Public Summary of Training Content for Xiaomi MiMo V2.6 Family Models

Version of the Summary: *v1*
Last update: *22/09/2026*

1. General information

1.1. Provider identification

Provider name and contact details:	*Xiaomi Technologies Singapore Pte. Ltd.* *1 Fusionopolis Link #04-02/03 Nexus@one-north, Singapore 138542*
Authorised representative name and contact details:	*N/A*

1.2. Model identification

Versioned model name(s):	*Xiaomi MiMo-V2.6-Pro；Xiaomi MiMo-V2.6-Flash*
Model dependencies:	*Xiaomi MiMo-V2.5 Series*
Date of placement of the model on the Union market:	*22/09/2026*

1.3 Modalities, overall training data size and other characteristics

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…*	*The model was trained on text data covering a variety of types, including general web content, technical documentation, synthetic text, and code, with a deliberate shift toward data exhibiting long-range dependencies such as long-form web content and curated code corpora, all filtered for quality and safety.*
☒ Image	☐ Less than 1 million images ☐ 1Million to1 billion images ☒ More than 1 billion images	*The model was trained on image data covering a variety of types, including interleaved text and images from websites, synthetic images， computer graphics， all filtered for quality and safety.*
☒ Audio[1]	☐ Less than 10 000 hours ☐ 10 000 to1 million hours	*The model was trained on audio data covering a variety of types, covering human speech, environmental sounds, and other*

	☒ More than 1 million hours	*general audio content. The dataset spans a broad range of languages, speakers, acoustic environments, recording conditions, and audio durations, all filtered for quality and safety.*
☒ Video	☐ Less than 10 000 hours ☐ 10 000 to1 million hours ☒ More than 1 million hours	*The model was trained on video data drawn primarily from online videos spanning a wide range of domains, genres, and durations, with a deliberate shift toward richly annotated content such as fine-grained, temporally grounded event descriptions and higher-level video analyses of narrative and stylistic structure.*
☐ Other	*Specify the modality and for each one indicate approximate size and unit of measurement*	*N/A*

Latest date of data acquisition/collection for model training:	*The data used to train the models includes different datasets from varying time periods, with some data collected no later than January 2026.*
Description of the linguistic characteristics of the overall training data:	*Multilingual, with strong English coverage and substantial representation across EU official languages and 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. The corpus aims to provide broad topical, geographic, linguistic, and format coverage.*
Additional comments (optional):	*N/A*

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 training corpus draws on publicly available repositories, notably Common Crawl and curated datasets from GitHub and Hugging Face.*
General description of other publicly available datasets not listed above:	*The training data also draws on publicly available multimodal datasets provided by third parties through public repositories, online platforms, and domain-specific websites. These datasets include text, image-text pairs, and speech or video content with associated captions, transcripts, or metadata. The datasets are international in scope and multilingual. All such data is preprocessed through quality assessment, duplicate removal, and safety review.*
Additional comments (optional):	*N/A*

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
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	*We obtain private datasets from third-party providers and data intermediaries for the purpose of improving model performance. These datasets cover multiple modalities, including text, images, video, and audio, and span diverse domains and content types. All data is accessed and used in accordance with applicable laws.*
Additional comments (optional):	*N/A*

2.3 Data crawled and scraped from online sources

Were crawlers used by the provider or on behalf of?	☒ Yes ☐ No
If yes, specify crawler name(s)/identifier(s):	*XiaomiMiMoBot*
Purposes of the crawler(s):	*The XiaomiMiMoBot crawler collects web content that could be used to the model's training.*
General description of crawler behaviour:	*The XiaomiMiMoBot crawler strictly adheres to established web standards and ethical scraping policies. It automatically parses and respects robots.txt exclusion protocols, NOINDEX/NOFOLLOW directives, and HTTP header restrictions. It explicitly refrains from bypassing paywalls, login/authentication systems, password-protected gateways, or CAPTCHA mechanisms. Crawling frequency is dynamically rate-limited to avoid operational burden or DoS impacts on target hosts. Automated mechanisms are deployed to recognize and honor machine-readable Text and Data Mining (TDM) opt-outs..*
Period of data collection:	*Approximately December 2024 to January 2026*
Comprehensive description of the type of content and online sources crawled:	*The collected content encompasses a multi-lingual, multi-domain, and global cross-section of publicly accessible internet resources. Key source categories include open-access research databases, technical documentation, public forums and so on.*

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, technical repositories, legal and government resources, document-sharing platforms, and community sites. Top-level domains such as .com and .net are included alongside sites from a range of different countries.*
Additional comments (optional):	*N/A*

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
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

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:	*Xiaomi MiMo generated synthetic data using its own general-purpose AI models.*
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:	*Xiaomi MiMo uses internal models to generate synthetic data for targeted training, especially in domains, languages, tasks, or formats where data is scarce and includes instruction following, reasoning, coding, multimodal understanding, safety, and evaluation scenarios.*
Additional comments (optional):	*N/A*

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

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

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:	*Before and during data collection, Xiaomi honours machine-readable reservations of rights relevant to text and data mining. Opt-out signals are recognised and respected by Xiaomi's crawler and are re-checked on subsequent crawls, so that content newly subject to a reservation is excluded from further collection.*
Additional comments (optional):	*N/A*

3.2 Removal of illegal content

General description of measures taken:	*Xiaomi applies screening measures to identify and remove from the training data content that is illegal under Union law. These measures cover all modalities present in the training data (text, image, video, and audio) and include blacklists, keyword-based rules, hash matching, and model-based classifiers. These measures address the lawfulness of the content and are not data selection practices aimed at improving model capabilities.*

3.3. Other information (optional)

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