☐

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: *1.1*
Last update: *23rd of November 2026*

1. General information

1.1. Provider identification

Provider name and contact details: *Aleph Alpha GmbH*

Authorised representative name and contact details: *N/A*

1.2. Model identification

Versioned model name(s): *Kolibri-1.1*

Model dependencies: *None*

Date of placement of the model on the Union market: *23rd of November 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 *Select the modalities present in the training data, to the extent that they are identifiable*	**Training data size** *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.*	**Types of content** *For each selected modality, provide a general description of the type of content that has been included in the training data.*
☑ 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:	*The text data consists primarily of English and German web documents, covering a broad range of everyday, informational and technical topics. It further includes source code, scientific and technical documents, encyclopaedic content, legal and public-sector texts such as legislative and parliamentary documents, and mathematical and STEM question-answer data. In addition, it contains synthetically generated data.*

☐

☐ Image	☐ Less than 1 million images ☐ 1Million to1 billion images ☐ More than 1 billion images	*N/A*
☐ Audio[1]	☐ Less than 10 000 hours ☐ 10 000 to1 million hours ☐ More than 1 million hours	*N/A*
☐ Video	☐ Less than 10 000 hours ☐ 10 000 to1 million hours ☐ More than 1 million hours	*N/A*
☐ Other	*Specify the modality and for each one indicate approximate size and unit of measurement*	

Latest date of data acquisition/ collection for model training:	*06/2026* *The model is not continuously trained on new or dynamic data after this date.*
Description of the linguistic characteristics of the overall training data:	*The training data is predominantly in English and German, both official EU languages. The pre-training corpus consists of approximately 62.5% English, 23.9% German and 13.6% source code. Other languages were not deliberately included.*
Other relevant characteristics of the overall training data:	*The overall training data is mostly sourced from the public Internet and publicly-available data and is hence general-domain as a baseline. However, the data also contains specific data representing the European and especially the German region. Such data is mainly derived from European and German official texts.*
Additional comments (optional):	*The data size was measured based on the tokenizer the model was trained on. The tokenizer is released alongside the model weights.*

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

This Section requires information about datasets that were used to train the model and which have been compiled by a third party, are made available publicly for free, and are readily downloadable as a whole or in predefined chunks, such as datasets and collections available on public repositories and online platforms, specialised websites, or snapshots of common crawl. The public availability of the datasets for free does not mean that the content at issue is necessarily free of rights since it may be subject to licensing arrangements or conditions of use (e.g., certain free and/or open licenses may determine the scope of the uses, including prohibiting uses relating to model training).

A dataset is considered to be "large" if the total data size for any one of the modalities contained in the dataset exceeds 3% of the size of all publicly available datasets for that modality used for training. The size of the dataset should be based on its size after pre-processing (for example filtering), and without splitting the dataset to prevent reporting circumvention.

[1] Excluding audio that is part of video, as this should be reported under the "video" modality instead. Furthermore, the Commission understands the modality of 'audio' to include 'speech'.

☐

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:

All of the listed datasets were not used as is. After source-level and legal review (cf. also section 3 below) especially on opt-out and license adherence, only parts of the datasets or subsets (or parts of subsets) have been used after cleaning and filtering. Filtering entails dedicated filtering against illegal, unethical, infringing, harmful and pirated content as well as deduplication methods and sampling. Additionally the datasets undergo dedicated removal of personally identifiable information through filtering (cf. section 3.2)

- *Common Crawl (https://commoncrawl.org):*
- *Nemotron-CC-v2 (https://huggingface.co/datasets/nvidia/Nemotron-CC-v2)*
- *Nemotron-CC-v2.1 (https://huggingface.co/datasets/nvidia/Nemotron-CC-v2.1)*
- *Nemotron Pretraining SFT v1 https://huggingface.co/datasets/nvidia/Nemotron-Pretraining-SFT-v1*
- *HPLT 3.0 (https://huggingface.co/datasets/HPLT/HPLT3.0):*
- *FinePDFs (https://huggingface.co/datasets/HuggingFaceFW/finepdfs)*
- *The Stack v3 (https://huggingface.co/datasets/HuggingFaceCode/stack-v3-train)*
- *Nemotron Pretraining Specialized v1 https://huggingface.co/datasets/nvidia/Nemotron-Pretraining-Specialized-v1*

General description of other publicly available datasets not listed above:

Other datasets are text-only (English & German) and contain (i) curated data from targeted domains, such as knowledge bases and governmental, legal, source-code, mathematical & technical websites & publications, e.g. political proceedings and German legislation, (ii) synthetically generated data published by third parties, generated from web content, encyclopedic content, source code and existing fine-tuning datasets, (iii) chat, instruction-following, mathematics, science, code, tool-use and safety datasets used in supervised fine-tuning, and publicly available task datasets used in reinforcement learning environments. The most recent end date of the data collection is June 2026.

Additional comments (optional):

All datasets were not used as is. After source-level and legal review (cf. also section 3 below) especially on opt-out and license adherence, only parts of the datasets or subsets (or parts of subsets) have been used after cleaning and filtering as well as dedicated removal of personally identifiable information through filtering (cf. section 3.2)

☐

2.2 Private non-publicly available datasets obtained from third parties

This Section requires information about private non-publicly available datasets of third parties that are not publicly available and not disclosed under Section 2.1. These include:

1) *datasets for which transactional commercial licensing agreements were concluded between the provider and the rightsholders or their representatives, including by collective management organisations and legitimate content aggregators who have the right to collectively license works on behalf of rightsholders (Section 2.2.1);*
2) *other private datasets obtained through data intermediaries, non-publicly available databases and datasets of third parties for which transactional commercial licenses have not been concluded with rightsholders or their representatives (Section 2.2.2).*

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
If yes, specify the modality(ies) of the content covered by the datasets concerned:	☐ Text ☐ Image ☐ Video ☐ Audio ☐ Other *If so, please specify…*

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
If yes, specify the modality(ies) of the content covered by the datasets concerned:	☐ Text ☐ Image ☐ Video ☐ Audio ☐ Other *If so, please specify…*
If publicly known, list private datasets obtained from other third parties:	*N/A*
General description of non-publicly known private datasets obtained from third parties	*N/A*
Additional comments (optional):	*N/A*

2.3 Data crawled and scraped from online sources

This Section requires information about crawled, scraped data, or otherwise compiled from online sources directly by the provider of the model or on their behalf (i.e. excluding publicly available datasets already compiled by third parties and made available on platforms such as common crawl that are covered under Section 2.1).

Were crawlers used by the provider or on behalf of?	☐ Yes ☑ No
If yes, specify crawler name(s)/ identifier(s):	*N/A*
Purposes of the crawler(s):	*N/A*

☐

General description of crawler behaviour:	*N/A*
Period of data collection:	*N/A*
Comprehensive description of the type of content and online sources crawled:	*N/A*
Type of modality covered:	☐ Text ☐ Image ☐ Video ☐ Audio ☐ Other
Summary of the most relevant domain names crawled:	*N/A*
Additional comments (optional):	*N/A*

2.4 User data

This Section requires information about user data collected by all services and products of the provider, including through mail services, social media platforms, content platforms or interaction with the providers' AI models and/or systems. This does not cover data licensed by users based on commercial transactional agreements described in Section 2.2.1., or customer data to fine-tune models for specific purposes.

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
If yes, provide a general description of the provider's services or products that were used to collect the user data:	*N/A*
Type of modality covered:	☐ Text ☐ Image ☐ Video ☐ Audio ☐ Other *If so, please specify…*
Additional comments (optional):	*N/A*

2.5 Synthetic data

This Section requires information about synthetic data created by or on behalf of the provider for training the model directly on the outputs of another AI model, in particular through model distillation or model alignment (e.g. AI feedback through reinforcement learning). This does not include the use of AI models to clean or enrich data (e.g. AI-generated metadata to enrich or modify a dataset, such as creating depth maps or text descriptions of images). In case this concerns publicly available datasets as described in Section 2.1, these should be reported in that Section of the Template. In case this concerns synthetic datasets created by third parties on behalf of the provider, these should be reported in this Section of the Template instead of in Section 2.2.2.

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:	*- Gemma-4-26B-A4B (https://huggingface.co/google/gemma-4-26B-A4B)* *- Qwen3-32B (https://huggingface.co/Qwen/Qwen3-32B)* *- Qwen3.8-27B (https://huggingface.co/Qwen/Qwen3.8-27B)* *- Mistral-Nemo-12B (https://huggingface.co/mistralai/Mistral-Nemo-Instruct-2407)* *- GLM-5.2-fp8 (https://huggingface.co/zai-org/GLM-5.2-FP8)* *- GLM-5.3 (https://huggingface.co/zai-org/GLM-5.3)* *- Kimi-K2.6 (https://huggingface.co/moonshotai/Kimi-K2.6)* *- DeepSeek-V4.1-Flash (https://huggingface.co/deepseek-ai/DeepSeek-V4.1-Flash-1017)* *- DeepSeek-V4.1-Flash-Abliterated (https://huggingface.co/s-zaizen/DeepSeek-V4.1-Flash-Abliterated)* *- Qwen3.8-27B-Uncensored-Heretic (https://huggingface.co/llmfan46/Qwen3.8-27B-Uncensored-Heretic-Native-MTP-Preserved)*
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*
Additional comments (optional):	*The abliterated models were used to generate examples of harmful requests that our model is then trained to recognise and refuse.*

2.6 Other sources of data

This Section requires information about data that does not fall under any of the categories in the previous Sections, for example data collected from offline sources, self-digitised media (e.g., digitised analog text context, images), datasets labelled by humans commissioned by the provider, or human generated data through reinforcement learning.

Have data sources other than those described in Sections 2.1 to 2.5 been used to train the model?	☐ Yes ☑ No
If yes, provide a narrative description of these data sources and the data:	*N/A*
Additional comments (optional):	*N/A*

☐

3. Data processing aspects

3.1. Respect of reservation of rights from text and data mining exception or limitation

This Section concerns measures implemented by the provider to identify and comply with the reservation of rights from the text and data mining (TDM) exception or limitation expressed pursuant to Article 4(3) of Directive (EU) 2019/790, as outlined in the copyright policy put in place by the provider in accordance with Article 53(1)(c) AI Act.

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 do not employ our own webcrawlers nor have we employed webcrawlers on our behalf.* *We use data from third parties. Such data was collected by these third parties, e.g. by deploying web crawlers, and thereafter made available. We assess such datasets and only use them for the development of our models if we have information that these third parties (or entities acting on their behalf) deploy web crawlers which respect instructions expressed via robots.txt.*
Additional comments (optional):	*When using data compiled by third parties, we implement a number of additional measures in line with our copyright policy. These measures include manually assessing such third-party dataset before model training commences. By using information available to us, we investigate on the third-parties' data sources and assess potential indications on whether the party compiling the data (1) used copyright infringing sources, (2) circumvented technological measures as set out in Section 1.1. of the General-Purpose AI Code of Practice; and (3) ignored rights reservations expressed in an appropriate machine-readable format. As a result of such assessment, we exclude from training and development activities data which was presumably collected in these aforementioned ways.*

☐

3.2 Removal of illegal content

This Section concerns measures taken to avoid or remove illegal content under Union law from the training data (such as blacklists, keywords, and model-based classifiers), without requiring disclosure of specific details about the provider's internal business practices or trade secrets. Such measures are advisable if the training data is likely to include illegal or unlawful content under Union law, in particular child sexual abuse material and terrorist content and the non-authorised use of material protected by intellectual property rights. Such measures do not include data selection practices, for example to increase the capability of the model.

General description of measures taken:

In a filtering stage we remove documents from our dataset whose origin matched one of several blocklists. Such blocked sources may contain illegal, unethical, infringing, harmful and pirated material.

3.3. Other information (optional)

Other relevant information about data processing (optional):

A language-independent data preprocessing stage was implemented to remove personally identifiable information. Data-sources targeted at personally identifiable information have been entirely excluded from training.