Data Summary for MAI Image 2.7

Version of the Summary: *1.0*
Last update: *17 November 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 2.7

1.2.2 Model dependencies: N/A

1.2.3 Model release date: 23 October 2026

1.2.4 Date of placement of the model on the Union market: 17 November 2026

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	The MAI Image 2.7 text training corpus is primarily large-scale, multi-domain, safety-filtered, supervised fine-tuned (SFT)/synthetic image caption dataset collections curated for quality.

	_____Alternatively, specify the approximate size in a different measurement unit:	
✓ Image	☐ Less than 1 million images ☐ 1Million to1 billion images ✓ More than 1 billion images	The MAI Image 2.7 image training corpus is primarily large-scale, multi-domain, safety-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 November 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:

April 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:

Training data for MAI Image 2.7 combines publicly available, acquired, and web-crawled images and associated text. Pre-training uses large-scale image–text datasets covering objects, scenes, people, and images containing text. Post-training focuses on improving prompt following, visual quality, and image-editing capabilities. Data preparation includes caption generation and refinement to improve alignment between images and their textual descriptions. Synthetic examples, including rendered images containing text, support improved text-rendering accuracy. Microsoft-approved image-understanding and safety classifiers are used to filter sensitive content from training data, with additional safety guardrails applied in products where the model is deployed.

1.3.7 Rationale or purpose of data selection:

Training data for MAI Image 2.7 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.

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:

No large datasets were used for training the model.

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. The public data consists of the following types of images:

Realistic imagery – approx. 70.7%

Graphic & information design- approx. 23.2%

Art & illustration – approx. 6.1%

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. The non-public data consists of the following types of images:

Realistic imagery – approx. 80.4%

Graphic & information design – approx. 5.8%

Art & illustration – approx. 13.8%

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 August 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 2.7 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:

N/A

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:

N/A

2.5.6 Additional comments (optional):

Open-source 3D creation and image editing tools were used to 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. These images were not AI generated.

2.7 Other sources of data

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

☐ Yes ✓ No

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

2.7.3 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 respects robots.txt and meta tag instructions indicating certain online content should not be used to train Microsoft's foundation models. For more information see the robots meta tags and attributes Microsoft supports. Beginning February 1, 2027, our practices will include a shared AI training preference in robots.txt that Microsoft will extend to the page level. For more information see the frequently asked questions page here and associated announcement here. Web data is also only used from publicly accessible pages: content behind paywalls or logins is not used. Microsoft also does not use data from domains listed in the Office of the United States Trade Representative (USTR) Notorious

Markets for Counterfeiting and Piracy list. For more information, see Data for AI Training | Microsoft Trust Center. All training data is deduplicated for exact and fuzzy matches.

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 content that is illegal under EU law from training data. For example, before training, data is screened by hash matching, including Microsoft PhotoDNA, against databases of known illegal content: child sexual exploitation and abuse imagery (CSEAI, also called CSAM), nonconsensual intimate imagery (NCII), and terrorist and violent extremist content (TVEC). Matches are removed from training data. Domains associated with piracy and counterfeiting are excluded, as described in Section 2.3.4, as are web pages subject to valid DMCA takedown notices or court blocking orders and adult domains on the UT1 blocklist.

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