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: *2.0*
Last update: *19/10/2026*

1. General information

1.1. Provider identification

Provider name and contact details: *Nanonoble Pte. Ltd.*

Authorised representative name and contact details: *Prighter GmbH*

1.2. Model identification

Versioned model name(s): *MiniMax-M3; MiniMax-M3.1-Flash*

Model dependencies: *None.*

Date of placement of the model on the Union market: *01/06/2026;19/10/2026*

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: *Total pre-training corpus exceeds 1x10^14 (100+ trillion) tokens across interleaved text, image and video; text is the dominant modality.*	*Trained on a large-scale, multilingual mixture of publicly available data, data from partnerships, synthetic data, and human-generated text, this includes general web content, reference materials, technical documentation, source code, and other text, all carefully curated and filtered for quality and safety.*
☒ Image	☐ Less than 1 million images ☐ 1Million to1 billion images	*Trained on a broad corpus of images—sourced from publicly available channels, partnerships, and synthetic generation—*

	☒ More than 1 billion images	*spanning natural scenes, objects, people, and diagrams, with rigorous filtering to ensure quality, safety, and robust multimodal understanding.*
☐ 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	*Trained on publicly available, multi-domain audiovisual content—including spoken, narrated, instructional, and general-interest material—represented in part through transcripts, captions, or metadata, and thoroughly filtered for quality and safety.*
☐ Other	*Specify the modality and for each one indicate approximate size and unit of measurement*	

Latest date of data acquisition/collection for model training:	*The data used to train the Models includes differentdatasets from varying time periods, with some datacollected no later than January 2026.*
Description of the linguistic characteristics of the overall training data:	*Multilingual, with English coverage and substantial representation across EU official languagesand other languages from around the world.*
Other relevant characteristics of the overall training data:	*The overall training corpus is designed for a text-output model with multimodal understanding capabilities. Itincludes written materials, text-bearing images, andaudiovisual content used through captions, transcripts,metadata, or related text representations. The corpusaims to provide broad topical, geographic, linguistic, andformat coverage.*
Additional comments (optional):	*N/A.*

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:	*The training data for MiniMax-M3 includes text from Common Crawl.*
General description of other publicly available datasets not listed above:	*Other publicly available datasets include broad,multi-domain text and image datasets made available bythird parties through public repositories, onlineplatforms, and specialized websites, including referencematerials, scientific and technical content, source code,image-text datasets, and speech or audiovisual datasetsdistributed with captions, transcripts, metadata, orrelated text. These datasets are global in scope,multilingual, and subject to preprocessing such asquality filtering, deduplication, and safety filteringbefore training. We use advanced data filtering processesto reduce personal information from training data.*
Additional comments (optional):	*When processing public web datasets, MiniMax identifies and respects copyright reservations and opt-out signals such as robots.txt rules for our crawler on all listed domains. We also reference the USTR Notorious Markets List to exclude data from sites with repeated copyright infringements.*

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 ☒ Other (see below)
If yes, specify the modality(ies) of the content covered by the datasets concerned:	☒ Text ☒ Image ☒ Video ☐ Audio ☐ Other *If so, please specify…*
Additional comments (optional):	*MiniMax establishes extensive partnerships with third parties. Such collaborations may cover rights to display partners' content within our products and access to non-public materials including archives and metadata.*

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	☒ Yes ☐ No

obtained from providers of private databases, or data intermediaries?	
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	*MiniMax collaborates with third parties to obtain multi-domain, multi-scenario data for model improvement, with all data access conducted in full compliance with relevant laws.*
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):	*MINIMAX_UA*
Purposes of the crawler(s):	*MINIMAX_UA is used to crawl content that may be used to train MiniMax's generative AI foundation models and make them more useful and safe.*
General description of crawler behaviour:	*MINIMAX_UA respects website boundaries: it follows robots.txt instructions—such as opting out of AI training—and is built without the ability to bypass paywalls, captchas, or login screens.*
Period of data collection:	*April 2025 - December 2025*
Comprehensive description of the type of content and online sources crawled:	*Crawled data encompasses a wide spectrum of publicly available online materials—including reference, educational, scientific, technical, governmental, institutional, and general-interest content. These sources comprise text, images, and associated metadata like captions, alt text, transcripts, and descriptive text, all of which undergo rigorous quality and safety filtering prior to model training.*
Type of modality covered:	☒ Text ☒ Image ☒ Video ☐ Audio ☐ Other *If so, please specify…*
Summary of the most relevant domain names crawled:	*The most relevant crawled source domains include academic, research, patent, and other technicalrepositories, legal and government resources,document-hosting and sharing services, community and general-interest sites, and region-specific portals.*
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:	*Subject to users' privacy preferences, control options, opt-out requests and our official privacy policies, MiniMax may leverage user interaction data generated from our AI products to optimize our foundation models.We implement rigorous data filtering workflows to minimize personal identifiable information contained within all training datasets.*
Type of modality covered:	☒ Text ☒ Image ☒ Video ☐ Audio ☐ Other *If so, please specify…*
Additional comments (optional):	*Advanced de-identification and filtering mechanisms are applied to strip out personal data prior to model training. Users may submit privacy adjustment requests through our official product channels to manage how their interaction data is utilized. Relevant guidance on user privacy controls can be viewed inside our product client.*

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:	*MiniMax generated synthetic data using its own general-purpose AI models, including MiniMax-M2.7.*
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:	*We adopt self-developed internal dedicated models and cooperate with third-party model providers to generate synthetic data to supplement scarce training resources across various fields, languages, tasks and multimodal formats. The synthetic data covers instruction following, logical reasoning, code generation, multimodal comprehension, safety alignment and model evaluation scenarios.*
Additional comments (optional):	*N/A*

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:

MiniMax and our cooperative suppliers produce customized synthetic data to boost model performance across diverse tasks. We collaborate with industry professionals to generate datasets that reflect real-world professional scenarios, enabling our models to deliver better support for practical work demands.

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:

Rights honoured by the provider or, as applicable, by third parties from which datasets have been obtained: MiniMax fully respects all intellectual property rights and website access rules when collecting public data for our generative foundation models. Our web crawler will not bypass CAPTCHA verification, paid content paywalls, nor attempt to access password-locked private content.

Additional comments (optional):

N/A

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:	*MiniMax carries out multi-layer preprocessing and screening procedures to detect and eliminate content violating relevant laws from training datasets. We deploy automated filtering, keyword matching rules, hash comparison and AI classification models to identify and remove all illegal materials.*

3.3. Other information (optional)

Other relevant information about data processing (optional):	*N/A*