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

Version of the Summary: V1.1, first published version. No previously published version of this Summary exists.
Last update: 31/07/2026

1. General information

1.1. Provider identification

Provider name and contact details
LINAGORA — SIREN 431 473 669, registered under number 431 473 669 R.C.S. Nanterre.

Registered office: Villa Good Tech, 37 rue Pierre Poli, 92130 Issy-les-Moulineaux, France.

Contact for matters relating to this Summary: jplorre@linagora.com

Websites:

- https://www.linagora.com
- https://openllm-france.fr

LINAGORA is a small or medium-sized enterprise within the meaning of Commission Recommendation 2003/361/EC. The SME threshold referred to in Section 2.3 of this template is therefore the applicable one.

Authorised representative name and contact details
Not applicable. The provider is established in the Union (France); Article 54 AI Act does not apply.

1.2. Model identification

Versioned model name(s)
This Summary covers the following models, whose training content is identical, in accordance with point 30 of the Commission Explanatory Notice:

- Luciole-1B-Base: https://huggingface.co/OpenLLM-France/Luciole-1B-Base
- Luciole-8B-Base: https://huggingface.co/OpenLLM-France/Luciole-8B-Base
- Luciole-23B-Base: https://huggingface.co/OpenLLM-France/Luciole-23B-Base

All models are released under the Apache 2.0 licence with publicly available weights, including intermediate checkpoints available here: https://dl.labs.linagora.com/files/models/OpenLLM-France/

Model dependencies

The Luciole base models are pre-trained from scratch by the provider. Their training did not involve modification or fine-tuning of third-party model weights. There are accordingly no general-purpose AI models already placed on the Union market on which the base models depend. All models were trained using version 2.3.1 of the NeMo library.

- Luciole-1B-Base: a dense transformer model trained using a custom adaptation of the Nemotron3 4B architecture recipe.
- Luciole-8B-Base: a hybrid Mamba-transformer model trained using a custom adaptation of the NemotronH-8B recipe.
- Luciole-23B-Base: a dense transformer model trained using a custom adaptation of the Nemotron3 22B architecture recipe.

Date of placement of the model on the Union market

02 June 2026

1.3 Modalities, overall training data size and other characteristics

Modality

☒ Text

Text is the only modality present in the training data

Training data size

☒ 1billion to 10 trillion tokens

Approximately 4.65 trillion tokens after pre-processing.

Types of content

Encyclopaedic and reference content; scientific and academic text (arXiv, HAL, doctoral theses, PubMed); legal, parliamentary and institutional documents (EUR-Lex, French Parliament, OECD, WTO, INSEE); press and historical newspapers; public-domain literature; web pages; forum and question-and-answer content; source code; mathematical text; parallel translation corpora; transcribed speech (subtitles); instruction, reasoning and dialogue data.

Latest date of data acquisition/collection for model training

- Principal training phases (1 and 2): June 2025; the most recent web-derived material comes from the CommonCrawl dump CC-MAIN-2025-26.
- Annealing and context extension: December 2025.

Description of the linguistic characteristics of the overall training data

Raw pretraining dataset: The pre-training data is multilingual with a deliberate European emphasis. Composition of the corpus as assembled:

- English 53.4%
- French 16.3%
- German 5.6%
- Spanish 4.9%
- Italian 2.8%

- Portuguese 1.9%
- Dutch 1.4%
- Arabic 0.7%
- Parallel bilingual data 0.7%
- Programming languages 11.3%
- Mathematical content 4.7%.

Training proportions after applying sampling weights (including upsampling of French data):

- English 41.9%
- French 30.4%
- German 3.8%
- Spanish 3.5%
- Italian 1.9%
- Portuguese 1.3%
- Dutch 1.0%
- Arabic 0.5%
- Parallel bilingual data 1.7%
- Programming languages 9.2%
- Mathematical content 3.5%

Regional and minority languages of the Union account for approximately 0.4% of the corpus: Basque, Breton, Catalan, Corsican, Franco-Provençal, French-based creoles, Occitan, Picard and Walloon (together with Tahitian). Wikimedia-derived content covers 21 languages.

Other relevant characteristics of the overall training data

The corpus contains a substantial share of French institutional, legal, parliamentary, statistical and heritage content — proceedings and written questions of the French Parliament, EUR-Lex, INSEE, Gallica monographs and press, HAL, French doctoral theses and data.gouv.fr — reflecting the intended use of the models in French and other European-language settings.

Additional comments

Tokeniser vocabulary: 128 000 tokens.

Context length: 131 072 tokens for the base models, extended progressively during training from 4 096.

Training phases: the base models were exposed to approximately 4.9 trillion tokens across five training phases: 3.5 T, 1.5 T, 300 B annealing, 50 B and 50 B for context extension (for the 8B, context extension was performed as a single, fourth phase). The difference from the 4.65 trillion tokens of the raw corpus reflects the re-use and re-weighting of parts of the corpus, not additional content.

Training infrastructure: training was carried out on 256–512 NVIDIA H100 80 GB GPUs of the Jean Zay supercomputer (GENCI/IDRIS), representing 576,587 GPU-hours for the 23B model, 237,037 GPU-hours for the 8B and 41,912 GPU-hours for the 1B.

Funding: development was carried out by LINAGORA within the OpenLLM-France consortium with funding from Bpifrance under the France 2030 programme.

2. List of data sources

2.1. Publicly available datasets

Have you used publicly available datasets to train the model?

☒ Yes

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

☒ Text

List of large publicly available datasets

Substantially all of the training content consists of pre-packaged, openly licensed datasets compiled by third parties. Rather than apply the 3% materiality threshold, the provider discloses the complete list of the 87 source configurations used. That list, with per-source licence information, is published and kept up to date at: https://huggingface.co/datasets/OpenLLM-France/Luciole-Training-Dataset

The principal sources are the following.

- Web (filtered): FineWeb 2 and FineWeb2-HQ (ODC-BY); FineWeb-Edu (ODC-BY); DCLM Dolmino (ODC-BY); CulturaX (mC4 / OSCAR terms); HPLT 2 (CC0 1.0).
- Encyclopaedic: Wikipedia, Wikibooks, Wiktionary, Wikisource, Wikiquote, Wikinews, Wikivoyage, Wikiversity (GFDL / CC BY-SA); Vikidia (GFDL).
- Institutional and legal: Common Corpus subsets EUR-Lex, OECD, WTO, TED EU tenders and GATT library (public domain/open); Eurovoc (EUPL 1.1); data.gouv.fr open data (ODC-BY); INSEE publications (ODC-BY); French Parliament amendments, public speeches, interventions and written questions (CC BY-SA / Licence Ouverte Etalab 2.0); Europarl (open).
- Academic: Common Pile, PubMed, LibreTexts, StackExchange, arXiv papers and abstracts (mixed open); HAL (HAL licence); French doctoral theses (Licence Ouverte Etalab 2.0).
- Books and heritage: Project Gutenberg (public domain); Gallica monographs and press (public domain); Common Pile pre-1929 books (public domain); BNL Luxembourg newspapers 1841–1879 (public domain).
- Code: StarCoder Data (mixed open); StarCoder Olmomix (ODC-BY); Stack-Edu (mixed open); Common Pile GitHub Archive (mixed open); Open Code Reasoning (CC BY 4.0).
- Mathematics: FineMath 3+ and 4+ (ODC-BY); InfiMM-WebMath (ODC-BY); MegaMath Web (ODC-BY, more than 300 B tokens); MathPile Commercial (CC BY-SA 4.0).
- Instruction and dialogue: Aya Dataset (Apache 2.0); Claire dialogue corpora (CC BY-NC-SA 4.0)[1]; Nemotron Post-Training v2 (CC BY 4.0); OpenThoughts (Apache 2.0); OpenMathInstruct-1 (NVIDIA licence); PleiasSynth (CDLA-Permissive 2.0).
- Parallel corpora: Europarl parallel (open); CroissantAligned (CC BY-SA 4.0); ParaDocs (Apache 2.0); Translation-Instruct (CC BY-SA 4.0).
- Transcribed speech: subtitles of French-language public video content available under permissive terms, used as text.

[1] Only subcorpora whose licences allow for commercial use were selected from the Claire corpora.

Approach to selecting parts of datasets: several of these datasets were used only in part. Selection was made by language, by quality score (for example FineWeb2-HQ retains the top decile by classifier score), by mathematical or educational subset, and by the retrospective robots.txt filtering described in Section 3.1. The selection criteria applied to each source are documented in the published corpus card and in the processing scripts found at:
https://github.com/OpenLLM-France/Luciole-Training/tree/main/data/processing/pretraining.

General description of other publicly available datasets not listed above

None. All publicly available datasets used are enumerated in the published corpus documentation referred to above.
The corpus is restricted to material distributed under open licences or in the public domain (public domain dedications, CC0, CC BY, CC BY-SA, ODC-BY, Apache 2.0, EUPL, Licence Ouverte Etalab 2.0 and comparable terms). It includes copyright-protected content distributed under those open licences; personal data present in public web and institutional content, subject to the pseudonymisation described in Section 3.2; and machine-generated content originating from the third-party instruction datasets listed above.

Additional comments

The corpus itself is published under CC BY-SA 4.0 with 87 loadable configurations and per-source documentation and the processing scripts are public at
https://github.com/OpenLLM-France/Luciole-Training/tree/main/data/processing/pretraining
Verification of the statements in this Summary is therefore possible at source level.

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?
☒ No

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?
☒ No

2.3 Data crawled and scraped from online sources

Were crawlers used by the provider or on behalf of?
☒ No

2.4 User data

Was data from user interactions with the AI model (e.g. user input and prompts) used to train the model? ☒ No
Was data collected from user interactions with the provider's other services or products used to train the model? ☒ No No data from users of the provider's services or products was used to train the models. In particular, no LinTO, Twake or other LINAGORA product data, and no logs from public Luciole demonstrators, entered the pre-training or post-training corpora.

2.5 Synthetic data

Was synthetic AI-generated data created by the provider or on their behalf to train the model? ☒ Yes
If yes, modality of the synthetic data ☒ Text
If yes, specify the general-purpose AI model(s) used to generate the synthetic data if available on the market Synth FineWeb 2: using Qwen 3 8B, we synthetically augmented documents from FineWeb 2 by prompting the model to reformulate them using three different levels of difficulty: easy, medium, difficult. Synth Wikipedia: using Qwen 3 8B, we synthetically augmented documents from Wikipedia by generating question/response pairs based on the content of the Wikipedia document. The question/answer pairs were appended to the end of the document concerned.
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: N/A
Additional comments Several third-party post-training datasets used by the provider were themselves produced by AI models (e.g., Nemotron, OpenCodeReasoning, OpenThoughts, PleiasSynth). Because they were obtained as publicly available datasets rather than generated by or on behalf of the provider, they are reported under Section 2.1. Before use, the provider removed references to third-party model names and company names, and removed Chinese and Russian script, in order to limit the transfer of provenance artefacts and cultural bias.

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? ☒ No

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

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:

LINAGORA is a signatory to the General-Purpose AI Code of Practice and adheres to its Copyright chapter, including Measure 1.3 on the identification of and compliance with reservations of rights. The measures below implement that commitment.

1. Restriction of the corpus to openly licensed material. The corpus was assembled exclusively from sources distributed under open licences or in the public domain. Content whose rightsholders had reserved rights under Article 4(3) of Directive (EU) 2019/790 was therefore not a target of collection.
2. Retrospective application of robots.txt to all web-derived datasets. Robots.txt files were retrieved from the CommonCrawl dump CC-MAIN-2025-26, the most recent file for each host was retained, and a document was kept only where the robots.txt explicitly permitted crawling by CCBot or where the file was malformed. This filter was applied to FineWeb and its derivative datasets, CulturaX, DCLM Dolmino, FineMath, HPLT2 InfiWebMath and MegaMath. The processing code is public.
3. A standing opt-out mechanism. Rightsholders and data subjects who identify their protected work or personal data in the corpus may request its removal through the form published at https://openllm-france.fr/delete-data/. Requests are processed against the published corpus and removals are reflected in the next corpus release.

For datasets obtained from third parties, the provider additionally relies on the collection practices and rights-reservation compliance of the upstream compilers, as documented by those compilers.

Limits of these measures, stated for completeness: The robots.txt filter was applied after collection by the upstream compilers rather than during collection and relies on a single snapshot dated June 2025. It evaluates the CCBot user-agent only, and it does not capture reservations expressed by means other than robots.txt, such as the TDM Reservation Protocol, metadata assertions or contractual terms of use. The provider has undertaken, in its copyright policy, to re-run the evaluation at each corpus release, to extend it to further user-agents and to read TDMRep assertions where present.

3.2 Removal of illegal content

General description of measures taken

Sourcing constraint: By restricting source corpora to openly licensed, largely institutional, encyclopaedic, academic and heritage sources, we substantially reduce exposure to illegal material; the main sources of illegal content are expected to come from web crawled data, for which we took additional measures described below.

Pseudonymisation of personal data: E-mail addresses were replaced by placeholders (for example email@example.com), IP addresses by the token <IP_ADDRESS>, and telephone numbers (detected with the phonenumbers library) by the token <PHONE_NUMBER>. Pseudonymisation was applied to CulturaX, DCLM, FineWeb 2, FineWeb-Edu, FineWeb-HQ, FineWeb2-HQ, HPLT 2 and Common Corpus.

Quality and toxicity filtering: English web data sources (DCLM Dolmino, FineWeb-edu, FineWeb-HQ) were annotated and filtered using model-based quality and content classifiers prior to publication. Multiligual web sources (FineWeb2, HPLT, CulturaX) were annotated for quality and toxicity using an in-house classifier trained following the FineWeb-edu approach. In later phases, data from FineWeb2-HQ, which were filtered for quality and toxicity prior to publication, were used. Later stages of training were restricted to data with high quality and educational scores.

Child sexual abuse material and terrorist content: All FineWeb corpora are filtered as described in the code here https://github.com/huggingface/datatrove/blob/main/src/datatrove/pipeline/filters/url_filter.py#L33. CulturaX and HPLT were filtered for adult content based on the blacklist provided by the University of Toulouse https://dsi.ut-capitole.fr/blacklists/.

Disclaimer: The provider states openly in the published corpus documentation that, despite these measures, toxic and offensive documents may remain in the web-derived portion, and that historical material carries period biases relating to gender, ethnicity, skin colour and religion.

3.3. Other information

Other relevant information about data processing

Deduplication was applied across the corpus; the methods are documented in the public processing scripts.

The complete corpus, the processing pipeline and the training configuration are published, so that the statements in this Summary can be independently verified rather than merely asserted.

This Summary is kept up to date and republished at least every six months, or sooner upon any material change such as further pre-training, a new post-training run or a new corpus release. Superseded versions are archived and remain accessible.

Requests concerning this Summary, and rights-related requests, may be addressed to the contact given in Section 1.1.