Template for the 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:	*Version of the summary, with link(s) to previous versions where applicable*
Last update:	*Click or tap to enter a date.*

1. General information

1.1. Provider identification

Provider name and contact details:	*Replace this with your response…*
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):	*Provide the unique identifier(s) for the model(s) or model version(s) covered by this Summary (e.g. Llama 3.1-405B). In accordance with point 30 of the Commission Explanatory Notice to the Template, the same Summary may be used for different model(s) or model version(s) provided the content of their respective Summaries is identical. Where available, provide link(s) to additional publicly available documentation, such as the model card, for the model(s) or model version(s).*
Model dependencies:	*If the model is the result of a modification, including fine-tuning, of one or more general-purpose AI models already placed on the Union market, specify the model (version) name(s) of that/those models and provide a link to their Summary(ies) where available.*
Date of placement of the model on the Union market:	*Indicate the date on which the model was placed on the Union market (including the dates each model (version(s)) was placed on the market, if the Summary applies to more than one model or version (see point 30 of the Commission Explanatory Notice to the Template).*

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: *Replace this with your response…*	*Examples of possible types of content include fiction and non fiction text, scientific text, press publications, legal and official documents, social media comments, source code.*
□ Image	□ Less than 1 million images □ 1Million to1 billion images □ More than 1 billion images	*Examples of possible types of content include photography, visual art works, infographics, social media images, logos, brands.*
□ Audio	□ Less than 10 000 hours □ 10 000 to1 million hours □ More than 1 million hours	*Examples of possible types of content include musical compositions and recordings, audiobooks, radio shows and podcasts, private audio communication.*
□ Video	□ Less than 10 000 hours □ 10 000 to1 million hours □ More than 1 million hours	*Examples of possible types of content include music videos, films, TV programmes, performances, video games, video clips, journalistic videos, social media videos.*
□ Other	*Specify the modality and for each one indicate approximate size and unit of measurement*	

Latest date of data acquisition/ collection for model training:	*Indicate the latest date when data was collected/obtained for the model training: MM/YYYY Additionally, indicate if the model is continuously trained on new or dynamic data after this date.*
Description of the linguistic characteristics of the overall training data:	*Where applicable, describe the languages covered by the training data (e.g., text, videos or speech), focusing in particular on EU official languages.*
Other relevant characteristics of the overall training data:	*Where such information is readily available and in so far as it is relevant and practicable, describe other relevant characteristics of the overall training data, such as national/regional or demographic specificities of the training data.*
Additional comments (optional):	*Providers may also disclose other relevant information on a voluntary basis (e.g. the compression or tokenization methodologies applied for the data size calculation, the sampling frequency/rate plays for audio or video content).*

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.

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 <u>large</u> publicly available datasets:	*For each large dataset, provide the identifier/name of the dataset and a link through which the dataset can be accessed. If a link is not available, provide a general description of the dataset, including the approximate start and end dates of the data collection if known (otherwise indicate "not known"). If only part of the datasets has been used for the training, indicate the general approach to selecting those parts.*
General description of other publicly available datasets not listed above:	*For other publicly available datasets that are not listed above, provide a general description of their content. The description could include indication of: (i) the types of modality (e.g. text, images), (ii) nature of the content (e.g. personal data, copyright protected content, machine generated data such as Internet of Things or synthetic data), (iii) its linguistic characteristics, where applicable. (iv) the approximate start and end dates of the data collection if known (otherwise indicate "not known").*
Additional comments (optional):	*Providers may also disclose other relevant information on a voluntary basis, e.g. size of the datasets and other relevant details.*

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 <u>not</u> 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:	*If publicly known, list the identifiers/names of the main private datasets from third parties that are not licensed as described in Section 2.2.1 and that are used to train the model, and provide links to relevant information, where available.*
General description of non-publicly known private datasets obtained from third parties	*For those private datasets used to train the model that are not publicly known and whose identifiers are not listed above, provide a general description of their content. The description should indicate (i) the modalities (e.g,, text, images), (ii) nature of the content (e.g., personal data, copyright protected content, machine generated data such as Internet of Things or synthetic data) and (iii) its linguistic characteristics, where applicable.*
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

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):	*Replace this with your response…*
Purposes of the crawler(s):	*Replace this with your response…*
General description of crawler behaviour:	*For example, this includes respect of captchas, password protected websites and paywalls, respect of robot.txt and other protocols, while crawling.*
Period of data collection:	*From MM/YYYY to MM/YYYY*

Comprehensive description of the type of content and online sources crawled:	*Provide a comprehensive description of the type of content crawled, including its geographical, linguistic or demographic characteristics, as well as an indication of the type of websites scraped (e.g. news, blogs, social media, forums, community websites, other user-generated content platforms, websites of cultural heritage institutions, educational sites, government portals, personal blogs, streaming, gaming platforms, online TV platforms, synthetic data libraries, etc).*
Type of modality covered:	□ Text □ Image □ Video □ Audio □ Other *If so, please specify…*
Summary of the most relevant domain names crawled:	*In so far as any content from internet domains has been crawled or scraped and used for the training of the model, provide a list of those most relevant internet domains names (top and second-level domain, e.g. "example.com") by listing the top 10 % of all domain names determined by the size of the content scraped (in a representative manner across all modalities where applicable). Small and medium-sized enterprises (SMEs), including start-ups, should disclose top 5% of all domain names or 1 000 internet domain names, whichever is lower. You can provide this list for instance as a downloadable file or provide here information on how to access it.*
Additional comments (optional):	*Providers may also disclose other relevant information on a voluntary basis, for instance more domain names than those required in the list above and/or URLs and the sources of individual works.*

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:	*Replace this with your response…*
Type of modality covered:	□ Text □ Image □ Video □ Audio □ Other *If so, please specify…*
Additional comments (optional):	*Providers may also disclose other relevant information on a voluntary basis.*

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:	*Specify the name of the general-purpose AI model(s) and provide a link to their Summary(ies) where available.*
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:	*Provide information about other AI models used to generate synthetic data, including provider's own models if not available on the market. This includes a general description of the model training data if known and in so far as this may be needed for the exercise of the rights of parties with legitimate interests and to avoid circumvention of the disclosure obligations in the other Sections of the Template.*
Additional comments (optional):	*Providers may also disclose on a voluntary basis other relevant information.*

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:	*Replace this with your response…*
Additional comments (optional):	*Providers may also disclose other relevant information on a voluntary basis.*

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:

Replace this with your response…

Additional comments (optional):

Providers may also disclose other relevant information on a voluntary basis. Providers are also encouraged to disclose a summary of their copyright policy under Article 53(1)(c) AI Act, if made publicly available (e.g., by providing a link to the relevant web-

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: *Replace this with your response…*

3.3. Other information (optional)

Other relevant information about data processing (optional):

Providers are also encouraged to disclose on a voluntary basis other relevant information about relevant data processing aspects and measures taken before or after the training of the model that is relevant for the respect and exercise of rights protected under Union law.