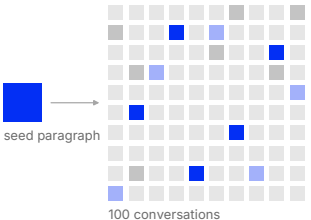
pleias

Paris · pleias.ai

SynthCommonCorpus

Building the largest open synthetic pretraining corpus for agentic models



Dr. ing. Pieter Delobelle, Lead AI scientist @ Pleias, Guest professor @ KU Leuven · GOSIM 2026 · 16 October 2026

SYNTH V1

We start from SYNTH

Accepted at NeurIPS 2026 as “It’s All Training”



arXiv:2609.37891v1 [cs.CL] 29 Sep 2026

It's All Training: A Fully Synthetic Single-Stage Recipe for LLMs

Pierre-Carl Langlais[1,2,3] Pierre Delobelle[1,4] Yannick Dotreb[1,5]
Pavel Chizhov[1,6] Carlos Rosas Hinostroza[1,7] Nell St Small[1]
Benjamin Burtke[1] Hanna Shcharbakova[1]
Ivan Yamshchikov[1,3] Anastasia Stasenko[1,2]

Pleias

[2]Softcome Center for Artificial Intelligence [3]Science Po Médialab [4]EPFL
[5]CARO, Technical University of Applied Sciences Würzburg-Schweinfurt
[6]Paris Dauphine-PSL [7]Lattice, ENS-PSL
[8]TU Munich, Munich Center for Machine Learning [9]KU Leuven
Correspondence: {pierre-carl, pierre}@pleias.ai

Abstract

Current pre-training datasets are derived from web crawls, with all their issues, and were not designed to support mid- and post-training pipelines — for instance, they contain little explicit reasoning. Thus, many frontier labs have begun to develop their own synthetic datasets, but the recipes behind them are largely undisclosed. Pre-training data mix, e.g., with reasoning traces to address cold-start problems. While demonstratively effective, none of these datasets are public, and the effect of this so-called *synthetic data on knowledge and skill acquisition of language models*, including small ones, remains poorly understood. We present SYNTH, the first fully synthetic pre-training dataset, built entirely with open models and open methods: pre-, mid-, and post-training into a single training stage via structured amplification of curated encyclopedic seeds. We evaluate SYNTH by training a suite of models: a 500M tiny model (Monad), 0.3B–0.8B dense models (Baguettotron, Nanbeige), and a 10B-total / 1B-active Mixture-of-Experts. At size scorpion, SYNTH outperforms all open-data baselines we tested, and matches or exceeds models trained on open-weight baselines. Because SYNTH is fully-simulated from grounded passages, SYNTH-trained models achieve high factual precision despite 35–140× fewer training tokens than models of the same size. Our results suggest that fully synthetic datasets, including our SYNTH dataset, are capable of producing competitive generalist models from a fraction of the training data, enabling rapid iteration at the frontier advances. These findings open up possibilities for both generalist models with significantly increased data efficiency, as well as domain-specific models for which sufficient training data is unavailable. We release the publicly release our SYNTH dataset and the suite of BALLETTOKEN models under a permissive license, thus supporting open-source language model development.

pleias/synth | pleias/baguettotron

1 Introduction

Curated data has been the most prominent data source for Large Language Model (LLM) pre-training since the release of The Pile [Gao et al., 2020] by Common Crawl. However, this data allows for little control over content. Furthermore, in response to AI scraping, some rule owners have taken countermeasures to prevent their content from being included in pre-training datasets. As such,

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

80B

synthetic tokens in 8 languages

58,698

Wikipedia articles as seeds

10–140×

fewer training tokens than models of the same size

Langlais, Delobelle et al., NeurIPS 2026. arXiv:2609.37891

pleias

SYNTH V1

We already released models trained on SYNTH

Four open models, from 56M parameters to a sparse 13B flagship

SYNTH

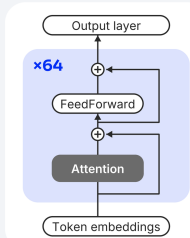
Fully synthetic, open dataset, 8 languages

PleIAs/SYNTH

Monad

56M · SMALLEST VIABLE

PleIAs/Monad

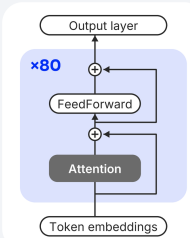


64 layers · hidden 256

Baguettotron-350M

350M · DENSE · 80 LAYERS

PleIAs/Baguettotron

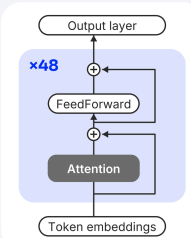


80 layers · hidden 576

Baguettotron-600M

600M · DENSE

PleIAs/Baguettotron-600M

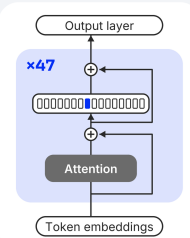


48 layers · hidden 1,024

Baguettotron-MoE

13BA1B · SPARSE FLAGSHIP

PleIAs/Baguettotron-MoE



47 layers · hidden 1,280

pleias

The model still learns in the second epoch

Baguettotron-600M saw SYNTH approximately two times: 158B tokens



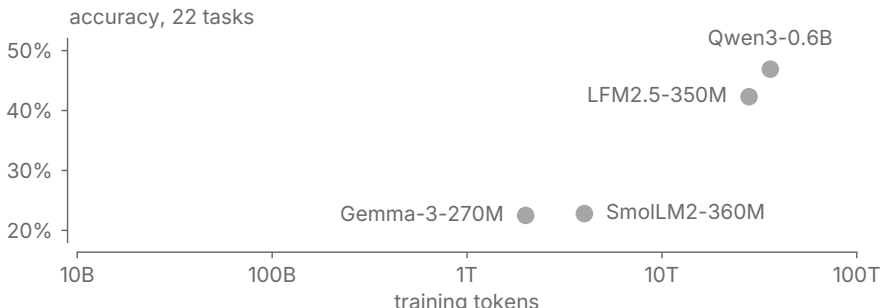
With synthetic data, we do not have to repeat the data. We can make new data.

Langlais, Delobelle et al., NeurIPS 2026. arXiv:2609.37891

SYNTH V1

Close to Qwen3 with 230 times fewer tokens

Both models have 600M parameters



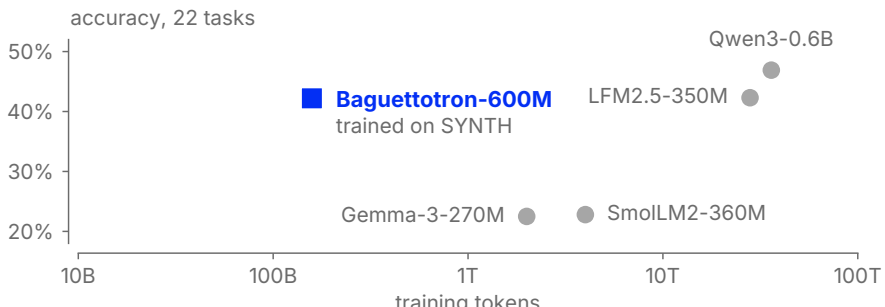
Models trained on web data use 2 to 36 trillion tokens.

Langlais, Delobelle et al., NeurIPS 2026, Figure 4. Baselines: pre-training tokens as reported.

pleias

Close to Qwen3 with 230 times fewer tokens

Both models have 600M parameters

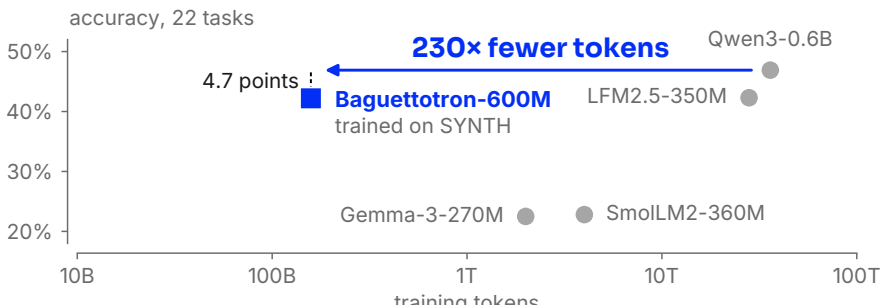


Baguettotron-600M used 158B tokens of SYNTH.

Langlais, Delobelle et al., NeurIPS 2026, Figure 4. Baselines: pre-training tokens as reported.

Close to Qwen3 with 230 times fewer tokens

Both models have 600M parameters

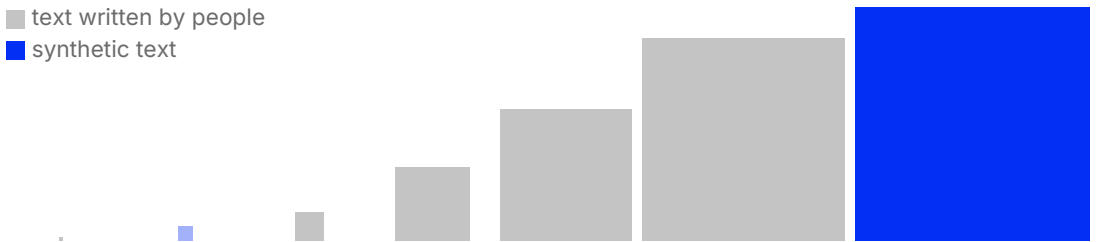


We shape synthetic data for training. Then each token teaches more.

Langlais, Delobelle et al., NeurIPS 2026, Figure 4. Baselines: pre-training tokens as reported.

The goal: 20 trillion synthetic tokens

250 times SYNTH v1, and more than FineWeb



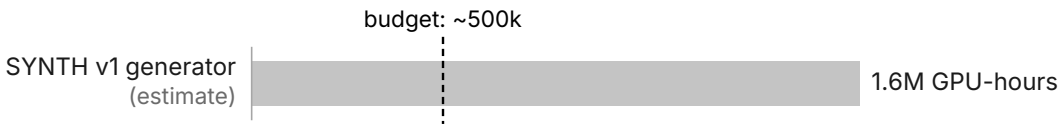
The area of a square shows the number of tokens. Each release counts with its own tokenizer.

pleias

THE PROBLEM

The SYNTH v1 generator is too slow for 20T

Each GPU must write approximately 11,000 tokens per second



We need a faster generator and a pipeline for hundreds of GPUs.

pleias

01

A small, fast generator

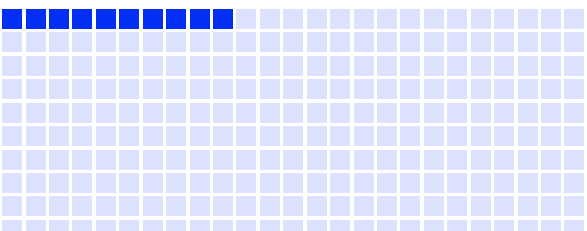
Large models teach, a small model writes

Teacher models wrote 13.2B tokens. Synthotron-8B writes the 20T.

13.2B ×1,500 **20T** tokens from Synthotron-8B



tokens from teacher models:
Gemma and DeepSeek-V4-Pro

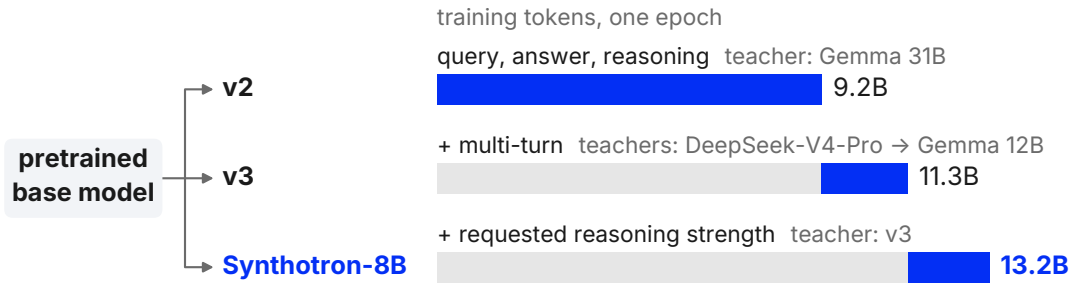


one square = 80B tokens blue: 830B already written

The teachers run one time. The small model writes all 20T tokens.

Synthotron-8B learned from three teachers

Three versions. Each version adds one type of training data.

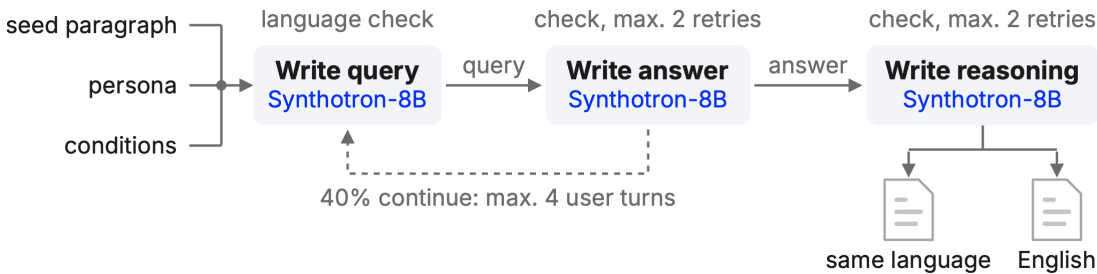


Each version trains again from the same base model, for one epoch.

Training: torchtitan, 32 H100 GPUs. Blue: new data in this version.

One model writes the full conversation

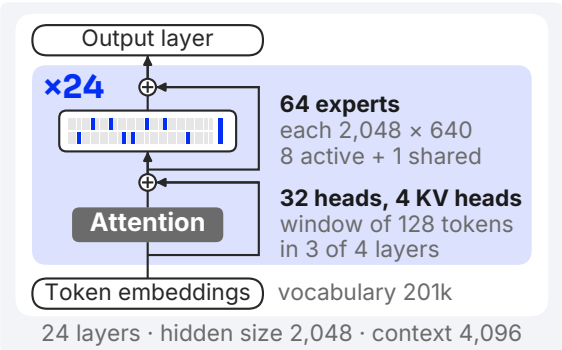
All facts come from one Wikipedia paragraph: the seed



One seed gives 100 conversations, with 100 different personas.

Small experts work best with many requests

Each step reads every expert one time. More requests share that read.

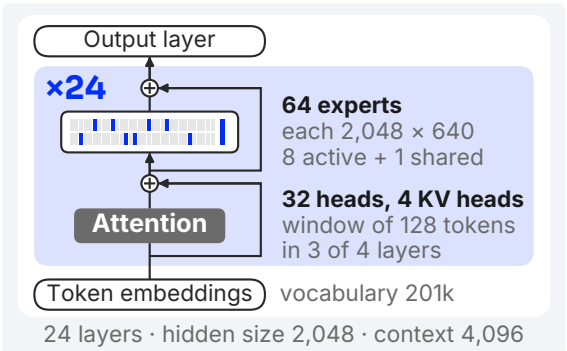


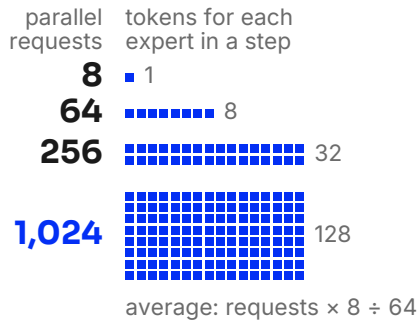
Each token uses 8 of 64 small experts: 1.5B of the 7.4B parameters.

Architecture: configuration of the production checkpoint.

Small experts work best with many requests

Each step reads every expert one time. More requests share that read.





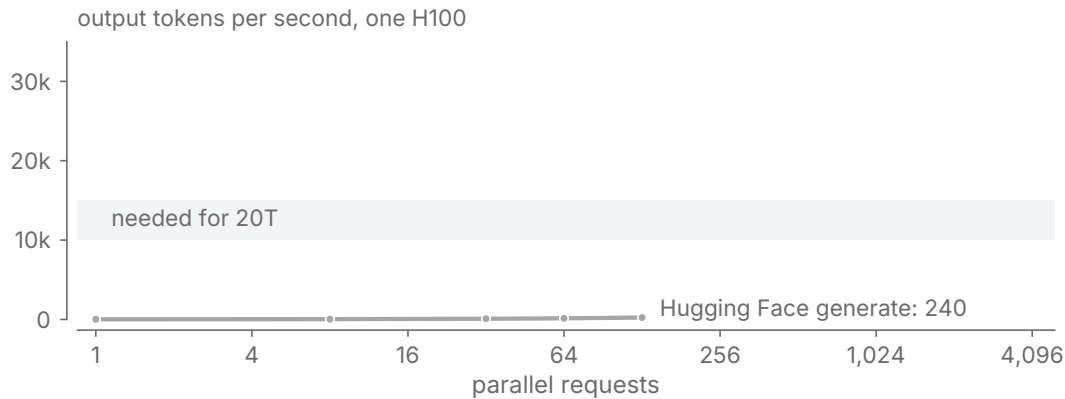
4 KV heads keep the cache small: one GPU keeps more than 500,000 tokens.

Architecture: configuration of the production checkpoint.

pleias

One H100 writes 19,000 tokens per second

The SYNTH v1 generator writes approximately 3,500 tokens per second

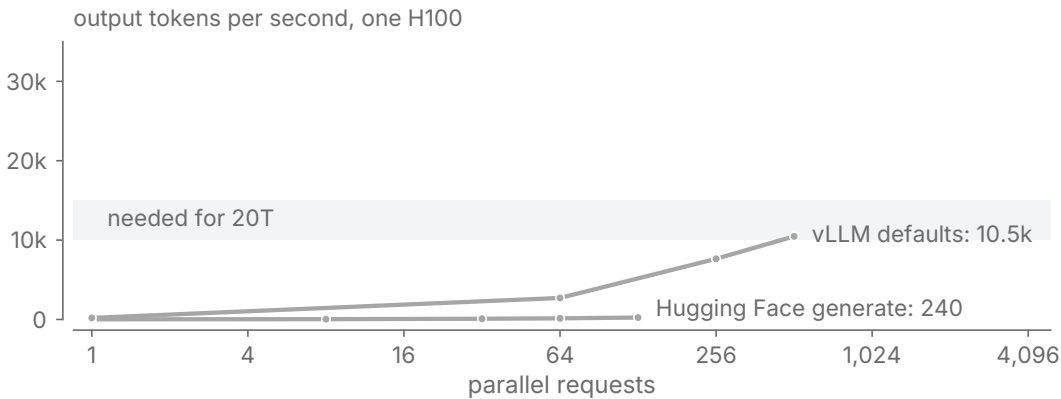


Synthotron-8B, vLLM 0.18 and our plugin. One MareNostrum 5 H100 (64 GB), query task.

pleias

One H100 writes 19,000 tokens per second

The SYNTH v1 generator writes approximately 3,500 tokens per second

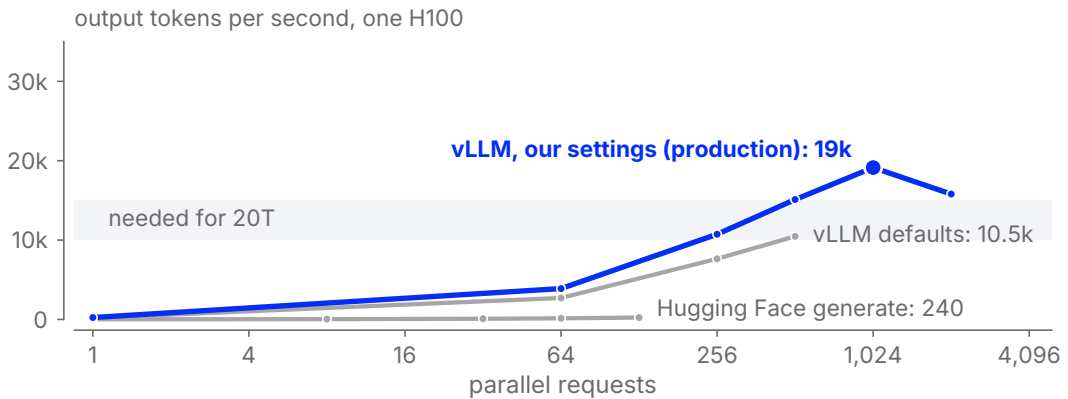


Synthotron-8B, vLLM 0.18 and our plugin. One MareNostrum 5 H100 (64 GB), query task.

pleias

One H100 writes 19,000 tokens per second

The SYNTH v1 generator writes approximately 3,500 tokens per second

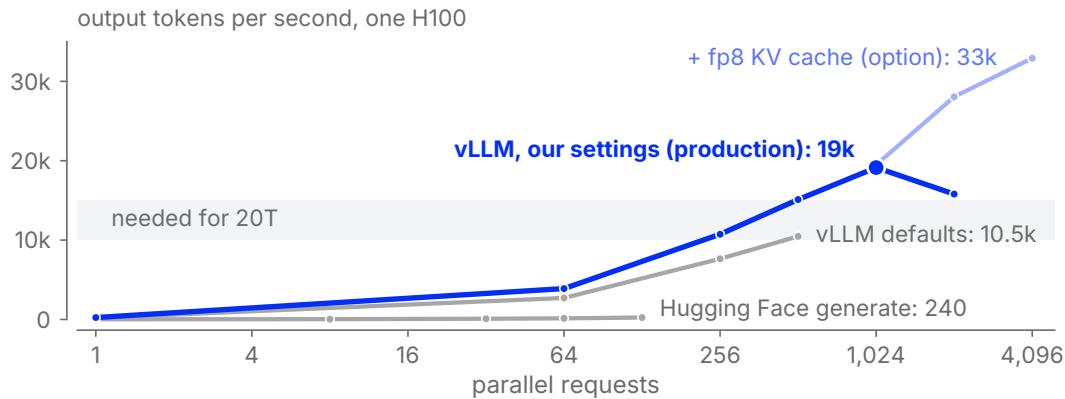


Synthotron-8B, vLLM 0.18 and our plugin. One MareNostrum 5 H100 (64 GB), query task.

pleias

One H100 writes 19,000 tokens per second

The SYNTH v1 generator writes approximately 3,500 tokens per second



Synthotron-8B, vLLM 0.18 and our plugin. One MareNostrum 5 H100 (64 GB), query task.

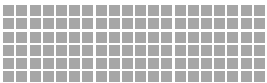
pleias

Fewer active parameters, faster generation

The dense SYNTH v1 generator and Synthotron-8B, each on one H100

parameters, 1 square = 100M

SYNTH v1 generator
Gemma-3-12B, dense



12B active
all parameters

Synthotron-8B
mixture of experts



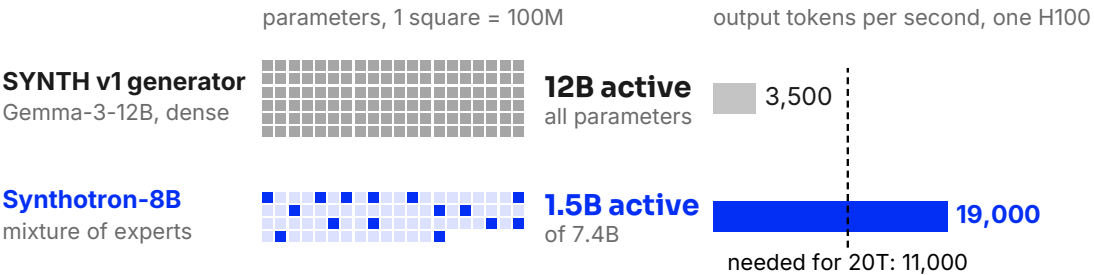
1.5B active
of 7.4B

Each token uses 1.5B parameters instead of 12B.

Different benchmarks. SYNTH v1 generator: H100 (80 GB). Synthotron-8B: H100 (64 GB), query task.

Fewer active parameters, faster generation

The dense SYNTH v1 generator and Synthotron-8B, each on one H100



Only Synthotron-8B writes the 11,000 tokens per second that 20T needs.

Different benchmarks. SYNTH v1 generator: H100 (80 GB). Synthotron-8B: H100 (64 GB), query task.

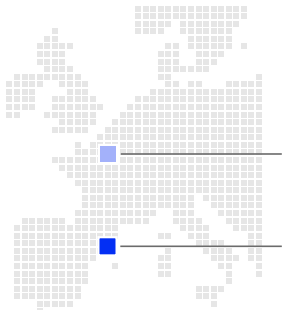
pleias

02

Scaling out

We use public European supercomputers

Shared clusters with SLURM, not a cloud



Jean Zay
GENCI · IDRIS, Paris area
1,456 H100 · **ready**

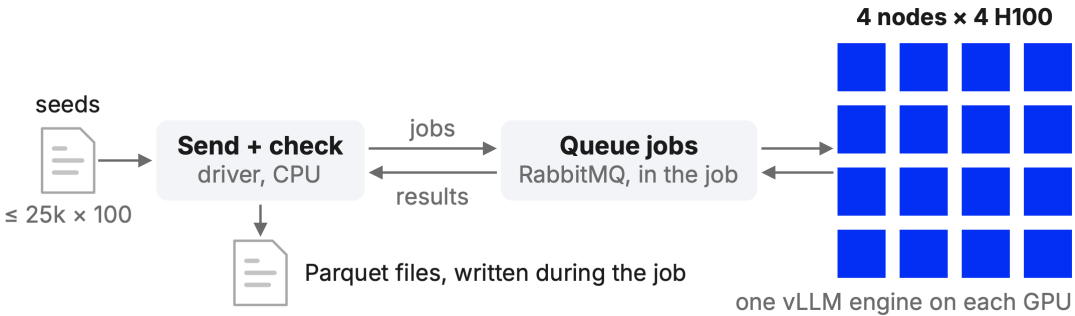
MareNostrum 5
EuroHPC · BSC, Barcelona
4,480 H100 · **production**

SLURM and nodes without internet make orchestration difficult.

GPU counts: BSC and CNRS/GENCI.

One job: one driver and 16 GPUs

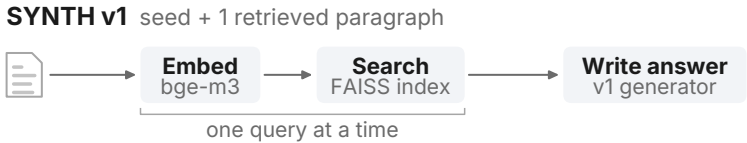
A driver sends the work to one vLLM engine on each GPU



One job makes approximately 2.4M conversations in 3.5 to 5 hours.

We removed retrieval from the pipeline

Seed only: a simpler and faster pipeline

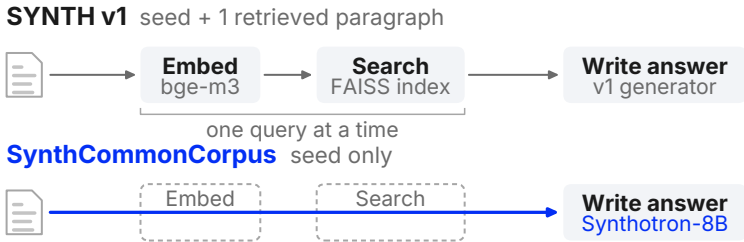


SYNTH v1 searched for one related paragraph for each query.

Fact attribution: 100 answers, judge gpt-oss-120b (August 2026). Speed: MareNostrum 5 (September 2026).

We removed retrieval from the pipeline

Seed only: a simpler and faster pipeline

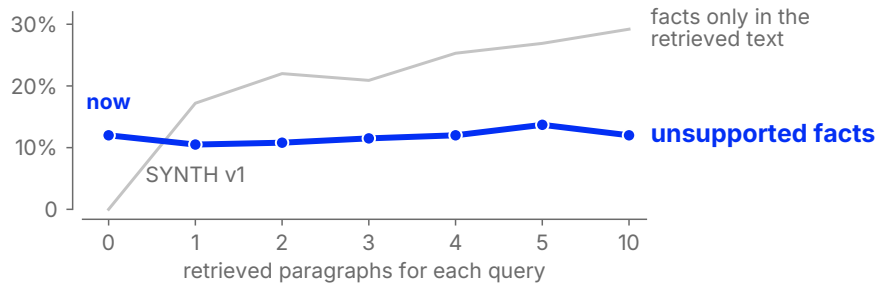


Without the serial search, a job is no longer limited to 48 pairs per second.

Fact attribution: 100 answers, judge gpt-oss-120b (August 2026). Speed: MareNostrum 5 (September 2026).

We removed retrieval from the pipeline

Seed only: a simpler and faster pipeline

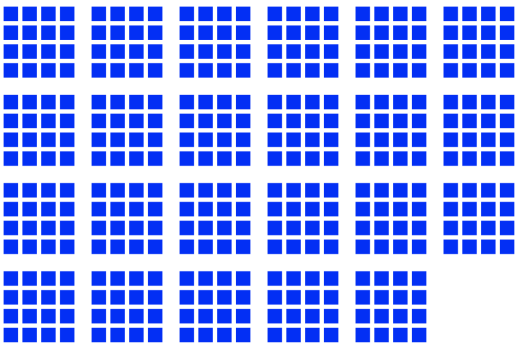


Unsupported facts stay at approximately 11–13%, with or without retrieval.

Fact attribution: 100 answers, judge gpt-oss-120b (August 2026). Speed: MareNostrum 5 (September 2026).

368 H100 GPUs at the same time

MareNostrum 5 (BSC). The speed increases linearly with the GPUs.



1 square = 1 GPU, 1 block = 1 job

12,700

question–answer pairs per second

~8M

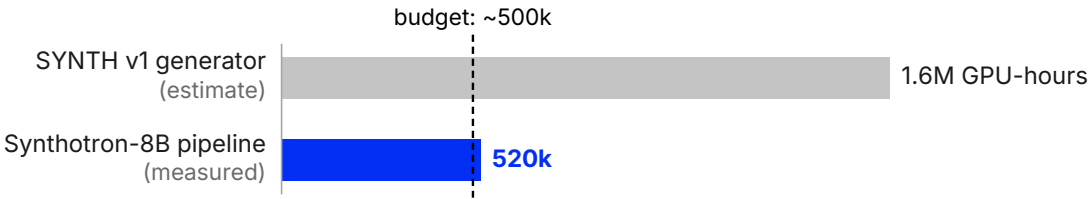
tokens per second

36

pairs per second for each GPU

20T needs approximately 520k GPU-hours

At the measured rate: 38M tokens per H100 GPU-hour



This is close to the budget. An fp8 KV cache can give more speed.

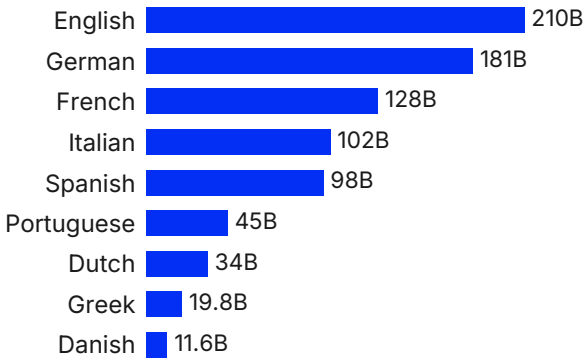
03

24 EU languages

pleias

Nine languages are done: 830B tokens

Every language has its own seeds, queries and reasoning



893.7M
conversations

21.7k
H100 GPU-hours

1.67 TB
of Parquet files

One conversation, two reasoning languages

A real Danish conversation from SynthCommonCorpus, with a Wikipedia seed

persona: a 10-year-old girl in Togo · seed: Wikipedia, H.C. Andersen

er det så rigtigt at han skrev det første bog i 17 år?

`<think>…</think>`

Ja, det er rigtigt. H.C. Andersen var 17 år gammel, da han udgav sin første bog, Ungdoms-Forsøg, i 1822. …

SynthCommonCorpus (synth_wiki_v5), Danish, row 63bdf370. Excerpts of `text` and `text_english_reasoning`.

One conversation, two reasoning languages

A real Danish conversation from SynthCommonCorpus, with a Wikipedia seed

Danish

```
…
→ "det første bog" = *Ungdoms-Forsøg*.
→ "17 år" = alder ved udgivelse.

Faktatjek:
● Bogen udgivet: 1822.
● Alders: 17 år.
☑ Ja, rigtigt.
…
```

SynthCommonCorpus (synth_wiki_v5), Danish, row 63bdf370. Excerpts of `text` and `text_english_reasoning`.

One conversation, two reasoning languages

A real Danish conversation from SynthCommonCorpus, with a Wikipedia seed

English

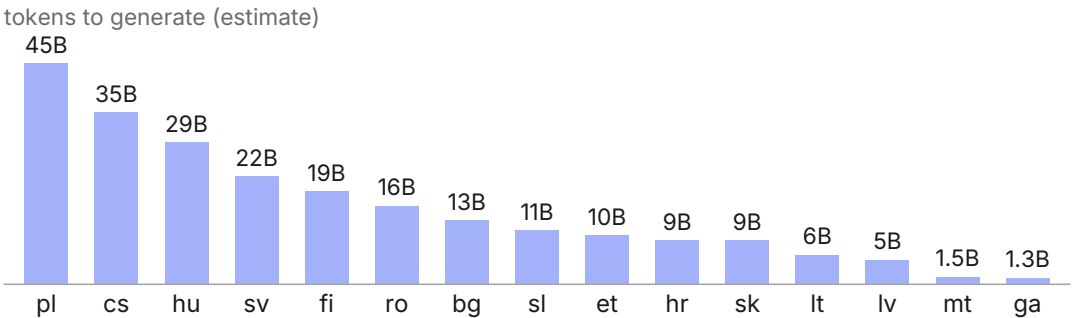
```
…
→ Translation: Is it true he wrote his first book at 17?
Task: Fact-check + historical context.

Entity: H.C. Andersen ●
Fact: First book "Ungdoms-Forsøg" @1822.
Age @1822: 17.
Premise ☑ True.
…
```

SynthCommonCorpus (synth_wiki_v5), Danish, row 63bdf370. Excerpts of `text` and `text_english_reasoning`.

15 more EU languages are next

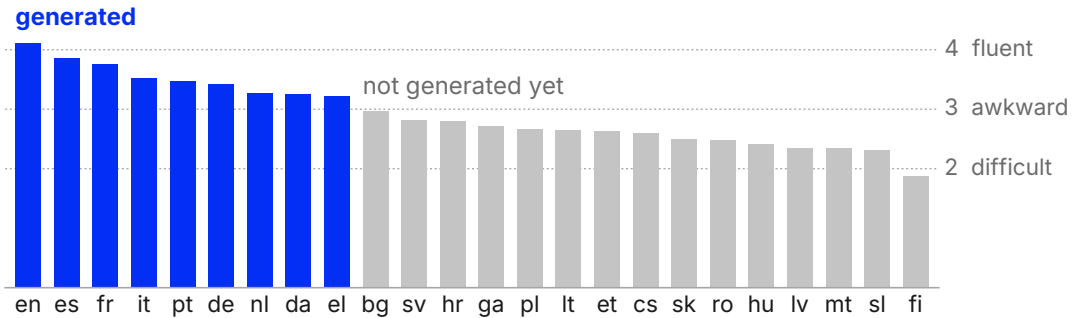
Approximately 232B tokens and 5.8k GPU-hours. The seeds are ready.



These values are estimates. For Danish and Greek, the real values were 5–16% higher.

Fluency follows the size of the language

Answer fluency (0–5) from one automated judge



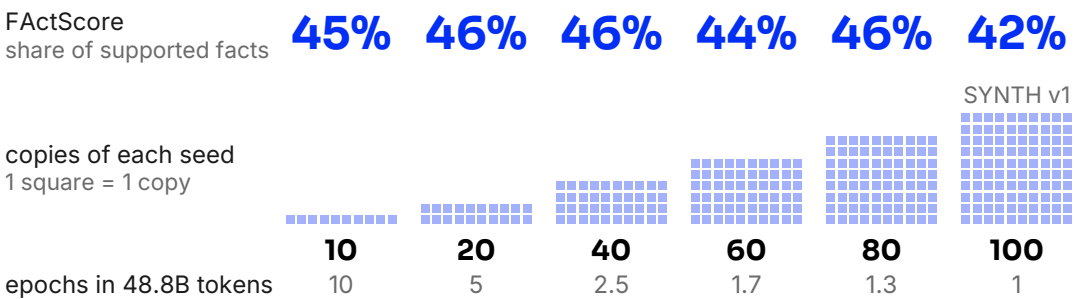
English is fluent. Most smaller EU languages are between 2 and 3.

04

From 830B to 20T

More copies of a seed do not teach more facts

Six 600M models. A copy is a different conversation about the same seed.



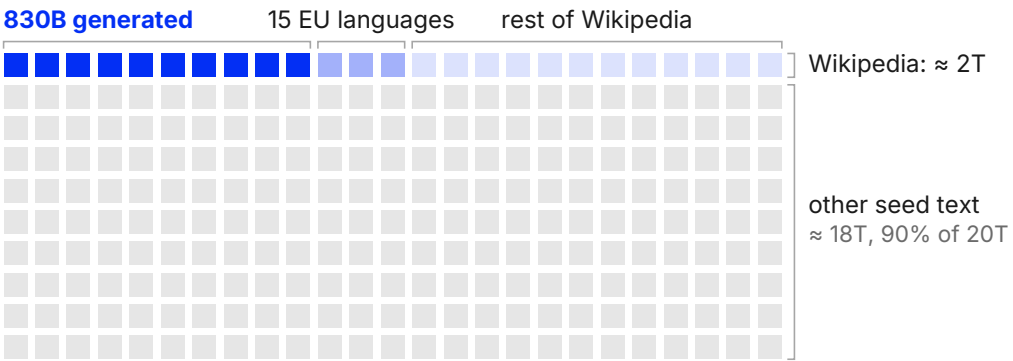
The facts come from the seeds. For more knowledge, we need more seeds.

Ablation after the paper (June 2026). FActScore on 1,000 seed topics, one run for each setting.

pleias

Wikipedia is only 10% of the target

One square is 80B tokens. All Wikipedia editions give approximately 2T.

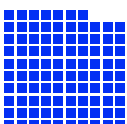


pleias

Common Corpus: licensed, traceable seeds

2.27T tokens of open text. Each document has a license and a source.

1 square = 10B tokens



OpenCulture
books, press
967B



OpenGovernment
legal, financial
579B



OpenSource
code
283B



OpenScience
academic
281B

OpenWeb
Wikipedia, Q&A
88B

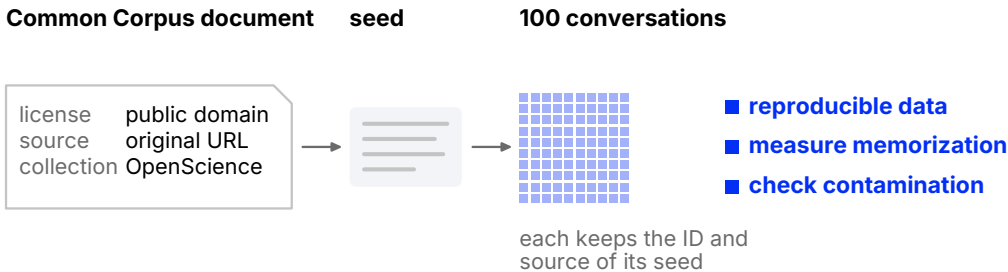
OpenSemantic
Wikidata
68B

The largest open licensed text dataset, made by Pleias.

Common Corpus, ICLR 2026 (oral). huggingface.co/datasets/PleIAs/common_corpus

Common Corpus: licensed, traceable seeds

2.27T tokens of open text. Each document has a license and a source.

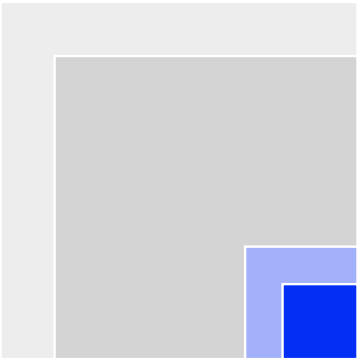


We can trace each training example back to a licensed seed.

Common Corpus, ICLR 2026 (oral). huggingface.co/datasets/PleIAs/common_corpus

Common Corpus has enough seed text

Tokens of seed text. The area of a square shows the quantity.



2.3T **Common Corpus**
open text, many languages

1.6T **kept by the classifier**
72% of the tokens

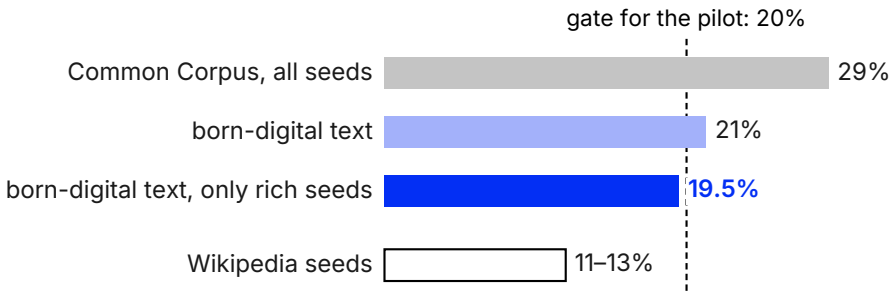
225B **first set**
born-digital and science text, only rich seeds

100B **needed for 20T**
with 100 conversations per seed

A small classifier gives each paragraph one of six labels. We keep only rich seeds.

Strict filters give fewer unsupported facts

Share of facts in the answers that the seed does not support



Seeds that are only a fragment of a text give up to 2 times more unsupported facts.

300 answers for each filter. Judge: gpt-oss-120b. August 2026.

First agentic data: single-turn tool calls

An early result: one request, one turn, no multi-step tasks yet

What is the etymological origin of the word 'drama', and what did Aristotle mean by it in his 'Poetics'?

2 tool calls, chosen from 3 tools in the system prompt

```
{"name": "Linguistics_1_GetEtymology", "arguments": {"word": "drama"}}
```

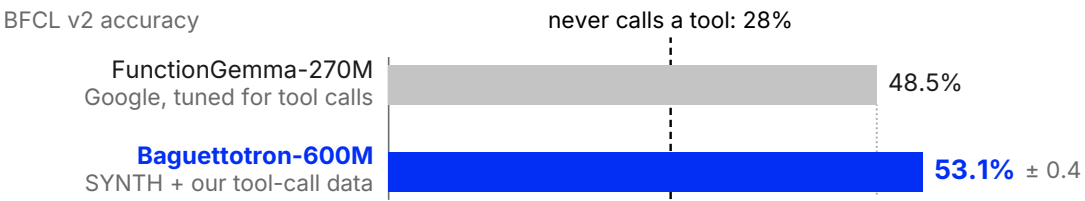
```
{"name": "Linguistics_1_GetDefinitionByContext",
 "arguments": {"word": "drama", "context": "Aristotle's Poetics"}}
```

Two Gemma-3-12B models wrote 532k samples from Wikipedia seeds.

Sample 352 of our tool-call corpus.

First agentic data: single-turn tool calls

An early result: one request, one turn, no multi-step tasks yet



With our data, Baguettotron-600M scores 4.6 points more than FunctionGemma.

BFCL v2, all 3,981 items, our scorer. Ours: mean of 3 training seeds. July 2026.

The next three steps to 20T

From 4% to 100% of 20T

1 seeds ready

More languages

15 EU languages:
232B tokens or more

2 classifier ready

More seed text

Filter the seeds from Common Corpus

3 planned

Longer reasoning

Long traces from a larger teacher model

The next three steps to 20T

From 4% to 100% of 20T

1 seeds ready

More languages

15 EU languages:
232B tokens or more

2 classifier ready

More seed text

Filter the seeds from
Common Corpus

3 planned

Longer reasoning

Long traces from a
larger teacher model

The next three steps to 20T

From 4% to 100% of 20T

1 seeds ready

More languages

15 EU languages:
232B tokens or more

2 classifier ready

More seed text

Filter the seeds from Common Corpus

3 planned

Longer reasoning

Long traces from a larger teacher model

The next three steps to 20T

From 4% to 100% of 20T

1 seeds ready

More languages

15 EU languages:
232B tokens or more

2 classifier ready

More seed text

Filter the seeds from Common Corpus

3 planned

Longer reasoning

Long traces from a larger teacher model

First release: November 2026. Follow @pleiasfr on X for the news.

pleias

Paris · pleias.ai

SynthCommonCorpus

Building the largest open synthetic pretraining corpus for agentic models

Slides: pieter.ai/talks/2026-10-16-gosim.html

Dr. ing. Pieter Delobelle, Lead AI scientist @ Pleias, Guest professor @ KU Leuven · GOSIM 2026 · 16 October 2026