

Data Summary for MAI-Thinking-1

Version of the Summary: 1.0
Last update: 12 August 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-Thinking-1
- 1.2.2 Model dependencies:** N/A
- 1.2.3 Model release date:** 2 June 2026
- 1.2.4 Date of placement of the model on the Union market:** 12 August 2026

1.3 Modalities, overall training data size and other characteristics

1.3.1 Size of dataset per modality (table)

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 to 10 trillions tokens ☒ More than 10 trillions tokens Alternatively, specify the approximate size in a different measurement unit:	General-purpose text training corpus spanning code, academic articles, PDF content, math and STEM web crawled/downloaded content, books, general web crawled content, and post-training synthetic text data. Datasets were safety-filtered and curated for quality.
☐ Image	☐ Less than 1 million images ☐ 1Million to 1 billion images ☐ More than 1 billion images	N/A
☐ Audio (Excluding audio)	☐ Less than 10 000 hours	N/A

<i>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’)</i>	☐ 10 000 to 1 million hours ☐ More than 1 million hours	
☐ Video	☐ Less than 10 000 hours ☐ 10 000 to 1 million hours ☐ More than 1 million hours	N/A
☐ Other	<i>Specify the modality and for each one</i> <i>indicate approximate size and unit of measurement</i>	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 July 2026 for model training. The model will not be continuously trained on new or dynamic data while in production, but subsequent releases may undergo additional fine-tuning, extending training runs, data refreshes, and added modality capabilities which would 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: March 2026

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

Text datasets used to train the model are primarily in English. Training data also includes German, Spanish, French, Italian, Portuguese, Chinese (Simplified), Russian, Hindi, Japanese, Korean, Arabic, and Hungarian. Additional limited coverage also includes Hebrew, Turkish, Persian, Thai, Vietnamese, Indonesian, and Ukrainian, among others. Coverage and quality vary by language.

1.3.6 Other relevant characteristics of the overall training data:

Training data combines publicly available, commercially acquired, and crawled web data. Microsoft-approved safety classifiers are used to filter out sensitive data used for training and provide additional safety guardrails to the frontend user experience where the model is deployed in products. User interaction data from the model itself is not used. Synthetic data is used in post-training (supervised fine-

tuning and reinforcement learning) and for quality classification, captioning, and metadata generation in pre-training data curation.

1.3.7 Rationale or purpose of data selection:

Training data for MAI-Thinking-1 was selected to provide broad general-purpose coverage for natural language understanding and generation across code, reasoning, academic, web, and multilingual use cases. The diversity in source types is technically important to support generalization across domains, improve reasoning depth, and maintain strong performance on both natural language and technical tasks.

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:

We used a variety of large, publicly available datasets including GitHub public repositories, Wikipedia, and CommonCrawl (English and multilingual).

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

Publicly available datasets are safety-filtered and curated for quality from a wide range of publicly available heterogeneous data sources. Approximate composition by category includes code, PDFs of academic journals/books, math and STEM web content (English and multilingual), general web content, structured knowledge bases, and images pertinent to VLM-related tasks. Global fuzzy deduplication was applied across all data sources simultaneously, and eval decontamination was applied against held-out evaluation benchmarks. Training data does not include domains listed in the Office of the United States Trade Representative (USTR) Notorious Markets for Counterfeiting and Piracy list.

2.2 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:

Acquired text data is vetted for compliance with use rights, applicable data privacy and security laws, and is covered by agreements describing the roles and responsibilities of the parties with respect to the data.

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 and meta tags 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 December 2025

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 web content spanning everyday topics, technical and STEM domains, news, blogs, forums, community Q&A websites, educational sites, and reference sources. Both English and multilingual content is included. Publicly available crawled data sources are safety-filtered, quality-filtered, and deduplicated (exact and fuzzy).

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% domains by volume crawled and used to train MAI-Thinking-1 span a broad mix of content types. These include publicly available reference and knowledge sites, blogging and personal-publishing platforms, social and professional networks, educational and study platforms, document-sharing services, technical Q&A and developer communities, scientific and academic publishers and archives, news outlets and news archives, creative-writing and design communities, e-commerce marketplaces, and large search portals. General web hosting, blogging, social, and e-commerce platforms make up a large share of the total. The content is substantially multilingual. Alongside English, it includes significant volumes of Spanish, Portuguese, Italian, French, German, and Russian, as well as a large East-Asian presence in Chinese, Japanese, and Korean. A substantial share of the top domains are China- or Japan-based. Prominent country-code TLDs include .jp, .cn, .co.uk, .es, .fr, .ca, and .nz, alongside the .eu institutional domain. The top domains include government and public-sector sources, academic (.edu) sources, and recognized authoritative reference and scholarly publishers.

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:

Conversation logs from Microsoft Consumer Copilot users who have not opted out of having their data used for model training (and excluding certain other users as set forth in the Microsoft Privacy Statement and [Privacy FAQ for Microsoft Copilot | Microsoft Support](#)) were used as inputs for reinforcement learning rollouts during post-training, as context inputs for supervised fine-tuning across multiple post-training domains, and in training internal reward models. Production log data was filtered to exclude identifying information, tool-use traces, safety-sensitive content, and non-web-surface / nonsensical prompts before use. The model generates rollout responses during training, which are scored by the reward model – production log responses are not used as completion targets for supervised fine-tuning and all SFT completions are model-generated.

2.4.4 Type of modality covered:

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

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:

Microsoft did not use publicly available general-purpose AI models to create synthetic outputs for direct training. Internal checkpoints of MAI-Thinking-1 (MAI-Base-1) were used to generate synthetic outputs which were directly trained on.

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:

Internal Microsoft AI models, (i.e., MAI-Thinking-1 base checkpoints (MAI-Base-1)), that are not available on the market were used to generate synthetic post-training data. These included internal versions of MAI text models used as teachers for mathematics, STEM reasoning, and software engineering training data, internal models used to generate chain-of-thought rationales appended to human-written responses, grade instruction-following, tool use, and agentic task-completion training data, and an internal safety-tuned model used to distill safety reasoning.

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 training data in domains where human-annotated data is scarce or where verifiable training signals at scale are needed. It supports model capabilities including reasoning, code, tool use, agentic task completion, and instruction-following, and is also used to support automated quality and topic classification during pre-training data curation. Synthetic data was selected because it provides scalable, targeted training signals that would not be feasible to obtain from human annotation alone at the required volume.

2.6 Other sources of data

2.6.1 Was personal data used to train the model? Microsoft follows applicable laws and best practices pertaining to personal data.

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

☐ Yes ✓ No

2.6.3 If yes, provide a narrative description of these data sources and the data: 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 Does this dataset include any data protected by copyright, trademark, or patent?

☒ Yes ☐ No

Microsoft follows applicable laws and best practices for processing data protected by copyright, trademark, or patent.

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