Data Summary for MAI-Image-2

Version of the Summary: *19 March 2026*
Last update: *19 March 2026*

1. General information

1.1 Model developer identification

1.1.1 Model developer name and contact details: Microsoft Ireland Operations Limited (MIOL) 70 Sir John Rogerson's Quay, Dublin 2, D02 R296, Ireland

1.1.2 Authorized representative name and contact details: MSFTAIActRequest@microsoft.com

1.2: Model identification

1.2.1 Versioned model name(s): MAI-Image-2

1.2.2 Model dependencies: N/A

1.2.3 Model release date: 19 March 2026

1.2.4 Date of placement of the model on the Union market: Coming soon.

1.3 Modalities, overall training data size and other characteristics

1.3.1 Size of dataset per modality (table)

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 to10 trillions tokens ☐ More than 10 trillions tokens _____Alternatively, specify the approximate size in a different measurement unit:	The MAI-Image-2 text training corpus is primarily large-scale, multi-domain, safety-filtered, supervised fine-tuned (SFT)/synthetic image caption dataset collections curated for quality.
✓ Image	☐ Less than 1 million images	The MAI-Image-2 image training corpus is primarily large-scale, multi-domain, safety-

	☐ 1Million to1 billion images ✓ More than 1 billion images	filtered, acquired, open source, or publicly available image dataset collections curated for quality.
☐ Audio (*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')*	☐ 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*	N/A

1.3.2 Latest date of data (acquisition/collection for model training):

The data used to train the model is composed of different datasets with varying publication and cutoff dates, with datasets collected as late as February 2026 for model training. The model will not be continuously trained on new or dynamic data while in production, but subsequent versions may undergo additional fine-tuning and be released as later versions.

1.3.3 Is data collection ongoing to update the model with new data collection after deployment? ☐ Yes ✓ No

1.3.4 Date the training dataset was first used to train the model: January 2026

1.3.5 Description of the linguistic characteristics of the overall training data:

Text datasets used to train the model are in English.

1.3.6 Other relevant characteristics of the overall training data:

N/A

1.3.7 Rationale or purpose of data selection:

Training data for MAI-Image-2 was primarily collected to provide general purpose coverage of objects, scenes, unidentified people, image–text pairs, and extensive recaptioning or Vision Language Model (VLM) auto-captioning to improve alignment and avoid sensitive content. This development pipeline supports arbitrary aspect ratios within reasonable bounds, ensuring flexibility for diverse image formats. The diversity in image subjects is technically imperative for several reasons:

- Unseen object and scene coverage: by including a wide range of everyday objects, natural environments, and human activities, the dataset enables the model to generalize effectively to new scenarios. This ensures robust performance not only on common items but also on rare or complex compositions, which is essential for general purpose image generation tasks. Special attention was paid to images with complex lighting (e.g., shadows) to improve output image quality with regard to lighting realistic-ness.
- People-focused data: special attention is given to images of people, including portraits and full-body shots. This is important for generating accurate depictions of human features, activities, and interactions.
- Sensitive content removal: Microsoft-approved image understanding and safety classifiers are used to filter out sensitive data used for training and add additional safety guardrails to the front-end user experience where the model is deployed in products.

2. List of data sources

2.1 Publicly available datasets

2.1.1 Have you used publicly available datasets to train the model?

✓ Yes ☐ No

2.1.2 If yes, specify the modality(ies) of the content covered by the datasets concerned:

✓ Text ✓Image ☐ Video ☐ Audio ☐ Other (please specify)

2.1.3 List of large publicly available datasets:

Training data includes text and images from Wikipedia, which is a large scale, multi-domain, open source, and publicly available collection from a reputable online source.

2.1.4 General description of other publicly available datasets not listed above:

Publicly available datasets are safety-filtered and curated for quality with an emphasis on people, scenes, and objects. Publicly available data excludes sources with paywalls or sites that have opted out of training using published web controls (for example, we respect robots.txt files on third party websites). Training data does not include domains listed in the Office of the United States Trade Representative (USTR) Notorious Markets for Counterfeiting and Piracy list.

2.1.5 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

2.2.1.A Have you concluded transactional commercial licensing agreement(s) with rightsholder(s) or with their representatives?

✓ Yes, we leveraged data acquisition agreements. ☐ No

2.2.1.B 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

2.2.2.A 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

2.2.2.B If yes, specify the modality(ies) of the content covered by the datasets concerned:

☐ Text ✓ Image ☐ Video ☐ Audio ☐ Other (please specify)

2.2.2.C If publicly known, list private datasets obtained from other third parties:

Relevant data acquisition deals are bound by confidentiality terms and conditions. If the parties mutually agree to publicize the partnership in the future, we will update this data summary to provide additional information.

2.2.2.D General description of non-publicly known private datasets obtained from third parties:

This data includes synthetic image-based text captions and images with a focus on objects, scenes, and people. Acquired data is vetted for compliance with use rights, applicable data privacy and security laws, and are covered by agreements describing the roles and responsibilities of the parties with respect to the data.

2.2.2.E Additional comments (optional):

N/A

2.3 Data crawled and scraped from online sources

2.3.1 Were crawlers used by the provider or on behalf of?

✓ Yes ☐ No

2.3.2 If yes, specify crawler name(s)/identifier(s): Bingbot

2.3.3 Purposes of the crawler(s): Index web content for both search and model training

2.3.4 General description of crawler behavior:

For training purposes, we filter publicly available crawled data sources to exclude content behind paywalls, content that violates Microsoft's Responsible AI (RAI) policies, or sites that have opted out of training using published web controls (for example, we respect robots.txt files on third party websites). This includes respect of captchas, password protected websites and paywalls, robots.txt, and other

protocols while crawling. Training data does not include domains listed in the Office of the United States Trade Representative (USTR) Notorious Markets for Counterfeiting and Piracy list.

2.3.5 Period of data collection: February 2024 to January 2026

2.3.6 Comprehensive description of the type of content and online sources crawled:

Crawled training data includes a wide variety of filtered, publicly available sources on general web imagery and their associated alt text descriptions, spanning everyday objects, scenes, and people with a focus on high-quality images. Publicly available crawled data sources are safety-filtered and curated for quality (deduplication, watermarks, quality, aesthetics, Bing safety filters, etc.).

2.3.7 Type of modality covered:

✓ Text ✓ Image ☐ Video ☐ Audio ☐ Other (please specify)

2.3.8 Summary of the most relevant domain names crawled:

The top 10% of all domains crawled and used to train MAI-Image-2 is a blend of publicly available image hosts, blogging and social platforms, e-commerce, and region-specific portals. Overall, the crawl is skewed towards a handful of large social platforms and hosts which was filtered as described in Section 3.2. Country codes like .jp, co.jp, .de, .ru, .cn, .fr, .br, .it, .uk, .pl, and .kr are prominent and reflect broad geographic range of the sourced images. Crawled text data includes alternative text descriptions (i.e., alt text) for the images described above. Most commonly, alt text is located in the HTML <img> tag's alt attribute.

2.3.9 Additional comments (optional):

N/A

2.4 User data

2.4.1 Was data from user interactions with the AI model (e.g. user input and prompts) used to train the model?

☐ Yes ✓ No

2.4.2 Was data collected from user interactions with the provider's other services or products used to train the model?

☐ Yes ✓ No

2.4.3 If yes, provide a general description of the provider's services or products that were used to collect the user data: N/A

2.4.4 Type of modality covered:

☐ Text ☐ Image ☐ Video ☐ Audio ☐ Other (please specify)

2.4.5 Additional comments (optional):

N/A

2.5 Synthetic data

2.5.1 Was synthetic AI-generated data created by the provider or on their behalf to train the model?

✓ Yes ☐ No

2.5.2 If yes, modality of the synthetic data:

✓ Text ✓Image ☐ Video ☐ Audio ☐ Other (please specify)

2.5.3 If yes, specify the general-purpose AI model(s) used to generate the synthetic data if available on the market:

Synthetic data used in training 2 was generated using a combination of Vision Language Models (VLMs) and manual rendering tools. Specifically, VLMs were employed to produce text captions for images. For example, GPT-4o was used to caption approximately 1000 images. The models were self-hosted on secure infrastructure within Microsoft's own cluster.

2.5.4 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.

2.5.5 Provide a description of the need or desired purpose for using synthetic data for a model or system's intended purpose:

Synthetic data was used to augment coverage in certain domains where real data is scarce. It has been useful to improve visual-text generation fidelity, caption quality, and alignment. Synthetic data is generated using Microsoft licensed or open-source models.

2.5.6 Additional comments (optional):

An open-source 3D creation suite was used to manually render synthetic visual (image) and text data to improve text rendering capability. For example, images that contain words such as billboards were created to improve image spelling.

2.6 Other sources of data

2.6.1 Have data sources other than those described in Sections 2.1 to 2.5 been used to train the model?

☐ Yes ✓ No

2.6.2 If yes, provide a narrative description of these data sources and the data: N/A

2.6.3 Additional comments (optional):

N/A

3. Data processing aspects

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

3.1.1 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

3.1.2 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:

Microsoft's Bing crawler bots respect the Robots Exclusion Protocol for text files placed in the header of webpages as a method for reserving TDM rights.

In addition, Microsoft's Bing crawler bots respect a number of meta-tag and HTML tag attributes, giving webmasters greater control over how their content is used and displayed. (See, Announcing new options for webmasters to control usage of their...) Crawled data has been filtered by point-in-time robots.txt when accessed for training to ensure compliance with reserved rights.

3.1.3 Additional comments (optional): N/A

3.1.4 Does this dataset include any data protected by copyright, trademark, or patent?

✓ Yes, see response to section 2.2.2. ☐ No

3.2 Removal of illegal content

3.2.1 General description of measures taken to avoid or remove illegal content: Microsoft follows applicable laws and best practices to avoid or remove illegal content.

3.3 Other information

3.3.1 Does the dataset include information about consumer groups without revealing individual consumer identities?

☐ Yes ✓ No

3.3.2 Was the dataset cleaned or modified before model training?

✓ Yes ☐ No

3.3.3 Other relevant information about data processing (optional):

N/A