Code of Practice for the Apertus LLM (GPAI)

This document contains the Code of Practice for the Apertus LLM published by the Swiss National AI Institute (SNAI), a partnership between the two Swiss Federal Institutes of Technology, ETH Zurich and EPFL, on 17 July 2026.

The model is released under a free and open-source license that allows for the access, usage, modification, and distribution of the model, and whose parameters, including the weights, the information on the model architecture, and the information on model usage, are made publicly available. Despite the exceptions for GPAI models released under these conditions, SNAI provides this Code of Practice **voluntarily for transparency purposes and for the users' convenience**. The Swiss Federal Institutes have not signed the signature form provided by the EU AI Office.

The Apertus LLM is a general-purpose AI model without foreseeable systemic risk. Therefore, this Code of Practice comprises two (2) chapters: A) Transparency and B) Copyright Policy.

Release Version of this CoP: v1.5
Last update: 17.07.2026

A) TRANSPARENCY DOCUMENTATION (Art. 53(1)(b) EU AI Act

The following information outlines the measures regarding transparency.

1) General Information

Legal name for the model provider: Both Swiss Federal Institutes of Technology, ETH Zurich and EPFL, are cooperation partners of the Swiss National AI Institute (SNAI)

https://www.swiss-ai.org/
llm-requests@swiss-ai.org

Authorised representative name and contact details (in EU): N/A
Model family: Apertus

Versioned model name: Apertus 1.5
Release date: 14. July 2026

Union market release: 20. July 2026

Model dependencies: Apertus v1.

2) Model Properties

Model architecture: Decoder-only transformer architecture with xIELU activations, QK-Norms, Pre-Norm by RMSNorms, and rotary positional embeddings.

Input modalities: Text, audio, and image

Maximum input size: Native context length 64k tokens, extensible to 128k
Output modalities: Text-only

Maximum output size: N/A (DP dependent)

Total model size: 70 billion parameters, and 8 billion respectively

3) Methods of Distribution and Licenses

Distribution channels: Hugging Face

Model License: Apache 2.0 (January 2004), accessible at https://www.apache.org/licenses/LICENSE-2.0

Additional assets made available, incl. description of access and additional licenses:

- Data processing code:
  - github.com/swiss-ai/pretrain-data
  - github.com/swiss-ai/posttrain-data
  - github.com/swiss-ai/multimodal-data
- Model pretraining code:
  - github.com/swiss-ai/pretrain-code
- Base & instruct models as well as intermediate training checkpoints
  - huggingface.co/swiss-ai/Apertus-v1.5-8B
  - huggingface.co/swiss-ai/Apertus-v1.5-70B

4) Use

Acceptable Use Policy: See huggingface.co/swiss-ai/Apertus-v1.5-70B
Intended uses: General-purpose AI model

Type and nature of AI systems in which the general-purpose AI model can be integrated: Conversational AI Systems, AI Workflows, Research & Development Tools

Technical means for model integration: Support for common inference frameworks, such as vLLM, SGlang, Hugging Face Transformers. See model card on Hugging Face for full details.

Required hardware: N/A
Required software: N/A
Export Regulations: N/A

5) Information on Data Used for Training, Testing, and Validation

Training Data Type/Modality: Text, audio, and image

Latest date of data acquisition: Main pretraining dataset knowledge cutoff is 03/2024, while some domain-specific parts of the dataset (math) and parts of the post-training datasets have a later date of collection.

Training Data Provenance: The following large pretraining datasets derived from CommonCrawl were used, with data from 2013 onward to a knowledge cut-off mainly of March 2024 (Apertus v1), though with smaller newer documents added from 2025 and 2026 for Apertus v1.5 (the newest being from 28. April 2026). Datasets were not used in raw form but additionally filtered for opt-out retrospectively, for toxicity, high quality, and other preprocessing as detailed below.

We refer to the Apertus v1 technical report for the list of training datasets for v1. Apertus v1.5 is a continuous pretraining of v1, on the following additional large datasets. The complete list of training datasets is accessible on our GitHub

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 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).

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.

Data curation methodologies: See Technical Report.

B) COPYRIGHT POLICY (Art. 53(1)(c) EU AI Act)

The following information outlines the measures of taken to comply with copyright. In particular it identifies, and complies with, reservation of rights expressed by rightsholders.

1) Reproduction and extraction of lawfully accessible copyright-protected content

SNAI has implemented and adheres to the following measures to reduce and extract only lawfully accessible content for the training Apertus LLM:

- The pre-training data used for the Apertus LLM was obtained in/from datasets licensed from Hugging Face, Open SLR, AISHELL-1, FHNW Institute for Data Science Datasets, Common Voice, Zenodo (CERN), NDL Lab, Swisstopo,

Geoservices, Geodaten BGDI, Kaggle, Figshare, HoloAssist, Our World in Data, NASA, Smithonians, National Library of Medicine (for details please refer to chapter A, above, and specifically the dataset list in the Public Summary). SNAI respects technological denial and/or restriction of access imposed by copyright holders. No content from subscription models or paywalls was used (or circumvented for access purposes) for training of the Apertus LLM.

To the best of knowledge, the Apertus LLM was not trained on data recognised as persistently and repeatedly infringing copyright and related rights on a commercial scale by courts or public authorities in the European Union and the European Economic Area.

2) Identification and compliance with rights reservations

SNAI has identified and complied with rights reservations, including through state-of-the-art technologies and machine-readable reservations, e.g. robots.txt. The Apertus v1.5 LLM was trained on web documents crawled by CommonCrawl while respecting standard machine-readable opt-out by websites. In addition, data from websites which have recently opted out by specifying at least one of the common AI crawlers, at the time of January 2025, was removed. Crucially, such removals were also applied 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 processed by PII removal (for details please refer to chapter A, above).

3) Mitigation of the risk of copyright-infringing outputs

To mitigate the risk that a downstream AI system, into which the Apertus LLM (AI model) is integrated, generates output that may infringe copyrights, SNAI

- implemented state-of-the-art mitigation techniques to avoid verbatim memorization in the model, by the Goldfish loss technique https://arxiv.org/html/2406.10209 , which avoids verbatim memorization of text sequences longer than 50 tokens. More precisely, every 50th token (on average) of the pretraining data is not provided a prediction target, i.e. has no loss function, and thus breaks any verbatim memorization beyond that sequence length. More detailed results on the success of this mitigation technique is provided in the model's technical report. SNAI considers these measures as appropriate and proportionate technical safeguards to prevent the Apertus LLM from generating outputs that reproduce training content (for details please refer to chapter A, above); provided, however, that downstream providers remain responsible for their AI system and its (prompted) output.
- prohibits its Apertus LLM being used for copyright infringing uses (for details

please refer to chapter A, above).

4) Contact for the lodging of complaints

Affected right holders, i.e. right holders whose copyrighted material has been used for training purposes of the Apertus LLM and who believe their copyright has been infringed during this training process, may lodge a complaint to:

llm-copyright-requests@swiss-ai.org

Complaints must be sufficiently precise and adequately substantiated regarding the non-compliance of SNAI with its commitments pursuant to this chapter B and provide easily accessible information about it. SNAI will act within reasonable time from receiving a complaint in a diligent and non-arbitrary manner. SNAI reserves the right not to respond if (i) a complaint is manifestly unfounded or (ii) has already been addressed to an identical or similar complaint by the same rightsholder.

****