Public Summary of Training Content for General-Purpose AI models

Version of the Summary:	*V1*
Last update:	*01/09/2025*

1. General information

1.1. Provider identification

Provider name and contact details:	*Swiss National AI Institute, a partnership between the two Swiss Federal Institutes of Technology, ETH Zurich and EPFL* *https://www.swiss-ai.org/* *llm-requests@swiss-ai.org*
Authorised representative name and contact details:	*N/A*

1.2. Model identification

Versioned model name(s):	*Apertus-8B, https://huggingface.co/swiss-ai/Apertus-8B* *Apertus-70B, https://huggingface.co/swiss-ai/Apertus-70B* *Apertus-8B-Instruct, https://huggingface.co/swiss-ai/Apertus-8B-Instruct* *Apertus-70B-Instruct, https://huggingface.co/swiss-ai/Apertus-70B-Instruct*
Model dependencies:	*N/A*
Date of placement of the model on the Union market:	*Sept 2nd, 2025*

1.3 Modalities, overall training data size and other characteristics

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: *15 trillion tokens*	*Public text-only datasets derived mainly from web documents, in over 1000 languages. The training data is fully transparent and reproducible (Apertus is an open-data open-weights model).*

Latest date of data acquisition/collection for model training:	*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.*
Description of the linguistic characteristics of the overall training data:	*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.
Other relevant characteristics of the overall training data:	*Data has been rigorously filtered for respecting consent by website owners (opt-out for AI crawlers, also retroactively), remove PII (e-mail, IP addresses, IBAN), remove toxic content, and avoid verbatim memorization during model training.*
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

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:	*The following large pretraining datasets derived from CommonCrawl were used, with data from 2013 onward to a knowledge cutoff of March 2024 (CC-MAIN-2024-10). The datasets were not used as is but additionally filtered for opt-out retrospectively, for toxicity, high quality, and other preprocessing as detailed below.* *English datasets:* *https://huggingface.co/datasets/HuggingFaceFW/fineweb-edu (v1.0.0)* *https://huggingface.co/datasets/epfml/FineWeb-HQ* *https://huggingface.co/datasets/HuggingFaceTB/dclm-edu* *Multilingual datasets:* *https://huggingface.co/datasets/epfml/FineWeb2-HQ* *https://huggingface.co/datasets/HuggingFaceFW/fineweb-2 (v2.0.1)* *Code & math datasets:* *https://huggingface.co/datasets/bigcode/the-stack-dedup (v1.2)* *https://huggingface.co/datasets/common-pile/stackv2_edu_filtered* *https://huggingface.co/datasets/HuggingFaceTB/finemath* *https://huggingface.co/datasets/LLM360/MegaMath*
General description of other publicly available datasets not listed above:	*Other smaller publicly available and permissively text-only datasets were used in training, in particular in the post-training phase.Those were also processed by additional licence filtering and decontamination, and include:* *Wikipedia* *https://huggingface.co/datasets/HuggingFaceTB/smoltalk2* *https://huggingface.co/datasets/utter-project/EuroBlocks-SFT-Synthetic-1124* *https://huggingface.co/datasets/allenai/tulu-3-sft-olmo-2-mixture-0225* *https://huggingface.co/datasets/DataProvenanceInitiative/Commercial-Flan-Collection-Chain-Of-Thought* *In addition, the pretraining data contained small amounts of canary/poisoning data were added by https://spylab.ai/, as well as memorization detection traces from public domain project Gutenberg books.*
Additional comments (optional):	*The full training data (for pre- and post-training) is publicly available and reproducible with our provided data filtering scripts, provided here:*

https://github.com/swiss-ai/pretrain-data
https://github.com/swiss-ai/posttrain-data

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): *Providers may also disclose other relevant information on a voluntary basis, e.g. the period of data collection, size of the datasets and further details.*

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:	*In curating the training data, standard machine-readable opt-out by all websites were respected. In addition, data from websites which had recently opted out by specifying at least one of the common AI crawlers, at the time of January 2025 were 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 (e-mail, IP addresses, and IBAN) removal.*
Additional comments (optional):	*Please refer to Chapter B of the Code of Practice for the Apertus LLM: https://huggingface.co/swiss-ai/Apertus-70B-2509/blob/main/Apertus_EU_Code_of_Practice.pdf*

3.2 Removal of illegal content

General description of measures taken:	*Before training, toxic documents from the pretraining corpora were removed by employing a deep learning classifer 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 are made transparently available, and which further help to reduce problematic content.*

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 (the Apertus models being open-data and open-weights), techniques to minimize memorization or potentially remaining copyrighted content were employed: During training, the Goldfish loss technique https://arxiv.org/html/2406.10209 , 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. More detailed results on the success of this mitigation technique in the model's technical report.For detailed information, please refer to the Code of Practice for the Apertus LLM:
https://huggingface.co/swiss-ai/Apertus-70B-2509/blob/main/Apertus_EU_Code_of_Practice.pdf