pleias

All summaries | Pleias 1.0 | Baguettotron and Monad

Template for the Public Summary of Training Content for General-Purpose AI models: Pleias-1.0

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 three 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/Pleias-350m-Preview** (also referred to as Pleias-pico-350m-Preview) — 353,424,384 parameters — https://huggingface.co/PleIAs/Pleias-350m-Preview • **PleIAs/Pleias-1.2b-Preview** (also referred to as Pleias-

	nano-1.2b-Preview) — 1,195,468,800 parameters — https://huggingface.co/PleIAs/Pleias-1.2b-Preview • **PleIAs/Pleias-3b-Preview** — 3,211,926,528 parameters — https://huggingface.co/PleIAs/Pleias-3b-Preview All three are decoder-only transformer text-generation models released under the Apache 2.0 licence. The model card for each is published at the URL given above. The three differ only in parameter count and in the number of passes made over the same corpus. Collectively these models are referred to in PLEIAS's published documentation as the "Pleias 1.0" family.
Model dependencies:	None. All three models were pre-trained from scratch by PLEIAS. None of them is a modification, fine-tune, distillation or other derivative of a pre-existing general-purpose AI model, and none of them declares a base model.
Date of placement of the model on the Union market:	All three models were placed on the Union market before 2 August 2025. They are therefore covered by point (33) of the Commission Explanatory Notice, under which the corresponding Summary is to be made publicly available no later than 2 August 2027. • Pleias-350m-Preview — repository created 19 November 2024; model card published 1 December 2024. • Pleias-1.2b-Preview — repository created 27 November 2024; model card published 1 December 2024. • Pleias-3b-Preview — repository created 1 December 2024; model card published 2 December 2024. The family was publicly announced in December 2024. None of the three models has been further trained since its release, and no new version has been placed on the market.

1.3 Modalities, overall training data size and other characteristics

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: • Corpus as constituted at the time of training: approximately 2 trillion tokens. • Filtered corpus used for the main training pass: 1,086,324,736,000 tokens. • Tokens processed per model, counting repeated passes: approximately 1.09 trillion (350m); 4 trillion (1.2b); 5 trillion (3b).	Long-form written text drawn exclusively from public-domain or openly licensed sources. The principal categories, by share of tokens, are: • books, monographs, newspapers and periodicals digitised by public cultural-heritage institutions — approximately 46%; • legal, legislative, judicial, parliamentary, administrative and financial-regulatory documents published as open data — approximately 19%; • source code and software documentation under permissive open-source licences —

		approximately 17%; • scientific and scholarly articles, abstracts and preprints under open-access licences — approximately 11%; • encyclopaedic and collaboratively authored web content under free licences, together with openly licensed speech transcripts — approximately 7%. Within those totals, approximately 1.5% of the corpus — some 30 billion tokens — is synthetic text generated to cover task formats under-represented in the openly licensed material (see Section 2.5). It is a subset of the categories above and not an additional category. Contemporary press publications under copyright were not used. Historical newspapers are included only where the digitising institution has determined them to be in the public domain; for United States newspapers published between 1929 and 1963 that determination rests on the Library of Congress's own assessment. No fiction or non-fiction work under active copyright protection was used. No social media content was used; the only user-generated-content sources present are Stack Exchange and YouTube-Commons, both under free licences.
☐ 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:	11/2024. The corpus was frozen before pre-training began. None of the three models is continuously trained on new or dynamic data after that date, and none has been further trained since release.
Description of the linguistic characteristics of the overall training data:	The training data is multilingual. Approximately 35–40% of tokens are in languages other than English: English 64.4%; French 14.8%; German 6.68%; Spanish 2.62%; Latin 2.03%; Dutch 1.45%; Italian 1.26%; Polish 0.67%; Greek 0.64%; Portuguese 0.53%.

	Dozens of further languages are present in smaller proportions, including Danish, Slovak, Czech, Estonian, Hungarian, Swedish, Finnish, Maltese, Bulgarian, Lithuanian, Romanian, Slovenian, Latvian, Croatian and Irish. All 24 official languages of the European Union are represented, principally through EUR-Lex and EuroVoc/ Cellar, which cover all 24 official languages; Europarl covers 21. Non-EU languages present in smaller quantities include Russian, Ukrainian, Japanese, Chinese, Arabic and Haitian Creole. Language identification was performed with fastText.
Other relevant characteristics of the overall training data:	More than half of the corpus predates the 21st century. A substantial share consists of historical printed material digitised by national libraries and archives and processed by optical character recognition; recognition artefacts were corrected with a purpose-built model (OCRonos) and long concatenated scans were re-segmented into coherent documents (Segmentext). Geographical and institutional coverage is weighted towards France, the wider European Union and the United States, reflecting the cultural-heritage institutions and open-data programmes from which the material originates. Cultural-heritage content was collected predominantly from institutions established in the European Union or the United States, together with the National Library of New Zealand and a small number of other non-EU European institutions. Every document carries structured provenance metadata — identifier, collection, curator, licence, date, title, creator, language, word count and token count — which is published together with the corpus. PLEIAS's published documentation states expressly that the use of public-domain and openly licensed sources does not by itself eliminate bias: historical texts can contain archaic prejudiced language, and the corpus reflects the collection priorities of the digitising institutions.
Additional comments (optional):	Token counts are computed with the PLEIAS tokenizer, a byte-pair-encoding tokenizer with a vocabulary of 65,536 entries trained on a representative sample of the corpus. Because tokenizer vocabularies differ between providers, these token counts are not directly comparable with figures reported elsewhere. Number of passes per model: • Pleias-350m-Preview — one pass over the filtered corpus (1,086,324,736,000 tokens). • Pleias-1.2b-Preview — one pass over the full corpus, then two further passes over the filtered corpus. The model card describes this as "over three epochs (nearly 5 trillions tokens)"; Langlais et al., Procedia Computer Science 267 (2025), section 3.3 describes it as one epoch on the full corpus and two on the filtered subset. • Pleias-3b-Preview — two passes over the full corpus (approximately 2 trillion tokens), then one pass over a more aggressively filtered subset of approximately 1 trillion tokens (an "annealing" phase). Context length at training: 2,048 tokens for the 350m and 1.2b models; 4,096 tokens for the 3b model.

1. List of data sources

2.1. Publicly available datasets

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:

Common Corpus — https://huggingface.co/datasets/PleIAs/common_corpus

Common Corpus is the only dataset used to pre-train these models and therefore accounts for effectively 100% of the publicly available text data used. It is a corpus of public-domain and openly licensed text compiled by PLEIAS and published publicly, free of charge and downloadable as a whole.

Version used. These models were trained on the first release of Common Corpus, published on 13 November 2024, comprising **2,003,039,184,047 tokens** in five collections. The models have not been retrained since. The dataset published on the Hugging Face Hub has been extended twice since that date: a second version added a sixth collection (Open Semantic, derived from Wikidata) and document-level metadata, and a third version substantially expanded language coverage and added a subset derived from Creative Commons–filtered Common Crawl. **Neither of those later additions was present in, or used for, the training of these models.** As stated on the model cards, the corpus used for these models deliberately excluded Common Crawl and comparable general web-crawl archives. The figures reported below are therefore those of the November 2024 release and not those of the current published dataset or of the later technical paper.

Dates of the data collection. The material itself dates from the 17th century to 2024; more than half of the corpus predates the 21st century. Compilation of the corpus by PLEIAS was completed in November 2024.

Composition (release of 13 November 2024). Each of the five collections exceeds 3% of the total for the text modality and is therefore itemised below.

- **OpenCulture — 926,541,096,243 tokens (46.3%).** Public-domain books, monographs, newspapers and periodicals digitised by cultural-heritage institutions, including Chronicling America (Library of Congress; US newspapers published 1690–1963), Gallica (Bibliothèque nationale de France), Europeana Newspapers (over 1,000 titles from 23 European libraries, published 1618–1990), the Deutsches Zeitungsportal / Deutsche Digitale Bibliothek (1794–1957), the Biblioteca Digital Hispánica (Biblioteca Nacional de España), Delpher (Koninklijke Bibliotheek, Netherlands), the National Library of Luxembourg (1841–1879), Papers Past (National Library of New Zealand), the Library of Congress Selected Digitized Books collection, the Internet Archive, Project Gutenberg and Wikisource. Cultural-heritage material was collected predominantly from institutions established in the European Union or the United States, together with the National Library of New Zealand and a small number of other non-EU European institutions.

- **OpenGovernment — 387,965,738,992 tokens (19.4%).** Legislative, judicial, parliamentary, administrative and financial-regulatory material published as open data: USPTO patent prosecution records 2019–2022 (obtained via Pile of Law), EUR-Lex, EuroVoc via Cellar (Publications Office of the European Union), Europarl, TED / Supplement to the Official Journal, French open data published by the Direction de l'information légale et administrative (DILA) and other French administrative bodies and courts, CourtListener (Free Law Project), the Caselaw Access Project (Harvard Law

School Library; cases from 1658 to 2020), the UN Digital Library, OECD publications, filings from the US Securities and Exchange Commission via EDGAR (1993–2024), WTO Documents Online (1995–2024), the GATT Digital Library (1946–1996) and open data of the French Autorité des marchés financiers.

- **OpenSource — 334,658,896,533 tokens (16.7%).** Source code and software documentation under permissive open-source licences, obtained from The Stack v1 and v2 (BigCode), which compile publicly hosted GitHub repositories (arXiv:2506.01732, section 4.4). Files were filtered to permissive licences only and further filtered for format and quality.

- **OpenScience — 221,798,136,564 tokens (11.1%).** Open-access scientific articles, abstracts and preprints: principally OpenAlex filtered to CC-BY, CC0/public domain and CC-BY-SA only, together with French, Spanish and German open-science repositories and arXiv.

- **OpenWeb — 132,075,315,715 tokens (6.6%).** Wikipedia and Wikisource obtained from Wikimedia Enterprise dumps; YouTube-Commons (transcripts of 2,063,066 videos published by their uploaders under CC-BY); Stack Exchange (CC-BY-SA).

Licence composition. For the corpus as documented in arXiv:2506.01732, Table 4: public domain 57.0%; CC-BY 14.4%; MIT 7.1%; CC-BY-SA 3.7%; Apache-2.0 3.4%; BSD-3-Clause 0.92%; other open licences for the remainder. That table describes a later release of the corpus; the distribution of licence families in the November 2024 release was materially similar, the majority being public domain. The licence applicable to each individual document is recorded in that document's own metadata and published with the corpus.

General description of other publicly available datasets not listed above:

None. Common Corpus is the only publicly available dataset used. Common Crawl and comparable general web-crawl corpora were deliberately excluded.

Additional comments (optional):

Why this dataset is reported here. Section 2.1 of the Template is addressed to publicly available datasets compiled by a third party. Common Corpus was compiled by PLEIAS itself rather than by a third party. PLEIAS nonetheless reports it in this Section, because it is publicly available, free of charge and downloadable as a whole, and additionally reports its own acts of retrieval under Section 2.3, so that the Summary is complete on either reading of the Template. The two Sections describe the same body of content and are not cumulative.

Third-party datasets incorporated into the corpus. A substantial part of Common Corpus consists of collections already compiled and published by third parties, which PLEIAS obtained as pre-packaged datasets rather than retrieving itself. These include: The Stack v1 and v2 (BigCode); Pile of Law (USPTO records); the Caselaw Access Project; CourtListener (Free Law Project); EDGAR-CORPUS (Loukas et al., 2021) for SEC filings up to 2020; Europeana Newspapers and the BnL Newspapers collection (both released through BigScience); the German public-domain newspaper and Library of Congress book collections curated by Sebastian Majstorovic; the EUR-Lex collections developed by Loza Mencía and Fürnkranz (2010) and Chalkidis et al. (2019); the EuroVoc collection compiled by Sébastien Campion; Europarl (Koehn, 2005); Wikimedia Enterprise dumps; OpenAlex; YouTube-Commons by pleias; and Stack Exchange as distributed in The Pile.

Provenance metadata. Per-document metadata published with the corpus records, for each document: identifier, collection, curator, licence, date, title, creator, language, word count and token count. A rightsholder can therefore determine directly whether a given work is present and on what legal basis.

Documentation. Dataset card https://huggingface.co/datasets/PleIAs/common_corpus; release announcement of 13 November 2024, https://huggingface.co/blog/Pclanglais/two-trillion-tokens-open; "Common Corpus: The Largest Collection of Ethical Data for LLM Pre-Training", arXiv:2506.01732 (ICLR 2026), which describes a later release of the corpus; "Pleias 1.0: the First Ever Family of Language Models Trained on Fully Open Data", Procedia Computer Science 267 (2025) 146–156, doi:10.1016/j.procs.2025.08.241.

2.2 Private non-publicly available datasets obtained from third parties

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. All data used to train these models is publicly available.

2.3 Data crawled and scraped from online sources

Were crawlers used by the provider or on behalf of?	☒ Yes ☐ No
If yes, specify crawler name(s)/identifier(s):	PLEIAS does not operate a general-purpose web crawler and has not crawled the open web at large. No crawler product or user-agent identifier is published by PLEIAS. Content was obtained from identified online sources by two means: (i) internal retrieval scripts written for each individual source, which query documented public interfaces — APIs, OAI-PMH endpoints, bulk-export and bulk-download services and open-data portals — and which retrieve an identified collection rather than following links across the web; (ii) for filings of the US Securities and Exchange Commission covering 2021–2024, the third-party open-source EDGAR-Crawler toolkit (https://github.com/nlpaueb/edgar-crawler), applied to the SEC's own EDGAR bulk-access service. Constitution of the Common Corpus pre-training dataset.
Purposes of the crawler(s):	Retrieval was directed exclusively at collections that the publishing institution makes available for reuse: public-domain digitised heritage collections, open-data portals of public administrations and courts, official bulk-download and API services of public bodies and international organisations, and openly licensed repositories. The purpose was to obtain complete, identified collections whose legal status could be established, not to discover or harvest content across the open web.

General description of crawler behaviour:	Retrieval was performed source by source against documented public interfaces rather than by link-following across the web. Consequently the behaviours normally associated with general-purpose crawling did not arise in the following respects: • No paywall, access-control measure, subscription barrier, password protection, CAPTCHA or other technical protection measure was circumvented. Only openly accessible interfaces were used. • No content was obtained from sources that make works available unlawfully. • Retrieval was kept within the rate limits and terms of use published by each institution for its API or bulk-download service. • No general-purpose crawl of the open web was performed, and Common Crawl and comparable web-archive corpora were excluded from the corpus used for these models. Robots.txt directives were read and honoured, however retrieval targeted documented bulk and API endpoints of institutions publishing open data.
Period of data collection:	From 01/2024 to 11/2024.
Comprehensive description of the type of content and online sources crawled:	The content retrieved directly by PLEIAS consists of: • digitised books, monographs, newspapers and periodicals in the public domain held by national libraries, archives and cultural-heritage institutions, predominantly in the European Union and the United States; • legislative, judicial, parliamentary and administrative documents published as open data by French, European Union, United States and international public bodies; • filings and publications of financial market authorities and international organisations; • open-access scientific literature. Geographically the material is concentrated on France, the wider European Union and the United States. Linguistically it is dominated by French and English, with substantial German, Spanish, Italian, Dutch, Latin, Polish, Greek and Portuguese content. Chronologically it ranges from the 17th century to 2024, with more than half of the material predating the 21st century. The categories of online source involved are: websites and open-data portals of cultural-heritage institutions; government portals and official journals; websites of courts and legal-information services; websites of financial market authorities and international organisations; open-access scientific repositories. The following categories of online source were NOT crawled or scraped by PLEIAS: news websites operating under copyright, social media, forums, community websites and other user-generated-content platforms, personal blogs, streaming platforms, gaming platforms, online TV platforms and synthetic data libraries. Two openly licensed user-generated sources — Stack Exchange and YouTube-Commons — are present in the corpus, but were obtained as pre-packaged third-party datasets reported in Section 2.1 rather than by crawling.

Type of modality covered:	☒ Text ☐ Image ☐ Video ☐ Audio ☐ Other *If so, please specify…*
Summary of the most relevant domain names crawled:	PLEIAS is a small enterprise within the meaning of recital 109 AI Act. The list below is given in two parts so that rightsholders can distinguish content retrieved by PLEIAS from content that entered the corpus through third-party datasets reported in Section 2.1. **(a) Domains from which PLEIAS retrieved content directly.** These account for the substantial majority of the material retrieved by PLEIAS: gallica.bnf.fr; api.bnf.fr; shiny.ens-paris-saclay.fr (Gallicagram); archive.org; gutenberg.org; wikisource.org; delpher.nl; bne.es; natlib.govt.nz; repos.ids-mannheim.de; echanges.dila.gouv.fr; legifrance.gouv.fr; data.gouv.fr; courdecassation.fr; conseil-etat.fr; data.europa.eu; ted.europa.eu; consilium.europa.eu; archives.eui.eu; sec.gov; docs.wto.org; amf-france.org; digitallibrary.un.org; oecd.org; openalex.org; arxiv.org. **(b) Domains whose content entered the corpus through third-party compiled datasets** (reported in Section 2.1) rather than through retrieval by PLEIAS: github.com (The Stack v1 and v2); uspto.gov (Pile of Law); case.law (Caselaw Access Project); courtlistener.com (Free Law Project); sec.gov for filings up to 2020 (EDGAR-CORPUS); eur-lex.europa.eu (collections of Loza Mencía and Fürnkranz, and Chalkidis et al.); op.europa.eu (EuroVoc, compiled by Sébastien Campion); europarl.europa.eu (Europarl, Koehn 2005); europeana.eu and data.bnl.lu (released through BigScience); deutsche-digitale-bibliothek.de and loc.gov (collections curated by Sebastian Majstorovic); chroniclingamerica.loc.gov; wikipedia.org and wikisource.org (Wikimedia Enterprise dumps); stackexchange.com (as distributed in The Pile); youtube.com (YouTube-Commons). In addition, the domain of origin of every individual document is recorded in the published corpus metadata, so that any rightsholder can determine directly and exhaustively whether material from a given domain is present.
Additional comments (optional):	The relationship between Sections 2.1 and 2.3 in this Summary reflects the fact that PLEIAS compiled its own training corpus. The corpus is reported under Section 2.1 because it is publicly available and downloadable as a whole; the acts of first-party retrieval by which PLEIAS assembled parts of it are reported here so that the Summary is complete. The two Sections describe the same content from two angles and are not to be added together. Beyond the domain list above, PLEIAS enables any party with a legitimate interest, including rightsholders, to establish upon request whether content from a specific internet domain has been used, by consulting the published per-document metadata or by writing to contact@pleias.fr.

2.4 User data

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
If yes, provide a general description of the provider's services or products that were used to collect the user data:	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. 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

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:	Approximately 30 billion tokens — about 1.5% of the corpus — of synthetic text were generated with the fine-tuned Qwen 2-7B (https://huggingface.co/Qwen/Qwen2-7B), not publicly released, to provide examples of task formats under-represented in the openly licensed material, principally conversational question-and-answer formats.
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 used the following models of its own in the preparation of the training data. Under Section 2.5 of the Template these uses constitute the cleaning and enrichment of existing data rather than the creation of synthetic training data; they are nonetheless disclosed here on a voluntary basis so that the picture is complete. • **OCRonos** — a post-correction model based on Llama 3 8B, used to repair optical-character-recognition errors in digitised print. It restores the text of the scanned original rather than generating new content. Published at https://huggingface.co/PleIAs/OCRonos. • **Segmentext** — a text-segmentation model used to split concatenated scans into coherent documents. Published at https://huggingface.co/PleIAs/Segmentext. • **Celadon** — a DeBERTa-v3-small classifier trained by PLEIAS on 2 million annotated samples, used to score and filter toxic content (see Section 3.2). Published at https://huggingface.co/PleIAs/celadon. • An unreleased internal small reasoning model, used to identify and drop parts of the French administrative material presenting a heightened risk of indirect personal identification. Passages flagged by Celadon in the highest toxicity band were synthetically rewritten rather than removed, and passages in the intermediate band were annotated with a generated content warning. That rewriting and annotation modify existing documents and are therefore reported as data processing under Section 3.2 rather than as the

	creation of synthetic data. Both were performed with **Llama 3.1 8B Instruct**, a general-purpose AI model made available by Meta Platforms, Inc. (arXiv:2410.22587; arXiv:2506.01732, Appendix F.5). Model documentation: https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct
Additional comments (optional):	None of the PLEIAS models listed above is itself placed on the market as a general-purpose AI model. OCRonos is derived from Llama 3 8B, and the rewriting described above used Llama 3.1 8B Instruct; both are general-purpose AI models made available by Meta Platforms, Inc. PLEIAS's published position is that the proportion of synthetic data was deliberately kept small in order to avoid the degradation associated with training predominantly on model-generated text.

2.6 Other sources of data

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:	Not applicable. No data source outside the categories described in 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.
Additional comments (optional):	These models were released as base models. They did not undergo a separate supervised fine-tuning, preference-optimisation or reinforcement-learning stage, and no human preference or annotation data was therefore collected or used.

1. **Data processing aspects**

3. Data processing aspects

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

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. The design of the training data makes the reservation-of-rights question structurally different from that arising for models trained on general web crawls. The measures implemented before and during data collection were as follows. **1. Licence-based sourcing.** Content was admitted to the training corpus only where it was verified to be in the public domain or released under a free licence permitting reuse, including for commercial purposes. Material whose copyright status could not be established was not included. This is a sourcing standard PLEIAS sets for itself; it operates in addition to, and not in substitution for, the measures

described below by which PLEIAS identifies and complies with reservations of rights. Nothing in this Summary is a waiver, disclaimer or limitation of any exception, limitation or defence available to PLEIAS under Union or national law.

2. Reservations of rights. PLEIAS does not acquire for training any content in respect of which a rightsholder has expressed a reservation of rights pursuant to Article 4(3) of Directive (EU) 2019/790, whether expressed by machine-readable means or in any other appropriate manner. Where a source publishes a reservation applying to a collection, that collection is excluded. Because PLEIAS performs no general-purpose crawl of the open web (see point 6 below), the circumstance in which such reservations are most commonly encountered at scale does not arise in its data acquisition; that is a fact about how PLEIAS collects data and not a reason to apply a lower standard.

3. Verification of public-domain status. The following criteria were applied: (i) for non-US authors, the author's death plus 70 years, established through an internal reconciliation pipeline against a complete dump of Wikidata; (ii) for US authors, publication plus 95 years; (iii) for several book collections where the author could not be identified, a conservative rule admitting only publications prior to 1884; (iv) for newspaper and other collective-work collections, the public-domain determination made by the digitising institution, which PLEIAS relied upon and recorded in the metadata.

4. Digitisation and rights. PLEIAS takes the position, consistent with the principle underlying Article 14 of Directive (EU) 2019/790 for works of visual art and with common practice among cultural-heritage reusers, that the faithful digitisation of a public-domain work does not of itself give rise to a new protected right. No restriction asserted by a digitising institution over a public-domain work was therefore treated as creating one. PLEIAS states this as its position rather than as a settled conclusion of law.

5. Licence filtering at source. Open-access scientific literature was filtered to CC-BY, CC0/public domain and CC-BY-SA only. Source code was filtered to permissive licences only. Web content was limited to material published by its authors under free licences (Wikimedia CC-BY-SA, YouTube-Commons CC-BY, Stack Exchange CC-BY-SA).

6. Exclusion of general web crawls. Common Crawl and comparable web-archive corpora were deliberately excluded from the corpus used to train these models, as stated expressly on the published model cards. PLEIAS performed no general-purpose crawl of the open web and therefore did not access domains carrying machine-readable reservations of rights. First-party retrieval was directed only at institutional repositories, official open-data services and openly licensed repositories whose published terms permit reuse.

7. No circumvention. No paywall, access-control measure, subscription barrier or other technical protection measure was circumvented, and no content was taken from sources making works available unlawfully.

---

8. Per-document provenance and licence metadata. The source and the applicable licence of every document are recorded in the corpus metadata and published with the corpus, so that any rightsholder can verify directly whether and on what basis a given work is present.

9. Complaints and removal. Rightsholders who consider that material has been included in error may write to contact@pleias.fr. PLEIAS examines such notifications and removes material from published versions of the corpus where the objection is well founded.

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

General description of measures taken:

Composition of the corpus as the primary measure. Because no general web crawl was used, and because every document was admitted only on the basis of a verified public-domain status or a free licence, the categories of illegal content most commonly associated with large-scale web scraping — child sexual abuse material, terrorist content and unauthorised reproductions of protected works — are largely excluded at source. The material originates overwhelmingly from national libraries, public archives, government open-data services, courts, international organisations, open-access scientific repositories and openly licensed collaborative projects, each of which applies its own publication controls. No social media platform was used. The only user-generated-content sources present are Stack Exchange and YouTube-Commons, both obtained as pre-packaged openly licensed datasets rather than by crawling.

Measures applied to the corpus before training.

- Toxicity classification and filtering. Documents were scored by Celadon, a DeBERTa-v3-small classifier trained by PLEIAS on approximately 640,000 samples drawn from a 2-million-sample annotated dataset, across five dimensions (race and origin-based bias, gender and sexuality-based bias, religious bias, ability bias, and violence and abuse). Documents in the highest toxicity band were removed or synthetically rewritten so as to remove the harmful language while preserving the informational content; documents in the intermediate band were annotated with a generated content warning. The rewriting and warning generation were performed with Llama 3.1 8B Instruct. The classifier and its training data are published (https://huggingface.co/PleIAs/celadon and https://huggingface.co/datasets/PleIAs/ToxicCommons) and the method is documented in arXiv:2410.22587.

- Keyword and slur filtering. Texts containing offensive terms and slurs were removed.
- Personal data. Personally identifiable information was detected using Microsoft Presidio with additional language- and country-specific rules, for example European telephone-number formats, which raised phone-number detection accuracy from approximately 55–60% to approximately 85%. Detected personal data — telephone numbers, email addresses, IBANs, IP addresses and URLs — was removed from the text. Parts of the French administrative material were dropped in their entirety, using an internal model, because of the heightened risk of transmitting indirectly identifying personal information.
- Quality and integrity filtering. Rule-based and model-based filters removed gibberish, extremely repetitive text and documents of unusable length; optical-character-recognition errors in digitised print were detected and corrected.

Limitations stated in good faith. PLEIAS states expressly, in its published documentation, that no curation method achieves complete accuracy and that residual problematic material may remain in a corpus of this size, and that the dataset card describes personal-data removal as attempted rather than exhaustive. These models were released as base models and have not undergone additional safety alignment; this is stated on each model card. Notifications concerning illegal content in the published corpus may be sent to contact@pleias.fr, and PLEIAS acts on well-founded notifications by removing the material from published versions of the corpus.

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 three model versions whose training content is identical. In accordance with point (33), these models were placed on the Union market before 2 August 2025.

Related Summaries. Pleias-RAG-350M and Pleias-RAG-1B are mid-trained variants of Pleias-350m-Preview and Pleias-1.2b-Preview respectively; they are covered by a separate Summary, which incorporates the content of this one and adds the mid-training data.

Openness and verifiability. The training corpus, the model weights, the tokenizer and the training configuration are published under open licences, so that the disclosures in this Summary can be verified independently rather than taken on trust:

- models, under Apache 2.0 — https://huggingface.co/PleIAs
- training corpus — https://huggingface.co/datasets/PleIAs/common_corpus
- training framework — Nanotron, https://github.com/huggingface/nanotron; the 1.2b model was trained with the TractoAI fork, https://github.com/tractoai/nanotron
- data-processing tools — https://github.com/Pleias/open_data_toolkit

Primary references.

- Langlais et al., "Pleias 1.0: the First Ever Family of Language Models Trained on Fully Open Data", Procedia Computer Science 267 (2025) 146–156, doi:10.1016/j.procs.2025.08.241.
- Langlais et al., "Common Corpus: The Largest Collection of Ethical Data for LLM Pre-Training", arXiv:2506.01732 (ICLR 2026).
- Arnett et al., "Toxicity of the Commons: Curating Open-Source Pre-Training Data", arXiv:2410.22587.

Compute and environmental information, disclosed voluntarily.

- Pleias-350m-Preview — 64 NVIDIA H100 GPUs for 46 hours at the Jean Zay supercomputer (GENCI/IDRIS); estimated 0.5 tCO2eq.
- Pleias-1.2b-Preview — 192 NVIDIA H100 GPUs for 5 days on TractoAI (ISEG cluster, Nebius AI); estimated 4 tCO2eq.
- Pleias-3b-Preview — 192 NVIDIA H100 GPUs for approximately 20 days at Jean Zay (compute grant GC011015451); estimated 16 tCO2eq.

Support acknowledged in the published documentation. étalab; GENCI Grand Challenge / Jean Zay / IDRIS; EVIDEN; NVIDIA Inception; TractoAI; Scaleway; the Mozilla Foundation Local AI Programme; the AI Alliance; LANGU:IA (French Ministry of Culture and DINUM); ALT-EDIC; Occiglot; Wikimedia Enterprise; Wikimedia Deutschland; Libraries Without Borders.

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