

Template for the Public Summary of Training Content for General-Purpose AI models required by Article 53 (1)(d) of Regulation (EU) 2024/1689 (AI Act)

Version of the Summary:

1.0

Last update:

24/09/2026

1. General information

1.1. Provider identification

Provider name and contact details:

Ant Intelligence (Hang Zhou) Technology Co., Ltd

Authorised representative name and contact details:

Not Applicable.

1.2. Model identification

Versioned model name(s):	Ming-UniAudio-16B-A3B
Model dependencies:	NA
Date of placement of the model on the Union market:	2025-09-29

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 <i>Select the modalities present in the training data, to the extent that they are identifiable</i>	Training data size <i>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.</i>	Types of content <i>For each selected modality, provide a general description of the type of content that has been included in the training data.</i>
☒ Text	☐ Less than 1 billion tokens ☐ 1billion to10 trillions tokens ☒ More than 10 trillions tokens Alternatively, specify the approximate size in a different measurement unit:	Trained on a large-scale, multilingual mixture of publicly available data, data accessed through partnerships, synthetic data, and human generated text, including general web content, reference materials, technical documentation, source code, and other diverse textual content
☐ Image	☐ Less than 1 million images ☐ 1Million to1 billion images	

☒ Audio ¹	☐ More than 1 billion images ☐ Less than 10 000 hours ☐ 10 000 to1 million hours ☒ More than 1 million hours	The model is trained on a large-scale, multilingual mixture of publicly available audio data, data accessed through partnerships, synthetic data, and human generated audio corpus,
--	---	---

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’.

		including general web content, reference materials, and other open-source datasets
☐ Video	☐ Less than 10 000 hours ☐ 10 000 to 1 million hours ☐ More than 1 million hours	
☐ Other	<i>Specify the modality and for each one indicate approximate size and unit of measurement</i>	

Latest date of data: acquisition/collection for model training:	Approximately Jul 2025.
Description of the linguistic characteristics of the overall training data:	Multilingual, with strong English / Chinese coverage, including substantial representation across EU official languages and other languages.
Other relevant characteristics of the overall training data:	The overall training corpus is designed for a text-output model. It aims to provide broad topical, geographic, linguistic, and format coverage. The dataset has been curated with the objective of maximizing representativeness and robustness.
Additional comments (optional):	NA

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 (please specify)
---	--

List of <u>large</u> publicly available datasets:	The training data includes text and audio clips from Common Crawl: Modality – text/audio; Nature of content – web-crawled text, video or audio data covering news, blogs, forums, government websites and other web content; Linguistic characteristics – primarily English, includes multilingual content.
General description of other publicly available datasets not listed above:	Other publicly available datasets used for training include academic paper repositories, filtered web-crawled data, programming Q&A community content, and multi-source diverse text/video/audio corpora. (i) The modality of all such datasets is text/audio. (ii) The nature of the content includes academic paper and metadata across disciplines such as physics, mathematics, computer science, quantitative biology, quantitative finance, statistics, and electrical engineering; cleaned and deduplicated web-crawled text/audio data derived from CommonCrawl or other open source datasets; programming-related technical questions, answers, code snippets, and user comments from developer community platforms; and large-scale diverse text/image/audio data compiled from web content, academic papers, books, code, and social media. (iii) Linguistic characteristics are primarily Chinese and English, with portions of multilingual content included. The data collection period is approximately 03/2024 to 07/2025.
Additional comments (optional):	NA

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 (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 (please specify)
If publicly known, list private datasets obtained from other third parties:	

General description of non-publicly known private datasets obtained from third parties	
Additional comments (optional):	

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):	Ant Spider
Purposes of the crawler(s):	Collecting publicly available information from the internet for model training.
General description of crawler behaviour:	We implement processes to respect rights reservations and opt-out signals relevant to text and datamining before and during data collection, placing great emphasis on compliance with matters relating to intellectual property, trade secrets, and personal privacy. The collection process is evaluated to verify alignment with robots.txt instructions and other standard web protocols, and is not designed to circumvent captchas, paywalls, or password-protected content. It is prohibited to employ collection methods that may cause adverse effects on the systems of the crawled parties.
Period of data collection:	Approximately 03/2024 to 07/2025.
Comprehensive description of the type of content and online sources crawled:	Crawled content includes a broad range of publicly accessible online material, including encyclopedic reference, educational, scientific, technical, government and institutional, financial disclosure, and general-interest content. Crawled sources include web page text/image/video/audio, open-source code, academic papers and metadata, scholarly literature indexes, encyclopedic knowledge entries, image-text paired data, biomedical literature, and listed company information disclosure announcements, along with associated metadata such as titles, descriptive text, and captions. In terms of geographic characteristics, the data sources cover a global scope, with mainland China and English-speaking countries as primary sources; .com and .cn country codes are prominent, reflecting a broad geographic range. In terms of linguistic characteristics, the content is primarily in Chinese and English, with multilingual content also included. In terms of website types, the sources encompass encyclopedic user-generated content platforms, code hosting platforms, academic preprint repositories, open scholarly literature databases, open dataset hosting platforms, government-designated information disclosure portals, and biomedical literature databases.
Type of modality covered:	☒ Text ☐ Image ☐ Video ☒ Audio ☐ Other (please specify)
Summary of the most relevant domain names crawled:	The top 10% of all domains crawled and used to train the model is a blend of publicly available web content platforms, code hosting platforms, academic repositories, encyclopedic platforms, and region-specific portals. Overall, the crawl is skewed towards a handful of large platforms and hosts which was filtered as described in Section 3.2. Crawled data encompasses a broad range of publicly accessible online content across educational, government, legal, research, financial, and encyclopedic sectors. The crawled content primarily includes web page text/image/video/audio, open-source code, academic papers and metadata, scholarly literature indexes, encyclopedic knowledge entries, image-text paired data, biomedical literature, and listed company

	information disclosure announcements. In terms of geographic characteristics, the data sources cover a global scope, with mainland China and English-speaking countries as primary sources. In terms of linguistic characteristics, the content is primarily in English and Chinese. Country codes like .com, .cn are prominent and reflect broad geographic range of the sourced images and audio clips.
Additional comments (optional):	

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:	
Type of modality covered:	☐ Text ☐ Image ☐ Video ☐ Audio ☐ Other (please specify)
Additional comments (optional):	

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 (please specify)
If yes, specify the general-purpose AI model(s) used to generate the synthetic data if available on the market:	We generate synthetic data using previously available generative AI models from Ling-lite, Ming-lite-omni, Ming-lite-omni-1.5. Synthetic data generated by these models include examples for instruction following, reasoning, coding, writing, knowledge.

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:	NA
Additional comments (optional):	

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:	
Additional comments (optional):	

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 place great emphasis on compliance with matters relating to intellectual property, trade secrets, and personal privacy. For publicly accessible web data, please refer to the descriptions in Section 2.3. For non-public data procured from third parties, we require the third parties to ensure the legality of the data sources and that they have sufficient rights to transfer the data, and that the data provided does not infringe upon the lawful rights of others (including intellectual property, trade secrets, and personal information).
Additional comments (optional):	NA

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:

We apply preprocessing and screening measures intended to avoid or remove illegal content from training data, including the following areas: child sexual abuse material, terrorist content & hate speech;

cyberviolence & harassment; intellectual property violations; illegal goods & services & substances; financial scams & fraud, etc.
Measures include automated filtering, keyword-based rules, hash-matching, and model-based classifiers to help identify and exclude unlawful material.

3.3. Other information (optional)

Other relevant information about data processing (optional):	We use advanced data filtering processes to reduce the presence of PII.
--	---