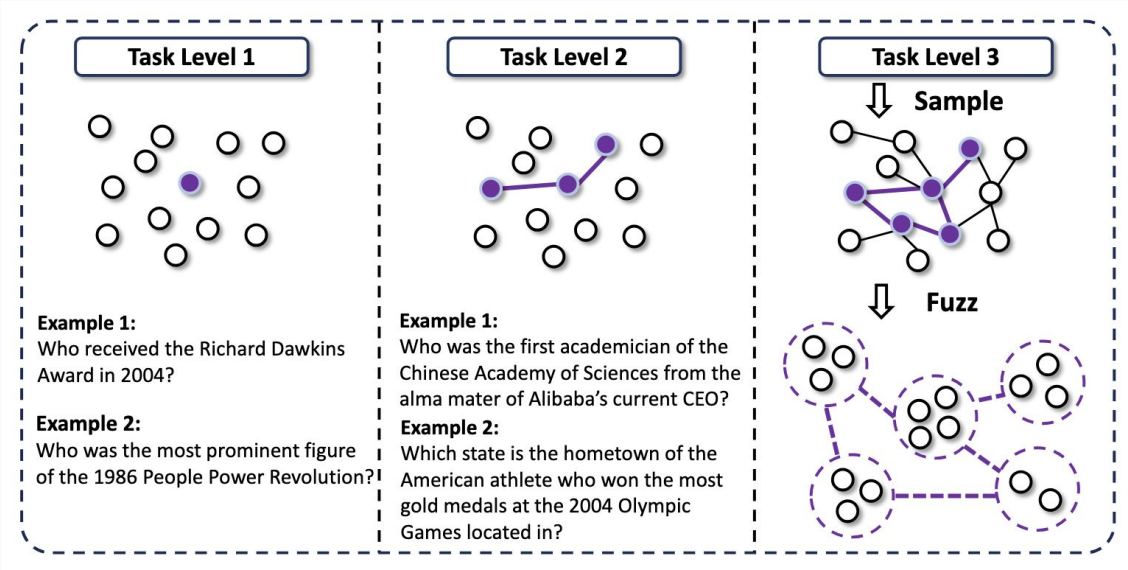
(a) CWM architecture: GQA with alternating local (8 k) and dynamic global (131 k) sliding window attention.

(b) Overview of the CWM training and inference types for inputs and outputs.

Meta's Code World Model likely relied on a large amount of well documented open source code to build up emulated programming environments.

Switching to environments

Prior to SYNTH the few openly documented synthetic environments, show actually an increasing needs for highly documented datasets **in the open**: since data can be indefinitely amplified, scale becomes less relevant than data quality, interconnection and extensive documentation. All things can that cannot happen with shadow datasets



Ali Baba Deep Research environment, WebSailor is based on the core of Semantic Web: Wikidata.

A shift in model design

Code and search were the first areas where we see a new wave of specialized agentic models, but won't be the only ones. As models are built as "effective agents" able to control their own workflow they re-integrate many features of deployment and become their own product.

The Model is the Product

There were a lot of speculation over the past years about what the next cycle of AI development could be. Agents? Reasoners? Actual multimodality?

I think it's time to call it: the model is the product.

All current factors in research and market development push in this direction.

- Generalist scaling is stalling. This was the whole message behind the release of GPT-4.5: capacities are growing linearly while compute costs are on a geometric curve. Even with all the efficiency gains in training and infrastructure of the past two years, OpenAI can't deploy this giant model with a remotely affordable pricing.
- Opinionated training is working *much* better than expected. The combination of reinforcement learning and reasoning means that models are suddenly learning tasks. It's not machine learning, it's not base model either, it's a secret third thing. It's even tiny models getting suddenly scary good at math. It's coding model no longer just generating code but managing an entire code base by themselves. It's Claude playing Pokemon with very poor contextual information and no dedicated training.
- Inference cost are in free fall. The recent optimizations from DeepSeek means that all the available GPUs could cover a demand of 10k tokens per day from a frontier model for... the entire earth population. There is nowhere this level of demand. The economics of selling tokens does not work anymore for model providers: they have to move higher up in the value chain.

This is also an uncomfortable direction. All investors have been betting on the application layer. In the next stage of AI evolution, the application layer is likely to be the first to be automated and disrupted.

A shift in model design

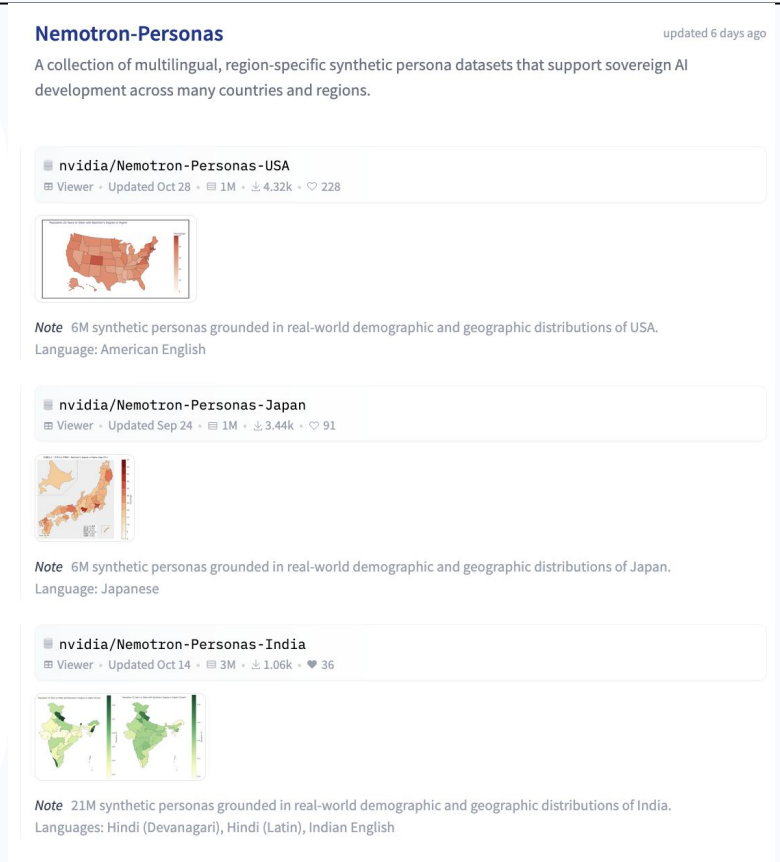
If the model is a product, can training data become a commodity?

3. Building synthetic environment in the open.

Synth is open by default?

While copyrighted pretraining data cannot be released publicly, the situation with synthetic data is very different:

- Copyright protects **original expression** and with well conceived original pipelines, content would be public domain by default (no author)
- Open datasets can be highly qualitative and extensively documented but to small. Synthetic allows for **indefinite data expansion** and make pretraining on small sources viable.
- Synthetic data generation allow to work on **sensitive data** by simulating personas and realistic documents



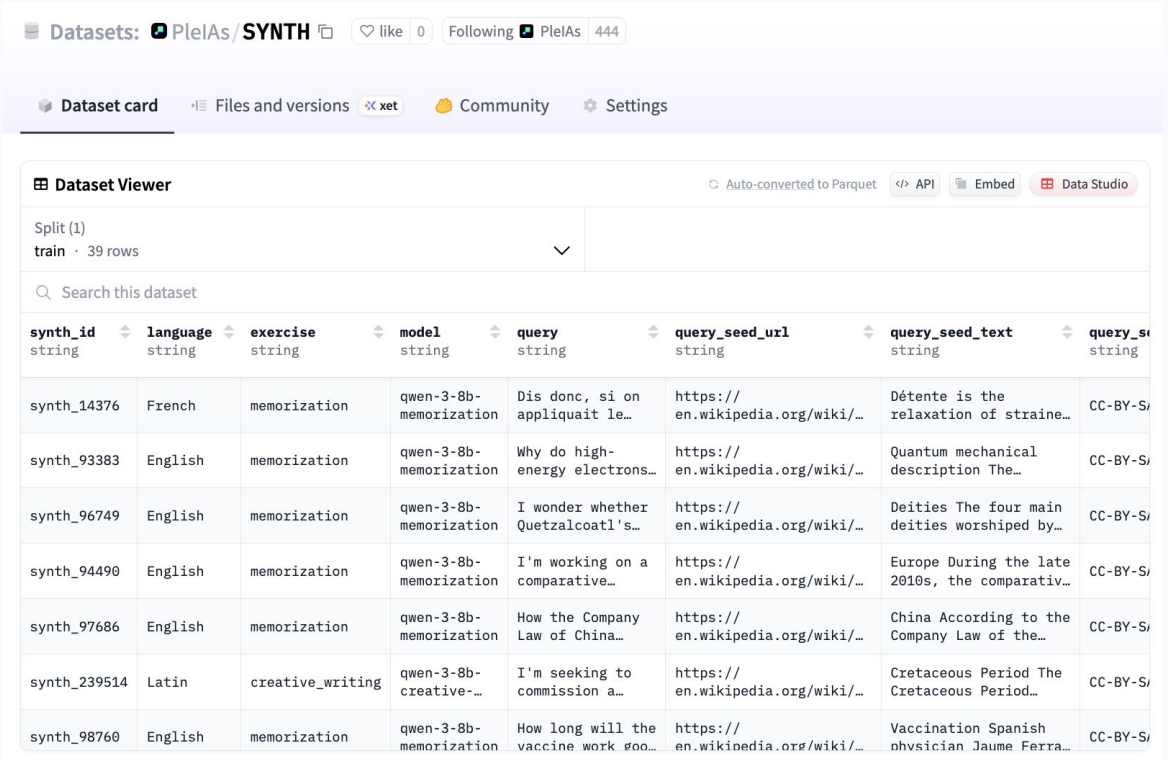
New incentives for openness

Beyond reducing liabilities for opennes, the synthetic turn also create new impetus for collaborations: specialized environments require specialized inputs as well as highly connected interoperable data to feed simulations.

Following on the examples set by Chinese labs like DeepSeek, even US industry leaders start open sourcing standards and intermediary artifacts like MCP.

Making environments open: SYNTH

Over the last few months, we have been working on a generalist synthetic pipeline to train more efficient small language models, thanks to the availability of high quality of open datasets and to the release of fully open weight models without restriction for data reuse



synth_id string	language string	exercise string	model string	query string	query_seed_url string	query_seed_text string	query_s string
synth_14376	French	memorization	qwen-3-8b-memorization	Dis donc, si on appliquait le…	https://en.wikipedia.org/wiki/…	Détente is the relaxation of straine…	CC-BY-S
synth_93383	English	memorization	qwen-3-8b-memorization	Why do high-energy electrons…	https://en.wikipedia.org/wiki/…	Quantum mechanical description The…	CC-BY-S
synth_96749	English	memorization	qwen-3-8b-memorization	I wonder whether Quetzalcoatl's…	https://en.wikipedia.org/wiki/…	Deities The four main deities worshiped by…	CC-BY-S
synth_94490	English	memorization	qwen-3-8b-memorization	I'm working on a comparative…	https://en.wikipedia.org/wiki/…	Europe During the late 2010s, the comparativ…	CC-BY-S
synth_97686	English	memorization	qwen-3-8b-memorization	How the Company Law of China…	https://en.wikipedia.org/wiki/…	China According to the Company Law of the…	CC-BY-S
synth_239514	Latin	creative_writing	qwen-3-8b-creative-…	I'm seeking to commission a…	https://en.wikipedia.org/wiki/…	Cretaceous Period The Cretaceous Period…	CC-BY-S
synth_98760	English	memorization	qwen-3-8b-memorization	How long will the vaccine work goo…	https://en.wikipedia.org/wiki/…	Vaccination Spanish physician Jaume Ferra…	CC-BY-S

Amplifying high quality open data

Seeding is not just relevant for grounding synthetic data: it allows to indefinitely expand the original training sources so that they get better memorized in the final model. This process is called **upsampled rephrasing**.

For this we reused parts of our synthetic RAG pipelines: texts are *backtranslated* into queries, and then matched with more texts to create more knowledge connections.

Wikipedia:Vital articles/Level/5

8 languages

Project page | Talk | Read | Edit source | View history | Tools

From Wikipedia, the free encyclopedia

< Wikipedia:Vital articles

Overview | Level 1 | Level 2 | Level 3 | Level 4 | Level 5

Shortcuts
WP:VITAL5
WP:VA5

The five nested vital article Levels are meant to give direction to the **prioritization of improvements** of English Wikipedia articles (e.g. which articles to bring to WP:GA and WP:FA status), to provide a **measurement of quality** of overall English Wikipedia (e.g. what proportion of the most important articles are at GA and FA status), and to serve as a **centralized watchlist** of English Wikipedia's most important articles. Unlike the list of articles every Wikipedia should have, they are tailored to the English Wikipedia and are actively maintained by the dedicated WikiProject Vital Articles. **This page contains links to the 50,000 articles of the Level 5 list.**

Any addition to or removal from these lists should ONLY BE MADE after a discussion on the relevant Level 5 sub talk pages.

Level 1 (10 articles) < Level 2 (100 articles) < Level 3 (1,000 articles) < Level 4 (10,000 articles) < **Level 5** (50,000 articles)

Level 5 sub-lists [edit source]

Because of its size, Vital articles Level 5 has been split into several sub-lists. If you spot a duplicate listing, please remove one of them; if you aren't sure in which section a topic belongs, please initiate a discussion on the talk page. Please do not duplicate items on the same level of the list.

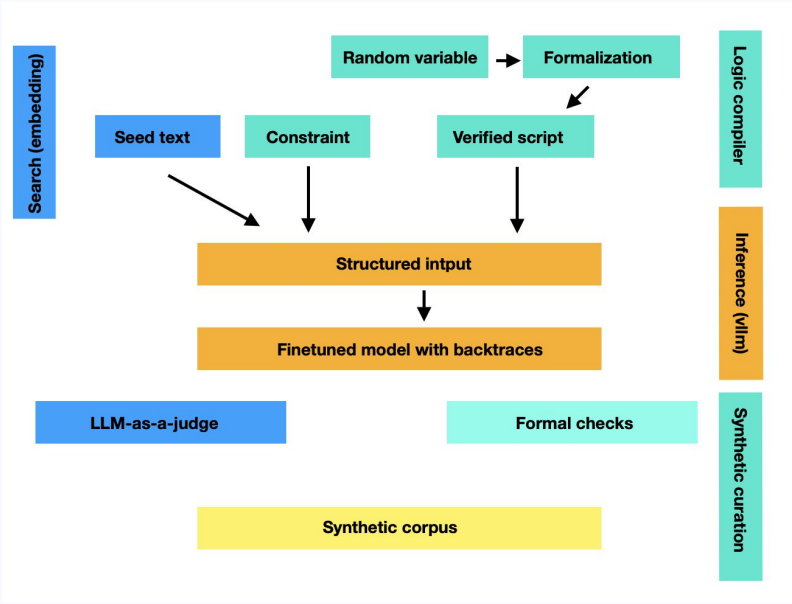
Overview | Level 1 | Level 2 | Level 3 | Level 4 | Level 5

Level 5 Sublists

Our nearly unique source of knowledge: the 50,000 articles Wikipedia:Vital Articles.

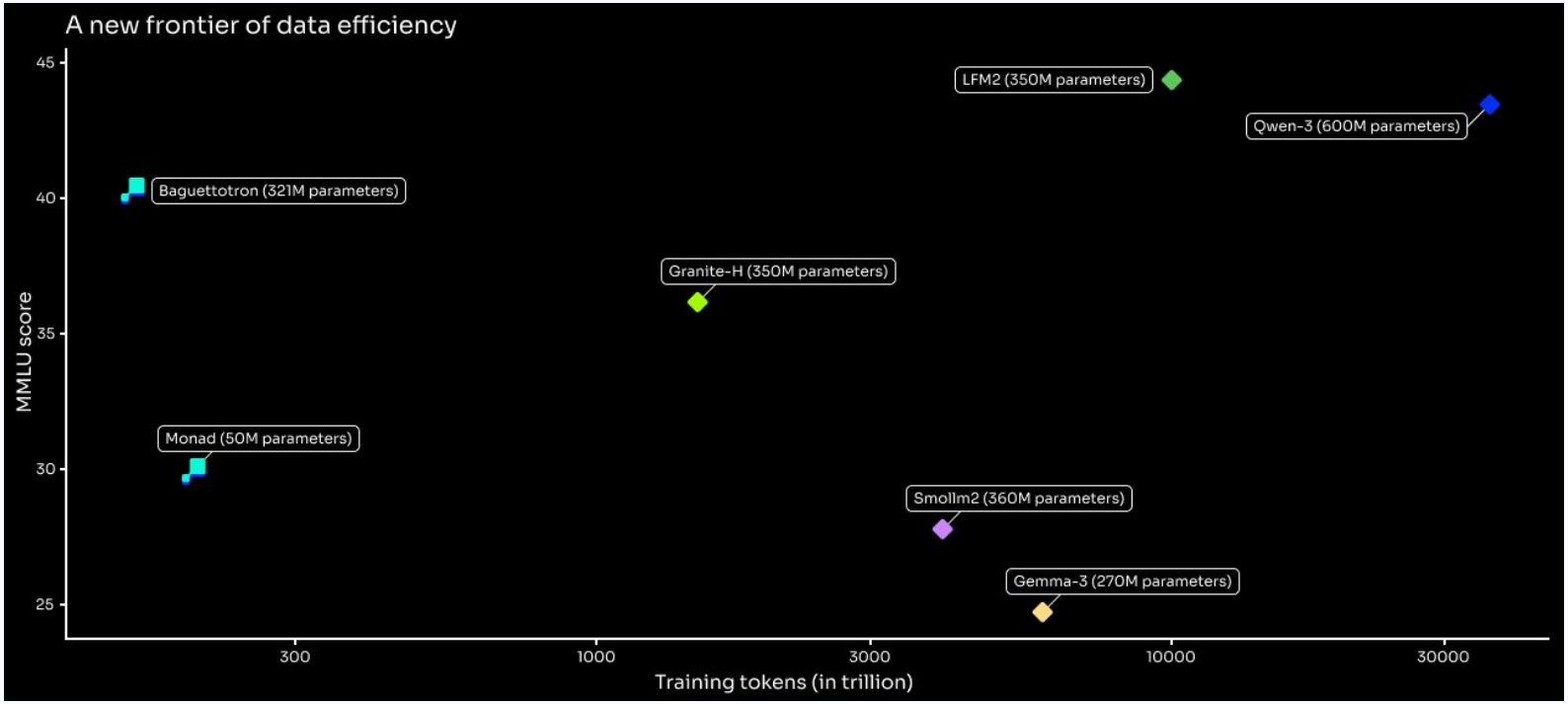
An engineering challenge

For SYNTH we generated about 100 billion tokens. While data efficiency meant we did not have to run inference at pretraining scale we had to make some key design choices to fit the project into our current compute plan: wide numbers of short slurm jobs (better for allocation/debugging if failure) and distillation of synth methods with smaller finetuned models.



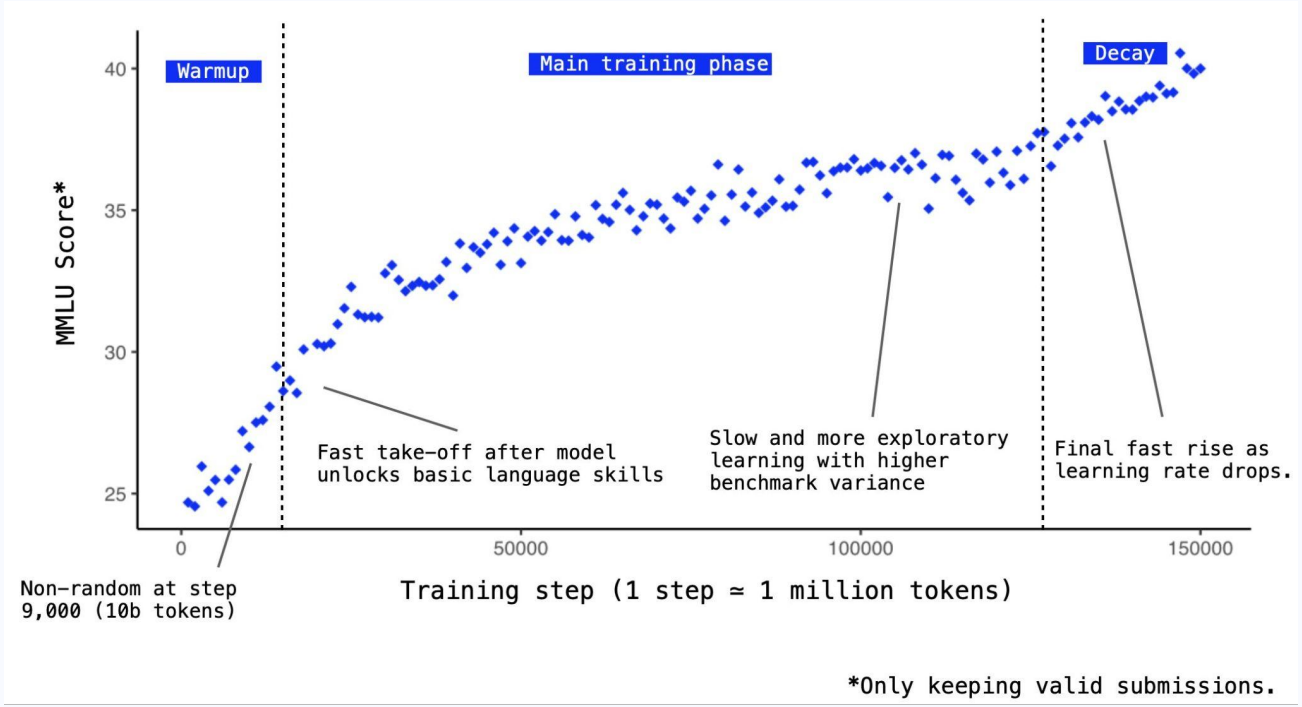
About half of synthetic data was generated in... three days. Due to an accidental data loss we had to recreate the entire set from scratch and the memorisation part

Setting a data-efficient state of the art



The first SYNTH models: a SOTA reasoning model in the 300M range and an entirely new category of “smallest viable model”.

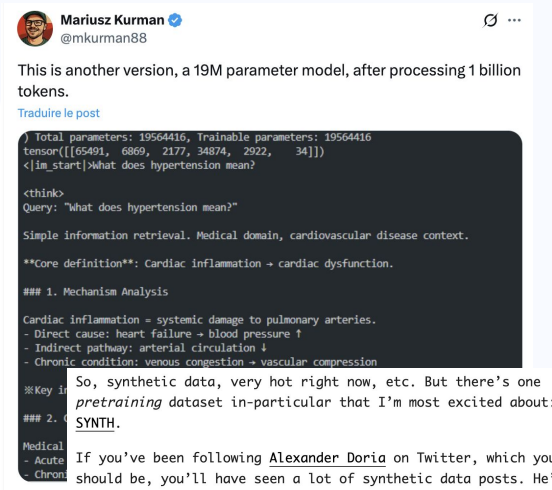
Setting a data-efficient state of the art



In contrast with classic pre-training datasets, reasoning signals appear very early in training. With less than 10B tokens, Baguettotron was non-random on MMLU.

Toward a new ecosystem of model training

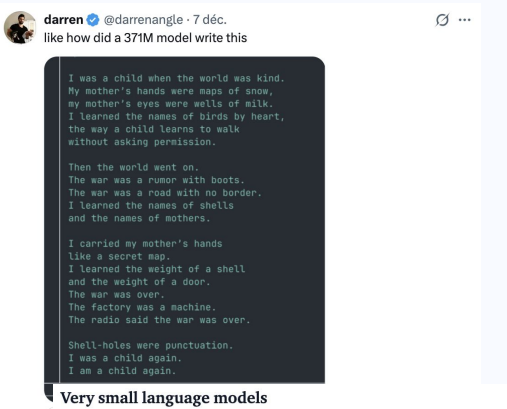
Although SYNTH was released less than one month ago, research community it has already led to a wide number of research experiments in the open community and essentially blurred the distinction between pre-/post- training research.



So, synthetic data, very hot right now, etc. But there's one *pretraining* dataset in-particular that I'm most excited about: SYNTH.

If you've been following Alexander Doria on Twitter, which you should be, you'll have seen a lot of synthetic data posts. He's the person I've followed who's most focused on taking Physics of Language Models seriously. One outcome from that school of thought is basically that you should be able to *engineer* data that will allow a model to efficiently learn the skills it needs to succeed in the inference environment you're interested in. You can also connect these ideas to the "Cognitive Core" Karpathy proposed: essentially finding the minimum number of parameters required to bake in basic world knowledge and the skills to search for what is needed at test-time "live". SYNTH is the result of pushing those ideas forward.

SYNTH starts with 50,000 Wikipedia articles and expands them into problem and resolution paths. Essentially turning those articles into the context for different skills being tested. To cut down on the inference costs of generating this data, relatively small open models (Qwen 8B, for example) are used to generate the data, with a complex harness around it to verify reasoning traces and improve



Very small language models

I also chose the English dataset because I wanted to test even smaller models. There are very few SLMs that are truly small enough to match encoder-only parameter sizes, which is important when you need inference via a single forward pass on CPU within a reasonable latency. I evaluated HuggingFaceTB/SmolLM2-360M-Instruct and ibm-granite/granite-4.0-350M. Additionally, Pleias AI achieved impressive results close to Qwen 0.6B on MMLU with Baguetotron (321M) and even more surprising performance with Monad (56M). These 2 last models underwent a single mode of training (no pretraining, mid training, post training paradigm, just training) on the SYNTH dataset and were not explicitly optimized for instruction following (mostly knowledge retrieval and creativity according to their blog). I still wanted to test them.

Model	Test Macro F1	Test Weighted F1
Baguetotron (3 epochs - 360M)	0.83096	0.88265
SmollLM2-instruct (3 epochs - 321M)	0.875	0.924
Granite (3 epochs - 350M)	0.76	0.77
Monad (5 epochs - 56M)	0.72865	0.86547
E5 small (3 epochs - 100M encoder only with MLP)	0.8646	0.9082

Toward a new ecosystem of model training

Since SYNTH was also highly data efficient on the input side, we are now replicating the methods in specialized regulated sectors with rich yet limited or even restricted data sources: healthcare, transportation, insurance, education...

SPINEDAO & PLEIAS PARTNERSHIP IN SPECIALISED AI FOR SPINE CARE

DEC 8, 2025

SpineDAO & Pleias are partnering to develop safe AI for wellness and future clinical deployment starting with back pain, the #1 cause of disability worldwide.

SpineDAO, the research collective of 200+ spine clinicians, scientists and engineers is joining forces with Pleias, the AI organisation, to build multi-agent reasoning systems for spine wellness. This collaboration is the first step towards solving the bottleneck of AI's hardest deployment challenge: clinical intelligence that scales without compromising safety for wellness and healthcare. SpineDAO brings the clinical expertise; Pleias brings the language model architecture and reasoning-first AI infrastructure.

Together we are tackling the scaling crisis in wellness and health AI by studying and researching in a very expert focus specialisation: the spine.

- In Back Pain, the #1 cause of disability worldwide the expert clinical judgment changes outcomes dramatically, but there aren't enough spine specialists and there never will be.
- Meanwhile, generic LLMs can scale but they're fundamentally unsafe for clinical deployment as they do not embed specific reasoning systems for safe, and efficient, clinical judgment.

The challenge consists into understanding one of the biggest AI bottlenecks of today not as a compute problem nor data quantity problem but as a *reasoning architecture* **problem.**

Conclusion