pleias

Public Summary of Training Content for General-Purpose AI models: Pleias-Baguettotron, Pleias-Monad

Version of the Summary:	Version 1.0 — first publication of this Summary. No previous versions.
Last update:	04/08/2026

General information

1. General information

1.1. Provider identification

Provider name and contact details:	PLEIAS, société par actions simplifiée (SAS), share capital EUR 5,000. Registered with the Paris Trade and Companies Register (RCS Paris) under number 982 899 023. Intra-EU VAT number FR95 982 899 023. Registered office: 91 rue des Maraîchers, 75020 Paris, France. General contact: contact@pleias.fr Contact point for rightsholders and for questions concerning this Summary: contact@pleias.fr Website: https://pleias.ai Model repositories: https://huggingface.co/PleIAs
Authorised representative name and contact details:	Not applicable. PLEIAS is established in the Union (France). Article 54 AI Act, which concerns providers established in third countries, does not apply.

1.2. Model identification

Versioned model name(s):	This Summary covers the following two model versions. Their training content is identical. They are reported together in a single Summary in accordance with point (30) of the Commission Explanatory Notice to the Template, under which the same Summary may be used for different models or model versions where the content of their respective Summaries is identical. The model-specific figures given below — parameter counts and the number of passes made over the corpus — are disclosed voluntarily and do not affect the content of the training data, which is the same in each case. • **PleIAs/Baguettotron** — 320,956,992 parameters, 80 layers — https://huggingface.co/PleIAs/Baguettotron • **PleIAs/Monad** — 56,656,128 parameters, 64 layers — https://huggingface.co/PleIAs/Monad

	Both are decoder-only transformer text-generation models released under the Apache 2.0 licence, and both were trained exclusively on the same dataset, SYNTH. The model card for each is published at the URL given above. The two models differ in size, in context length (4,096 tokens for Baguettotron, 2,048 for Monad), in tokenizer (Baguettotron uses the PLEIAS byte-pair-encoding tokenizer with a vocabulary of 65,536 entries, of which approximately 45 are reserved as control tokens for reasoning; Monad uses a purpose-built English-only tokenizer with a vocabulary of 8,192 entries trained on SYNTH itself), and in the languages they are documented as supporting.
Model dependencies:	None. Both models were pre-trained from scratch by PLEIAS. Neither is a modification, fine-tune, distillation or other derivative of a pre-existing general-purpose AI model, and neither declares a base model. Neither model underwent a separate supervised fine-tuning, instruction-tuning or reinforcement-learning stage. Instruction-following and reasoning behaviour was acquired during pre-training, because the training corpus consists of instruction and reasoning material throughout. The training data was itself produced with the assistance of third-party general-purpose AI models; those models are identified in Section 2.5. They are generators of the training data, not components of, or dependencies for, the models covered by this Summary.
Date of placement of the model on the Union market:	• Baguettotron — repository created and weights published 10 November 2025; model card published 10 November 2025. • Monad — repository created and weights published 10 November 2025; model card published 10 November 2025. Both models were publicly announced on 10 November 2025. Neither has been further trained since its release, and no new version has been placed on the market. A revision to Baguettotron's tokenizer files on 27 April 2026 did not alter the model weights or the training data.

1.3 Modalities, overall training data size and other characteristics

This Section requires general information about the overall training data after pre-processing and before the training of the model.

Modality *Select the modalities present in the training data, to the extent that they are identifiable*	**Training data size** *For each selected modality, select the range within which the estimated total training data size for that modality falls. Dynamic datasets may be excluded from the estimation.*	**Types of content** *For each selected modality, provide a general description of the type of content that has been included in the training data.*
☒ Text	☐ Less than 1 billion tokens ☒ 1billion to 10 trillions tokens ☐ More than 10 trillions tokens Precise figures, disclosed voluntarily: • Training dataset: 79,648,272 text samples, comprising over 41	Synthetic instructional and reasoning text, generated in its entirety by AI models from a fixed set of openly licensed encyclopaedic seed documents. No text was taken from the open web, and no naturally occurring text other than the seed

	billion words, equivalent to approximately 75 billion tokens with the PLEIAS tokenizer. • Tokens processed per model: approximately 200 billion for each of Baguettotron and Monad, as stated on the model cards. For Baguettotron this corresponds to between two and three passes over the dataset. Monad was trained with a different tokenizer, so the dataset's tokenised length — and therefore the number of passes — differs.	passages quoted within the generated material is present. Distribution of the dataset by exercise type, measured over the released dataset: • memorisation — question-and-answer pairs with reasoning traces, grounded in an encyclopaedic passage — 90.6%; • multiple-choice questions with distractors — 2.1%; • constrained writing exercises — 1.7%; • mathematical exercises with symbolic solutions — 1.6%; • retrieval-augmented-generation exercises with cited sources — 1.0%; • mathematical multiple-choice questions — 1.0%; • creative writing, including lipograms, layout poems and style or persona exercises — 0.8%; • editing exercises, covering translation, structured extraction, orthographic and factual correction and style reformulation — 0.8%; • practical-knowledge exercises based on cooking recipes — 0.5%. Each record carries the question, the seed passage from which it was derived, the URL and licence of that seed, the constraints applied, the generated reasoning trace and the generated answer. Source code was deliberately excluded from the dataset. No image, audio or video content was used. No press publications, fiction or other material under active copyright protection were used.
☐ Image	☐ Less than 1 million images ☐ 1Million to1 billion images ☐ More than 1 billion images	*Examples of possible types of content include photography, visual art works, infographics, social media images, logos, brands.*
☐ Audio[1]	☐ Less than 10 000 hours ☐ 10 000 to1 million hours ☐ More than 1 million hours	*Examples of possible types of content include musical compositions and recordings, audiobooks, radio shows and podcasts, private audio communication.*
☐ Video	☐ Less than 10 000 hours ☐ 10 000 to1 million hours ☐ More than 1 million hours	*Examples of possible types of content include music videos, films, TV programmes, performances, video games, video clips, journalistic videos, social media videos.*
☐ Other	Not applicable. The training data comprises text only.	

Latest date of data acquisition/collection for model training:	11/2025. The dataset was generated and frozen before pre-training

began. The underlying encyclopaedic seed material was taken from a Wikipedia snapshot preceding that date. Neither model is continuously trained on new or dynamic data after that date, and neither has been further trained since release.

Description of the linguistic characteristics of the overall training data:	The dataset is multilingual, with approximately 20% of samples in languages other than English. Distribution by language, measured over the released dataset: English 80.9%; German 3.16%; Spanish 3.16%; French 3.15%; Polish 3.15%; Italian 3.14%; Dutch 1.60%; Latin 1.60%. The non-English languages were selected from those best represented in PLEIAS's earlier Common Corpus dataset. Reasoning traces are written in English even where the question and answer are in another language; a question may be quoted verbatim in its original language within an English trace. Of the seven non-English languages present in material quantity, six are official languages of the European Union (German, Spanish, French, Polish, Italian, Dutch) and one is a historical European language (Latin).
Other relevant characteristics of the overall training data:	The dataset is derived from a deliberately small and fixed seed corpus of 62,555 documents, each amplified at least one hundred times, so that the factual knowledge the models can acquire is bounded by, and traceable to, an enumerable set of source documents. The seed corpus is composed of 58,698 Wikipedia articles — 50,000 from the community-curated "Vital articles" lists (levels 1 to 5) and 8,698 further articles added to reinforce coverage of law, medicine and chemistry — together with 3,727 Wikibooks pages on cooking and practical knowledge and 130 documents written internally by PLEIAS. The 130 internal documents were amplified approximately ten thousand times. PLEIAS states expressly in its published documentation that the seed selection carries the structural biases of Wikipedia contribution and editing, and that the selection was made from the perspective of Western European and United States culture. Coverage of knowledge specific to other regions is correspondingly limited. Because the dataset contains encyclopaedic information about well-known historical and public figures and no naturally occurring personal data, PLEIAS's published documentation states that no personal-data curation was required. See Section 3.2. The dataset is published in full at https://huggingface.co/datasets/PleIAs/SYNTH under the CC-BY-4.0 licence, so that the entire training content of these two models is publicly inspectable, record by record. Each record additionally reproduces its encyclopaedic seed passage verbatim in a dedicated field, together with that passage's own URL and licence; those seed passages remain subject to their own licence, which is CC-BY-SA 4.0 for 99.46% of records and CC0 for 0.54%.

Token counts are computed with the PLEIAS tokenizer, a byte-pair-encoding tokenizer with a vocabulary of 65,536 entries. Monad was trained with a different, purpose-built tokenizer of 8,192 entries; the dataset is identical in both cases and only its tokenised length differs. Because tokenizer vocabularies differ between providers, these token counts are not directly comparable with figures reported elsewhere.

Additional comments (optional):

The dataset occupies approximately 240 GB in its published Parquet form.

Baguettotron's tokenizer reserves approximately 50 control tokens used within reasoning traces: markers for logical, epistemic and verification steps, and simulated entropy annotations. These are training-time annotations present in the data; they do not correspond to any inference-time mechanism.

---

1. List of data sources

This Section requires information about specific sources of data used to train the general-purpose AI model. In this section "dataset" should be understood as a single, pre-packaged collection of data. The filtering and pre-processing of data collected from the same pre-packaged collection should not be considered a new dataset to be disclosed separately in the sections below. If a particular dataset can be assigned to more than one of the categories below, providers should select the most relevant category and only report the dataset in that category, except in the case of synthetic data (see Section 2.5).

2.1. Publicly available datasets

This Section requires information about datasets that were used to train the model and which have been compiled by a third party, are made available publicly for free, and are readily downloadable as a whole or in predefined chunks, such as datasets and collections available on public repositories and online platforms, specialised websites, or snapshots of common crawl. The public availability of the datasets for free does not mean that the content at issue is necessarily free of rights since it may be subject to licensing arrangements or conditions of use (e.g., certain free and/or open licenses may determine the scope of the uses, including prohibiting uses relating to model training).

A dataset is considered to be "large" if the total data size for any one of the modalities contained in the dataset exceeds 3% of the size of all publicly available datasets for that modality used for training. The size of the dataset should be based on its size after pre-processing (for example filtering), and without splitting the dataset to prevent reporting circumvention.

Have you used publicly available datasets to train the model? ☒ Yes ☐ No

If yes, specify the modality(ies) of the content covered by the datasets concerned: ☒ Text ☐ Image ☐ Video ☐ Audio ☐ Other *If so, please specify…*

List of large publicly available datasets:

The dataset actually used for training, SYNTH, is synthetic and is therefore reported in Section 2.5, in accordance with the instruction in Section 2 of the Template that synthetic data be reported in Section 2.5 even where it could also be assigned to another category. This Section reports the publicly available datasets that were used as the seed material from which the synthetic data was generated.

1. Structured Wikipedia (Wikimedia Enterprise) — https://huggingface.co/datasets/wikimedia/structured-wikipedia
Machine-readable dumps of Wikipedia articles published by the Wikimedia Foundation

through its Wikimedia Enterprise service. PLEIAS used 58,698 articles from this dataset: 50,000 from the community-curated "Vital articles" lists at levels 1 to 5 (https://en.wikipedia.org/wiki/Wikipedia:Vital_articles/Level/5) and 8,698 further articles selected by category-tree and Wikidata-graph expansion to reinforce coverage of law, medicine and chemistry. Content is licensed CC-BY-SA 4.0. This is the largest seed source and accounts for the substantial majority of the seed material.

2. Wikibooks — https://www.wikibooks.org

3,727 pages on cooking and practical knowledge, obtained through the official Wikimedia API, licensed CC-BY-SA 4.0. This category was included because practical and procedural knowledge is under-represented in encyclopaedic articles.

3. Formalised mathematics templates, mostly from the Kimina dataset - https://huggingface.co/datasets/AI-MO/Kimina-Prover-Promptset — approximately 3,000 formalised mathematical exercises, of which most came from the Kimina dataset, used as templates for the mathematical portion of the dataset. Variable values were randomised and the symbolic solutions recomputed, so that the released mathematical exercises are generated rather than reproduced.

Approximate dates of data collection: the seed material was retrieved in 2025 from the then-current Wikimedia dumps and API. The licence recorded for each individual seed record is published in the dataset: CC-BY-SA 4.0 for 99.46% of records and CC0 for 0.54%.

General description of other publicly available datasets not listed above:

No other publicly available dataset was used. Common Crawl and comparable general web-crawl corpora were not used.

The seed corpus was deliberately kept small and enumerable. PLEIAS's published rationale is that bounding the seed set bounds the factual knowledge the models can memorise, and makes it possible to state exactly which documents a model's parametric knowledge derives from — a property that is not available for models trained on web-scale corpora.

Additional comments (optional):

Every record generated from an encyclopaedic seed carries the URL and the licence of that seed, so that the provenance of the individual training example can be traced back to its source article. The mathematical exercises, approximately 1.6% of records, are generated from formalised templates rather than from an encyclopaedic passage and carry no seed URL.

Documentation: dataset card https://huggingface.co/datasets/PleIAs/SYNTH; "It's All Training: A Fully Synthetic Single-Stage Recipe for LLMs" (Langlais et al., preprint); blog post "SYNTH: the new data frontier", 10 November 2025, https://pleias.fr/blog/blogsynth-the-new-data-frontier.

2.2 Private non-publicly available datasets obtained from third parties

This Section requires information about private non-publicly available datasets of third parties that are not publicly available and not disclosed under Section 2.1. These include:

1) *datasets for which transactional commercial licensing agreements were concluded between the provider and the rightsholders or their representatives, including by collective management organisations and legitimate content aggregators who have the right to collectively license works on behalf of rightsholders (Section 2.2.1);*
2) *other private datasets obtained through data intermediaries, non-publicly available databases and datasets of third parties for which transactional commercial licenses have not been concluded with rightsholders or their representatives (Section 2.2.2).*

2.2.1. Datasets commercially licensed by rightsholders or their representatives

Have you concluded transactional commercial licensing agreement(s) with rightsholder(s) or with their representatives?

☐ Yes ☒ No

If yes, specify the modality(ies) of the content covered by the datasets concerned:

Not applicable. No transactional commercial licensing agreements were concluded with rightsholders or their representatives.

2.2.2. Private datasets obtained from other third parties

Have you obtained private datasets from third parties that are not licensed as described in Section 2.2.1, such as data obtained from providers of private databases, or data intermediaries?	☐ Yes ☒ No
If yes, specify the modality(ies) of the content covered by the datasets concerned:	Not applicable.
If publicly known, list private datasets obtained from other third parties:	Not applicable. No private dataset was obtained from any third party.
General description of non-publicly known private datasets obtained from third parties	Not applicable.
Additional comments (optional):	PLEIAS has not obtained any dataset from a data intermediary, a private database provider or any other third party on a non-public basis for the training of these models.

2.3 Data crawled and scraped from online sources

This Section requires information about crawled, scraped data, or otherwise compiled from online sources directly by the provider of the model or on their behalf (i.e. excluding publicly available datasets already compiled by third parties and made available on platforms such as common crawl that are covered under Section 2.1).

Were crawlers used by the provider or on behalf of?	☐ Yes ☒ No
If yes, specify crawler name(s)/identifier(s):	Not applicable. No crawler was used by PLEIAS or on its behalf for the training of these models.
Purposes of the crawler(s):	Not applicable.
General description of crawler behaviour:	Not applicable. No content was crawled or scraped from online sources for the training of these models.
Period of data collection:	The seed material was obtained from machine-readable dumps published by the Wikimedia Foundation through its Wikimedia Enterprise service and from the official Wikimedia API. Both are interfaces provided by the rightsholder for the express purpose of bulk reuse of freely licensed content. Obtaining data through them does not involve crawling or scraping, and no link-following, rate-limited harvesting or access to any site other than those official interfaces took place.
Comprehensive description of the type of content and online sources crawled:	Not applicable.
Type of modality covered:	Not applicable. No content was crawled or scraped from online sources for the training of these models.
Summary of the most relevant domain names crawled:	☐ Text ☐ Image ☐ Video ☐ Audio ☐ Other *If so, please specify…*
Additional comments (optional):	Not applicable. No internet domain was crawled or scraped.
	For completeness, the seed material originates from two domains, accessed through the rightsholder's own bulk-distribution interfaces rather than by crawling: wikipedia.org (via Wikimedia Enterprise) and wikibooks.org (via the official Wikimedia API).
	The absence of any crawling is a deliberate design property of these models rather than an incidental fact. PLEIAS's stated purpose in building a fully synthetic training corpus from a small seed set was to demonstrate that a competitive

	general-purpose model can be trained without recourse to web-scraped data.

2.4 User data

This Section requires information about user data collected by all services and products of the provider, including through mail services, social media platforms, content platforms or interaction with the providers' AI models and/or systems. This does not cover data licensed by users based on commercial transactional agreements described in Section 2.2.1., or customer data to fine-tune models for specific purposes.

Was data from user interactions with the AI model (e.g. user input and prompts) used to train the model?	☐ Yes ☒ No
Was data collected from user interactions with the provider's other services or products used to train the model?	☐ Yes ☒ No
	Not applicable. No data from user interactions with these models, and no data collected from any other PLEIAS service or product, was used to train them.
If yes, provide a general description of the provider's services or products that were used to collect the user data:	The models are released as open-weight models under the Apache 2.0 licence and are downloaded and run by third parties on their own infrastructure. PLEIAS does not operate a consumer-facing service that collects prompts, conversations or other user content for training purposes, and has never used data of that kind in any training stage of these models.
Type of modality covered:	Not applicable.
Additional comments (optional):	PLEIAS confirms that no interaction data of any kind — including prompts, conversations, uploaded documents, feedback signals or telemetry — has been collected for or used in the training of these models at any stage.

2.5 Synthetic data

This Section requires information about synthetic data created by or on behalf of the provider for training the model directly on the outputs of another AI model, in particular through model distillation or model alignment (e.g. AI feedback through reinforcement learning). This does not include the use of AI models to clean or enrich data (e.g. AI-generated metadata to enrich or modify a dataset, such as creating depth maps or text descriptions of images). In case this concerns publicly available datasets as described in Section 2.1, these should be reported in that Section of the Template. In case this concerns synthetic datasets created by third parties on behalf of the provider, these should be reported in this Section of the Template instead of in Section 2.2.2.

Was synthetic AI-generated data created by the provider or on their behalf to train the model?	☒ Yes ☐ No
If yes, modality of the synthetic data:	☒ Text ☐ Image ☐ Video ☐ Audio ☐ Other *If so, please specify…*
If yes, specify the general-purpose AI model(s) used to generate the synthetic data if available on the market:	Yes. The entire training corpus of these two models is synthetic: the SYNTH dataset, comprising 79,648,272 generated samples totalling approximately 75 billion tokens, published at https://huggingface.co/datasets/PleIAs/SYNTH under the CC-BY-4.0 licence. The generation pipeline has two stages. In the first, PLEIAS fine-tunes small auxiliary models on curated examples. In the second, those fine-tuned auxiliaries are run at scale over the seed corpus to produce the dataset. The generation was therefore performed by PLEIAS's own fine-tuned derivatives, but those derivatives are based on third-party general-purpose AI models available on the market, which are identified below.

General-purpose AI models on which the generators are based, all available on the market:

- **Qwen 3 8B**, provider Alibaba Cloud (Qwen Team) — the basis of the task-specific fine-tuned models that generated approximately 97% of the dataset, covering memorisation, multiple-choice questions, constrained writing, retrieval-augmented generation, creative writing, editing and practical-knowledge exercises, and the symbolic re-evaluation of the mathematical exercises. Model documentation: https://huggingface.co/Qwen/Qwen3-8B
- **DeepSeek-Prover (7B)**, provider DeepSeek — used in a drafter-and-solver configuration for the mathematical exercises, approximately 2.5% of the dataset. Model documentation: https://huggingface.co/deepseek-ai/DeepSeek-Prover-V2-7B

Links to the Summaries of those models, where published by their respective providers, may be obtained from the model documentation linked above.

The identity of the model that generated the answer and reasoning trace is recorded in a dedicated `model` field of every record of the published dataset, so that the Qwen 3 8B and DeepSeek-Prover attributions can be verified directly and per record.

Non-generative models used in the pipeline. A frozen bge-m3 text-embedding model and a FAISS index were used to retrieve the nearest-neighbour seed paragraph accompanying each generated question. These perform retrieval, not generation, and are not general-purpose AI models.

Information about other AI models, including provider's own AI model(s) not available on the market, used to generate synthetic data to train the model to which this Summary applies:

PLEIAS's own fine-tuned derivatives, not placed on the market. The task-specific fine-tunes of Qwen 3 8B, the DeepSeek-Prover drafter and solver configuration described above were created by PLEIAS for the sole purpose of generating this dataset. They have not been placed on the market. Each was fine-tuned on curated triplets of seed passage, constraints and target output drawn from the same openly licensed seed corpus described in Section 2.1, together with the distilled traces referred to in the confirmation note above.

Selection of the generating models. PLEIAS's position is that small fine-tuned generators integrated into structured pipelines give better control over diversity and factual grounding than prompting a single large model.

Additional comments (optional):

The design feature most relevant to the interests this Summary is intended to serve is that generation proceeds by *back-translation*: a question is generated from an authentic seed passage, so that every training target is anchored to a verifiable statement in a freely licensed encyclopaedic source, rather than to the unconstrained output of a large model. PLEIAS's stated purpose is to limit the propagation of teacher-model errors and to keep the factual content of the models traceable.

Approximately 20% of the generated queries deliberately target a refusal, a correction or a hedge rather than a confident answer, so that the models learn to decline when

a question is ill-posed or unanswerable, as described in the accompanying technical report.

The generated question, reasoning trace and answer are paraphrases of the seed passage rather than reproductions of it. Each published record additionally carries the seed passage verbatim in a dedicated field, together with the seed URL and licence, so that the freely licensed source of each record is directly inspectable.

The dataset was released jointly by PLEIAS and the AI Alliance, and was designed around a set of open standards for synthetic data developed with the AI Alliance.

2.6 Other sources of data

This Section requires information about data that does not fall under any of the categories in the previous Sections, for example data collected from offline sources, self-digitised media (e.g., digitised analog text context, images), datasets labelled by humans commissioned by the provider, or human generated data through reinforcement learning.

Have data sources other than those described in Sections 2.1 to 2.5 been used to train the model?	☒ Yes ☐ No
If yes, provide a narrative description of these data sources and the data:	Yes, in one limited respect. The seed corpus includes **130 documents written internally by PLEIAS staff**, covering documentation of the models themselves, information about their training conditions, general information on AI research, and recent events postdating the Wikipedia snapshot used for the other seeds. These are human-authored documents created by the provider, and so fall outside Sections 2.1 to 2.5. Although they represent only 130 of the 62,555 seed documents, these texts were amplified approximately ten thousand times in the generation process — a substantially higher amplification factor than the minimum of one hundred applied to the encyclopaedic seeds — because the knowledge they carry is not otherwise present in the seed corpus. They are disclosed here notwithstanding their small number, because their influence on the models' self-description and on their statements about recent events is disproportionate to their size. The documents are original works authored by PLEIAS. They contain no third-party content, no personal data of third parties, and no material obtained from any external source.
Additional comments (optional):	No other data source outside Sections 2.1 to 2.5 was used. In particular, no offline or self-digitised material, no dataset labelled by human annotators commissioned by PLEIAS, and no human-generated data obtained through reinforcement learning was used to train these models. The manual curation applied to the fine-tuning sets of the auxiliary generators was performed internally on machine-generated text and introduced no new source of data.

1. **Data processing aspects**

3. Data processing aspects

3.1. Respect of reservation of rights from text and data mining exception or limitation

This Section concerns measures implemented by the provider to identify and comply with the reservation of rights from the text and data mining (TDM) exception or limitation expressed pursuant to Article 4(3) of Directive (EU) 2019/790, as outlined in the copyright policy put in place by the provider in accordance with Article 53(1)(c) AI Act.

Are you a Signatory to the Code of Practice for general-purpose AI models that includes commitments to respect reservations of rights from the TDM exception or limitation?	☒ Yes ☐ No
Describe the measures implemented before model training to respect reservations of rights from the TDM exception or limitation before and during data collection, including the opt-out protocols and solutions honoured by the provider or, as applicable, by third parties from which datasets have been obtained:	PLEIAS is a signatory to the General-Purpose AI Code of Practice and applies its Copyright chapter commitments. For these two models the reservation-of-rights question arises in a materially narrower form than for models trained on crawled data, because no content was crawled or scraped and the entire training corpus was generated from a fixed set of 62,555 seed documents, of which 62,425 are freely licensed third-party documents and 130 are original works authored by PLEIAS. **1. No crawling or scraping of the open web.** No crawler was operated by PLEIAS or on its behalf for these models, and no content was collected by crawling or scraping any website. Content was obtained exclusively through the rightsholder's own bulk-distribution channels. It follows that PLEIAS did not access any content in respect of which a rightsholder had expressed a reservation of rights pursuant to Article 4(3) of Directive (EU) 2019/790, whether by machine-readable means or otherwise. **2. Seed material obtained under a free licence, through the rightsholder's own distribution channel.** The seed corpus consists of Wikipedia and Wikibooks content, licensed CC-BY-SA 4.0 (99.46% of records) or CC0 (0.54%), obtained from machine-readable dumps published by the Wikimedia Foundation through Wikimedia Enterprise and from the official Wikimedia API. Both are channels operated by the rightsholder for the express purpose of bulk reuse. The reproduction of that content therefore rests on the licence granted by the rightsholder and not on the text-and-data-mining exception, so that a reservation of rights under Article 4(3) has no application to it. **3. Generated content is paraphrase, not reproduction.** The questions, reasoning traces and answers that constitute the training targets are paraphrases of the seed passages rather than reproductions of them. Each published record additionally reproduces its seed passage verbatim in a dedicated field, so that the source of the record is inspectable; that passage is freely licensed encyclopaedic text and travels with its own licence notice. **4. Per-record provenance.** Every record generated from an encyclopaedic seed carries the URL and the licence of that seed, together with the identity of the model that generated the answer. A rightsholder can therefore establish, record by record, what seed material was used and under what licence. The mathematical exercises, approximately 1.6% of records, are generated from formalised templates and carry no seed URL.

5. Reuse of third-party model outputs. The general-purpose AI models used to generate the dataset were selected on the basis that their licences permit the reuse of their outputs, including for training purposes, so that no restriction attaching to a third-party model's terms of use was disregarded.

6. Complaints and removal. Rightsholders who consider that material has been included in error may write to contact@pleias.fr: PLEIAS examines such notifications and acts on those that are well founded, including by removing material from the published dataset.

PLEIAS's copyright policy under Article 53(1)(c) AI Act is published at pleias.ai/training-content

Additional comments (optional):

The licensing analysis underlying the corpus is documented in arXiv:2506.01732, and in particular in its Appendix E on the verification of public-domain status. PLEIAS notes that the same Appendix records that it did not attempt to establish public-domain status on the basis of non-renewal of United States copyright; where United States material published after 1929 is present, its public-domain status rests on the determination made by the digitising institution.

3.2 Removal of illegal content

This Section concerns measures taken to avoid or remove illegal content under Union law from the training data (such as blacklists, keywords, and model-based classifiers), without requiring disclosure of specific details about the provider's internal business practices or trade secrets. Such measures are advisable if the training data is likely to include illegal or unlawful content under Union law, in particular child sexual abuse material and terrorist content and the non-authorised use of material protected by intellectual property rights. Such measures do not include data selection practices, for example to increase the capability of the model.

General description of measures taken:

Composition of the training data as the primary measure. The training corpus was generated from 62,555 seed documents drawn almost entirely from community-curated Wikipedia and Wikibooks articles. No content was taken from the open web, from social media, from forums or from any user-generated-content platform. The categories of illegal content associated with large-scale web scraping — child sexual abuse material, terrorist content and unauthorised reproductions of protected works — are therefore excluded at source rather than filtered out after the fact. PLEIAS's published documentation states that the systematic grounding of every record in Wikipedia gives the dataset a very low risk of toxic or otherwise problematic content.

Personal data. The seed corpus consists of encyclopaedic material concerning well-known historical and public figures, drawn from articles that Wikipedia's own editorial and biographies-of-living-persons policies already govern. PLEIAS's published documentation states expressly that, for this reason, no personal-data curation step was required for this dataset. No naturally occurring personal data — such as would arise from scraped social media, forums or web pages — is present, because no such source was used. The 130 internal seed documents were authored by PLEIAS and contain no third-party personal data.

Independent quality assessment. As reported in the accompanying technical report, the dataset was scored

alongside a range of open pre-training corpora (including Nemotron-CC, FineWeb-2, FinePDFs, FineWiki, HPLT-4 and Common Corpus) using the Propella-1 multilingual document annotator, a model developed by a third party. On the integrity and safety axes of that assessment SYNTH scores at the ceiling, matching Wikipedia-derived corpora and above web-cleaned corpora.

Exclusion of source code. Code data was deliberately excluded from the dataset, removing a category of content that carries distinct licensing and security risks.

Filtering of generations. Generations were filtered during production to remove defective outputs. The generating models — Qwen 3 8B, DeepSeek-Prover and Gemma 3 12B — carry their own provider-level safety mitigations.

Limitations stated in good faith. PLEIAS states in its published documentation that the seed selection carries the structural biases of Wikipedia contribution and editing and reflects a Western European and United States perspective; and that at these parameter counts, and particularly for Monad at 56 million parameters, factual hallucination is to be expected. These models were released as base models and have not undergone a separate safety-alignment stage. Notifications concerning illegal content in the published dataset may be sent to contact@pleias.fr.

3.3. Other information (optional)

Other relevant information about data processing (optional):

Scope of this Summary. In accordance with point (30) of the Commission Explanatory Notice, this single Summary covers two model versions whose training content is identical.

Related Summaries. PLEIAS publishes two further Summaries covering its other model families: one for the Pleias 1.0 Preview family (Pleias-350m-Preview, Pleias-1.2b-Preview, Pleias-3b-Preview) and one for the Pleias-RAG family (Pleias-RAG-350M, Pleias-RAG-1B). Those models were trained on a different corpus and are not covered here.

Openness and verifiability. The entire training corpus of these two models is published, record by record, under the CC-BY-4.0 licence. Each record carries the seed URL, the seed licence, the generating model, the constraints applied, the generated reasoning trace and the generated answer. Every factual assertion in this Summary concerning the composition, provenance, language distribution and generation of the training data can therefore be verified directly against the published dataset rather than taken on trust:

- dataset — https://huggingface.co/datasets/PleIAs/SYNTH
- models, under Apache 2.0 — https://huggingface.co/PleIAs/Baguettotron and https://huggingface.co/PleIAs/Monad
- training framework — Nanotron, https://github.com/huggingface/nanotron

Primary references.

- Langlais, Delobelle, Detrois et al., "It's All Training: A Fully

Synthetic Single-Stage Recipe for LLMs" (preprint).
- "SYNTH: the new data frontier", PLEIAS blog, 10 November 2025, https://pleias.fr/blog/blogsynth-the-new-data-frontier.

Compute, disclosed voluntarily. Both models were trained on 16 NVIDIA H100 GPUs at the Jean Zay supercomputer (GENCI/IDRIS, compute plan A0191016886) using the Nanotron framework. Monad's full pre-training took a little under six hours. The final training runs for both models together represent fewer than 1,000 H100-hours; the project as a whole, including the generation of the dataset and all preparatory experiments, amounted to approximately 20,000 H100-hours.

Governance. The dataset was released jointly by PLEIAS and the AI Alliance, and was designed around a set of open standards for synthetic data developed in collaboration with the AI Alliance.

Updating. This Summary will be updated in accordance with point (29) of the Commission Explanatory Notice if these models are further trained on additional data. No such further training has taken place to date.

---

[1] Excluding audio that is part of video, as this should be reported under the "video" modality instead. Furthermore, the Commission understands the modality of 'audio' to include 'speech'.

Contact

Station F, 5 Parv. Alan Turing 75013, Paris
contact@pleias.fr

LinkedIn

GitHub

HuggingFace

pleias