

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

Version of the Summary: v1.0

Last update: 31 July 2026

1. General information

1.1. Provider identification

Provider name and contact details: EuroLLM Team, <https://euollm.io/>
Contact: andre.t.martins@tecnico.ulisboa.pt, a.birch@ed.ac.uk

Authorised representative name and contact details: N/A

1.2. Model identification

Versioned model name(s): EuroLLM-9B, <https://huggingface.co/utter-project/EuroLLM-9B>
EuroLLM-9B-2512, <https://huggingface.co/utter-project/EuroLLM-9B-2512>
EuroLLM-9B-Instruct, <https://huggingface.co/utter-project/EuroLLM-9B-Instruct>
EuroLLM-9B-Instruct-2512, <https://huggingface.co/utter-project/EuroLLM-9B-Instruct-2512>
EuroLLM-22B-2512, <https://huggingface.co/utter-project/EuroLLM-22B-2512>
EuroLLM-22B-Instruct-2512, <https://huggingface.co/utter-project/EuroLLM-22B-Instruct-2512>

Model dependencies: N/A

Date of placement of the model on the Union market: EuroLLM-9B, EuroLLM-9B-Instruct: 9 December 2024
EuroLLM-9B-2512, EuroLLM-9B-Instruct-2512, EuroLLM-22B-2512, EuroLLM-22B-Instruct-2512: 14 December 2025

1.3 Modalities, overall training data size and other characteristics

Modality <i>Select the modalities present in the training data, to the extent that they are identifiable</i>	Training data size <i>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.</i>	Types of content <i>For each selected modality, provide a general description of the type of content that has been included in the training data.</i>
☒ Text	☐ Less than 1 billion tokens ☒ 1billion to 10 trillions tokens ☐ More than 10 trillions tokens Alternatively, specify the approximate size in a different measurement unit: <i>4 trillion tokens</i>	<i>Public text-only datasets derived mainly from web documents in 35 languages. The training data is fully transparent and reproducible (EuroLLM is a fully open model).</i>

Latest date of data acquisition/collection for model training:	<i>The main pre-training dataset knowledge cutoff is 02/2024. Some domain-specific parts of the dataset (math) and parts of the post-training datasets have a later date of collection.</i>
Description of the linguistic characteristics of the overall training data:	<i>Languages covered: Bulgarian, Croatian, Czech, Danish, Dutch, English, Estonian, Finnish, French, German, Greek, Hungarian, Irish, Italian, Latvian, Lithuanian, Maltese, Polish, Portuguese, Romanian, Slovak, Slovenian, Spanish, Swedish, Arabic, Catalan, Chinese, Galician, Hindi, Japanese, Korean, Norwegian, Russian, Turkish, Ukrainian</i>
Other relevant characteristics of the overall training data:	<i>Data for each language reflects the amount of available web resources in that language. Filtering was applied to ensure high quality data.</i>
Additional comments (optional):	<i>For further details, refer to the 9B and 22B model technical reports.</i>

2. 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:

English datasets:
<https://huggingface.co/datasets/HuggingFaceFW/fineweb-edu>
<https://huggingface.co/datasets/nvidia/Nemotron-CC-v2> (22B models)

Multilingual datasets:
<https://huggingface.co/datasets/togethercomputer/RedPajama-Data-V2>
<https://hpl-project.org/datasets>
<https://huggingface.co/datasets/allenai/MADLAD-400>
<https://huggingface.co/datasets/uonlp/CulturaX>

Code & math datasets:
<https://huggingface.co/datasets/bigcode/the-stack>
<https://huggingface.co/datasets/HuggingFaceTB/SmolLM-corpus>
<https://huggingface.co/datasets/HuggingFaceTB/finemath> (22B models)

Other smaller datasets were used in training, in particular during the post-training phase.

General description of other publicly available datasets not listed above:

9B models:
<https://huggingface.co/datasets/utter-project/EuroBlocks-SFT-Synthetic-1124>

22B models:
<https://huggingface.co/datasets/utter-project/EuroBlocks-SFT-2512>

Additional comments (optional): *For further details, refer to the [9B](#) and [22B](#) model technical reports.*

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

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

2.3 Data crawled and scraped from online sources

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

Was data collected from user interactions with the provider's other services or products used to train the model? ☐ Yes ☒ No

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:

Qwen2.5-Math-7B,
<https://huggingface.co/Qwen/Qwen2.5-Math-7B>

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

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

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:

In curating the training data, standard machine-readable opt-out by all websites were respected.

3.2 Removal of illegal content

General description of measures taken:

The data sources used are not expected to contain illegal content under Union law. Nevertheless, proportionate safeguards were applied, including perplexity and heuristic filtering, as well as educational score filtering using a model-based classifier.
