Public Summary of Training Content for General-Purpose AI models

This template is provided by the European Commission and required to be filled in by providers of general-purpose AI models prior to their placing on the Union market in order to comply with their obligation under Article 53 (1)(d) of Regulation (EU) 2024/1689 (AI Act).
For more information and guidance see Commission's Explanatory Notice and Template for the Public Summary of Training Content for general-purpose AI models | Shaping Europe's digital future.

Version of the Summary: *v1.5*
Last update: *17.7.2026*

1. General information

1.1. Provider identification

Provider name and contact details: ***Swiss National AI Institute*** — *https://www.swiss-ai.org/* — *llm-requests@swiss-ai.org*

Authorised representative name and contact details: *Only applicable if the provider is established outside the Union (see Article 54 AI Act).*

1.2. Model identification

Versioned model name(s):
Apertus-1.5-8B
Apertus-1.5-70B
Apertus-1.5-8B-Instruct
Apertus-1.5-70B-Instruct
https://huggingface.co/collections/swiss-ai/apertus-1.5

Model dependencies: *N/A*

Date of placement of the model on the Union market: *July 14, 2026*

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	Training data size	Types of content
☒ 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: *17 trillion tokens*	*Public text-only datasets derived mainly from web documents written in over 1000 languages, while making significant efforts were made to respect consent by right holders. We ensure training data is fully transparent and reproducible, so as to make Apertus an open-data and open-weights model. Data is filtered for robots.txt compliance and personal identification information when applicable.*

☒ Image	☐ Less than 1 million images ☒ 1Million to1 billion images ☐ More than 1 billion images	*Public image-only and image-text-pairs, gathered from permissive datasets, themselves derived mainly from open image repositories and web data. Data is filtered for robots compliance and personal identification information when applicable.*
☒ Audio	☐ Less than 10 000 hours ☒ 10 000 to 1 million hours ☐ More than 1 million hours	*Public audio-only and audio-transcripts-pairs, gathered from permissive datasets, themselves derived mainly from open initiatives (such as commonvoice) and web data.*
☐ ~~Video~~	☐ Less than 10 000 hours ☐ 10 000 to1 million hours ☐ More than 1 million hours	
☐ ~~Other~~	*Specify the modality and for each one indicate approximate size and unit of measurement*	

Latest date of data acquisition/collection for model training:	*Pre-training dataset knowledge cutoff is April 28th, 2026.*
Description of the linguistic characteristics of the overall training data:	***Languages**. The pretraining dataset includes more than **1000 languages** (1782 language-script pairs), as provided by the FineWeb-2 and FineWeb-2-HQ datasets respectively. The amount of data per language reflects the natural frequency of web data in each language, thus improving representation of many communities with languages not present in most leading LLMs yet.* ***Domain-specificity.** Emphasis has been placed on **legal** (with datasets such as Multi-Legal-Pile or Swiss-caselaw), **medical** (with datasets such as MedTrinity), as well as **science-related** (finemath) contents for this release.*
Other relevant characteristics of the overall training data:	*Selected datasets (or dataset subsets) adhere to our strict policy of full license permissiveness (excluding sharealike, non-commercial and any other restrictive licenses). Data has been rigorously filtered for respecting consent by website owners (opt-out for AI crawlers, also retroactively), remove PII, remove toxic content, and avoid verbatim memorization during model training (see below).*
Additional comments (optional):	*The Apertus tokenizer builds upon the Mistral v3 (tekken), and is used for data size statistics (see also the Apertus technical report).*

2. 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

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…*

The following large pretraining datasets were not used as is, but were additionally filtered for opt-out retrospectively, for toxicity, high quality, and other preprocessing as detailed below.

The same versions and filtering techniques were used for the datasets used both in Apertus v1 and v1.5[1] and are detailed in our first technical report. For datasets used only in v1.5, filtering and pre-processing was very similar to v1. We used the same codebase[2] and the same list of robots.txt. The same PII-removal filter was used. Slight overhauls mainly concerned technical implementation details. For robots-filtered image datasets (MINT-1T, pixmo-cap), we only excluded the sources that were within our list and had robots.txt indications (i.e. not excluding sources that were not gathered by our list).

With respect to training from a purely technical perspective, the most important change is naturally the addition of multimodal data.

In addition to scale, we focused on data quality through curation, deduplication, quality filtering and domain relevance. Our resulting data-mixture supports robust, general-purpose, and specialised capabilities while advancing open and responsible AI development.

List of large publicly available datasets:

Text Datasets

HuggingFaceFW/fineweb-2: Large-scale, high-quality multilingual web text corpus derived from filtered Common Crawl data. Serves as a primary general-purpose pretraining source for LLMs, with strong emphasis on diversity and reduced noise. License: ODC-BY. (Subject to PII removal and robots.txt filtering.)

HuggingFaceTB/dclm-edu: Curated educational web dataset from the DataComp-LM project. Provides high-signal, knowledge-rich text optimised for reasoning and factual learning in language models. License: CC-BY-4.0. (PII removal and robots.txt filtering applied.)

nvidia/Nemotron-CC-v2.1: NVIDIA-curated Common Crawl dataset (v2.1) featuring quality scoring and filtering for large-scale LLM pre-training. License: Nvidia's custom data and model license (permissive, see dataset page for terms).

HuggingFaceFW/finePDFs-edu: High-quality collection of educational, scientific, and technical PDF documents with extracted clean text. Enhances long-context and domain knowledge capabilities. License: ODC-BY. (Filtered for quality and compliance.)

joelniklaus/Multi_Legal_Pile: Specialized multilingual legal corpus aggregating statutes, case law, contracts, and regulatory texts from multiple jurisdictions. Strengthens legal reasoning and domain adaptation. License: only compliant (non-SA, non-NC) subsets of this

[1] The following datasets are concerned: dclm-edu, fineweb-2, finetranslations, finemath, MegaMath, stack-edu.
[2] Public version accessible at: https://github.com/swiss-ai/pretrain-data

compound dataset were used. (PII removal and robots.txt filtering applied.)

nvidia/Nemotron-Pretraining-Code-v1: Large-scale, curated code corpus from NVIDIA designed to boost programming, software engineering, and logical reasoning abilities.License: Nvidia's custom data and model license (permissive, see dataset page for terms).

Audio Datasets

mozilla/CommonVoice24: Mozilla's crowdsourced multilingual speech corpus with validated transcriptions across many languages and accents. A cornerstone dataset for inclusive, robust automatic speech recognition (ASR) and text-to-speech (TTS). License: CC-BY-1.0.

speechcolab/gigaspeech: Large-scale English speech recognition corpus (~10k hours) sourced from audiobooks, podcasts, and YouTube with high-quality transcriptions. Supports general-domain ASR and audio understanding. License: Apache-2.0.

MLCommons/peoples_speech: One of the largest publicly available multilingual speech datasets, containing tens of thousands of hours of diverse, real-world speech. License: CC-BY-SA / CC-BY.

facebookresearch/voxpopuli: Large multilingual speech corpus extracted from European Parliament sessions, covering numerous EU languages with aligned transcripts. Excellent for cross-lingual and parliamentary-domain speech tasks. License: CC-BY-1.0.

k2-fsa/libriheavy: Massive clean English read-speech corpus (tens of thousands of hours) built on LibriSpeech audiobooks with precise alignments. Known for high acoustic quality and utility in ASR/TTS research. License: Apache-2.0.

facebook/omnilingual-asr-corpus: Massive multilingual speech corpus spanning a wide range of languages and dialects, created to advance open ASR systems globally. License: CC BY 4.0.

Image Datasets

mlfoundations/MINT-1T: Landmark large-scale image-text dataset scaled to trillions of tokens, specifically designed to dramatically expand open-source multimodal pretraining data. License: CC-BY-4.0. (Includes PII removal and robots.txt filtering.)

UCSC-VLAA/Recap-DataComp-1B: Billion-scale image-text dataset featuring high-quality recaptions of the original DataComp-1B collection. Optimized for vision-language pretraining and detailed visual understanding. License: CC-BY-4.0.

dclure/laion-aesthetics-12m-umap: Curated 12-million-image subset of LAION focused on high aesthetic quality (via CLIP and aesthetic scoring). Popular for training visually pleasing image understanding and generation models. License: MIT.

mvp-lab/LLaVA-OneVision-1.5-Mid-Training-85M: Large-scale (85M samples) mid-training dataset for vision-language models, emphasizing instruction tuning and multimodal alignment. License: Apache-2.0.

	DeepGlint-AI/DanQing100M: Large-scale Chinese image-text pretraining dataset (100M pairs) supporting enhanced multilingual and culturally relevant visual-language capabilities. License: CC-BY-4.0. *UCSC-VLAA/MedTrinity-25M: Large medical image-text dataset (25M samples) with rich annotations and captions. Key resource for building specialized medical vision-language understanding and diagnostic assistance features. License: only compliant (non-SA, non-NC) subsets of this compound dataset were used (see full list on the dataset's page).*
General description of other publicly available datasets not listed above:	*Other smaller publicly available and permissively licensed datasets were used for the purposes described above. The exhaustive list of pre-training datasets is accessible on our GitHub. Generally, all republished datasets used for Apertus v1.5 will be made publicly available in the dedicated collection on Hugging Face.*
Additional comments (optional):	*Training data preparation code (for pre- and post-training) is entirely reproducible and is or will be made publicly available in the following repositories:* - *Version 1 text pretraining data preparation pipelines (v1.5 soon to come)* - *Version 1 text posttraining data preparation pipelines (v1.5 soon to come)* - *Version 1.5 multimodal data preparation pipelines*

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
Additional comments (optional):	-

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

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:	*We respect standard machine-readable opt-out by all websites. In addition, we remove data from websites which have recently opted out by specifying at least one of the common AI crawlers, at the time of January 2025. Crucially, we have applied such removals also retroactively in all earlier crawls since 2013, of each corresponding website present in our datasets. Pretraining and posttraining datasets were additionally filtered for licence compliance, and text datasets are processed by PII removal.*
Additional comments (optional):	*Our Acceptable Use Policy of Apertus LLM is documented Usage Policy statement. A summary of our copyright policy can be found on the model card (Legal aspects), and in our Code of Practice.*

3.2 Removal of illegal content

General description of measures taken:	*Before training, we remove toxic documents from the pretraining corpora. For text data, we do so by employing a deep learning classifier trained on top of XLM-Roberta multilingual embeddings. The classifiers were trained using the multilingual datasets provided by https://github.com/Pleias/toxic-commons . In addition to toxicity filtering, most datasets also were filtered by additional quality classifiers, which we make transparently available, and which further help to reduce problematic content. Image dataset are deduplicated and converted to appropriate formats and dimensions, among other common preprocessing practices.*

3.3. Other information (optional)

Other relevant information about data processing (optional):	*In addition to the full transparency of ensuring all training data of our models is openly available and reproducible (Apertus models being open-data and open-weights), we also employed techniques to minimize memorization or potentially remaining copyrighted content: During training, we employ the Goldfish loss technique , which disables verbatim memorization of text sequences longer than 50 tokens. More precisely, every 50th token (on average) of our pretraining data is not provided a prediction target, i.e. has no loss function, and thus breaks any verbatim memorization beyond that sequence length. We provide more detailed results on the success of this mitigation technique in the model's technical report.*