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-Instruct-1.1: https://huggingface.co/OpenLLM-France/Luciole-1B-Instruct-1.1
- Luciole-8B-Instruct-1.1: https://huggingface.co/OpenLLM-France/Luciole-8B-Instruct-1.1
- Luciole-23B-Instruct-1.1: https://huggingface.co/OpenLLM-France/Luciole-23B-Instruct-1.1

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 Instruct 1.1 models are post-trained versions of the Luciole base models: Luciole-1B-Base, Luciole-8B-Base, Luciole-23B-Base found here: https://huggingface.co/collections/OpenLLM-France/luciole-llm. Training took place in three stages, with each stage starting from the last checkpoint of the previous phase:

- Luciole-xB-SFT-Thinking: a supervised fine-tuning of the corresponding base model using data with thinking traces.
- Luciole-xB-SFT-Instruct: a supervised fine-tuning of the corresponding base model using data without thinking traces.
- Luciole-xB-Instruct: alignment of the SFT-Instruct model using DPO.

Date of placement of the model on the Union market
09 July 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
☒ Less than 1 billion tokens
Approximately 4B for SFT Thinking, 2.3B for SFT, 0.75B for SFT DPO after pre-processing

Types of content
SFT: Examples of instructions with responses in various domains including: code, math, STEM, chat, RAG, tool calling, translation and NLI.
DPO: Examples of instructions with pairs of responses where one is labelled “chosen” and the other “rejected”. Same domains as SFT, excluding translation and including safety

Latest date of data acquisition/collection for model training
June 2026

Description of the linguistic characteristics of the overall training data
The Luciole-PostTraining-Dataset v1.1 consists almost exclusively of English language data, with the exception of some multilingual representation in the Nemotron datasets and safety datasets.

Other relevant characteristics of the overall training data
The categories of data used during training break down roughly as follows:

- Chat: 25%
- Math: 13%
- STEM: 14%
- Code: 16%
- Tool calling: 11%
- NLI: 13%

- Safety: 5%
- Other (including multilingual): 3%

Additional comments

Tokeniser vocabulary: 128 000 tokens.
Context length: 16,384 tokens.
Training phases: SFT Thinking used 2.6 million samples, SFT Instruct used 2.1 million, and Instruct DPO used 200,000 samples.
Training infrastructure: training was carried out on NVIDIA H100 80 GB GPUs of the Jean Zay supercomputer (GENCI/IDRIS).
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

Much of the training content for the supervised fine-tuning (SFT) phases of Instruct 1.1 models consists of pre-packaged, openly licensed datasets compiled by third parties or by OpenLLM partners. Rather than apply the 3% materiality threshold, the provider discloses the complete list datasets used. That list is published and kept up to date at: https://huggingface.co/datasets/OpenLLM-France/Luciole-PostTraining-Dataset-1.1

For SFT Thinking, we used the following third-party datasets:

- DOLCI [Think Persona Precise IF, Think Precise IF, SYNTHETIC-2-SFT-Verified] (ODC-BY)
- Nemotron Posttraining v3 [Chat English, Math v2, Science MCQ] (CC-BY-4.0)
- Nemotron Posttraining v2 [Code, Math fr] (CC-BY-4.0)
- Open Code Reasoning (CC-BY-4.0)
- Nemotron agentic sft v2 (Apache-2.0)
- smolagent tool calling (Apache-2.0)
- Pleias RAG (CC-BY-4.0)

For SFT without thinking:

- DOLCI [Think Precise IF, Sciriff, Python Algos, Flan2, Logic puzzles] (ODC-BY)
- Nemotron Posttraining v3 [Chat English] (CC-BY-4.0)
- Nemotron Posttraining v2 [Code, Stem] (CC-BY-4.0)
- Pleias RAG (CC-BY-4.0)
- Linagora [Personas Math, HotpotQA, TatQA, Hardcoded] (CC-BY-4.0)
- XLAM (CC-BY-4.0)
- Hermes (Apache-2.0)

- When2call (CC-BY-4.0)
- Paradocs (Apache-2.0)
- Croissant (CC-BY-SA-4.0)
- Smol [instruct, rewrite, summarize] (Apache-2.0)

For DPO, we used prompts from the following datasets:

- DOLCI [Think Persona Precise IF, Think Precise IF, Sciriff, Python Algos, Flan2] (ODC-BY)
- Nemotron Posttraining v3 [Chat English, Math v2, Science MCQ] (CC-BY-4.0)
- Nemotron Posttraining v2 [Code, Stem] (CC-BY-4.0)
- Nemotron Safety (CC-BY-4.0)
- OpenMathInstruct (NVIDIA License)
- Pleias RAG (CC-BY-4.0)
- Smol [instruct, rewrite, summarize] (Apache-2.0)
- When2call (CC-BY-4.0)
- XLAM (CC-BY-4.0)

To preprocess the split datasets, we checked for the presence of names of LLMs and companies (Claude, Amazon, etc.) as well as for Chinese and Russian. Given that most of the data was generated with open source models from Qwen and DeepSeek, there was a preponderance of Chinese and Russian script intermingled with our languages of interest. Samples containing any of these were removed entirely. The prepocessing scripts can be found in this folder of the Luciole-Training repository: https://github.com/OpenLLM-France/Luciole-Training/tree/main/data/processing/posttraining

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.

Additional comments

The corpus itself is published under CC BY-SA 4.0. The processing scripts are public at https://github.com/OpenLLM-France/Luciole-Training/tree/main/data/processing/posttraining
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 Generation was used in the DPO phase. With the exception of data for safety alignment, the alignment pairs were generated synthetically with a delta learning approach: all pairs were generated with Qwen3-32B and Qwen3-0.6B and the former were labelled as the accepted responses. Safety data were generated with a mixture of Qwen3-14B , Ministral-3-14B-Instruct, and an interim checkpoint of Luciole-8B-Instruct, after SFT. Pairs were judged with both Ministral-14B-Reasoning and Qwen3-14B. A pair was included only if both models agreed on their labels.
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

An an interim checkpoint of Luciole-8B-Instruct, after SFT, was used to make some safety data. The training data was the same as described above.

Additional comments

Several third-party post-training datasets used by the provider were themselves produced by AI models. 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

Annotation of preference pairs for DPO was determined automatically by using a larger model to produce the accepted responses and a very small model to generate the rejected responses.

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.

This commitment is maintained through testriction 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.

Because the data comes entirely from synthetic generation or standard NLP datasets

(e.g., HotPot, FLAN), no particular measures were taken to filter protected work or personal data (as this is not expected to be a problem).

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.

3.2 Removal of illegal content

General description of measures taken
Data is either generated using open models or chosen from open datasets published by third parties after extensive filtering or careful url selection to target specific domains such as math, code and science. It is taken not from massive, uncontrolled web content, thus illegal content is not expected to be a problem. Highly toxic or dangerous content is not expected to be present either except in the case of safety data used in the DPO phase. This content cannot be heavily filtered due to its purpose in training models what not to say.

3.3. Other information

Other relevant information about data processing
To preprocess the split datasets, we checked for the presence of names of LLMs and companies (Claude, Amazon, etc.) as well as for Chinese and Russian. Given that most of the data was generated with open source models from Qwen and DeepSeek, there was a preponderance of Chinese and Russian script intermingled with our languages of interest. Samples containing any of these were removed entirely. The prepocessing scripts can be found in this folder of the Luciole-Training repository: https://github.com/OpenLLM-France/Luciole-Training/tree/main/data/processing/posttraining