Data Summary for MAI-Cyber-1-Flash

Version of the Summary: *1.0*

Last update: *27 July 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-Cyber-1-Flash

1.2.2 Model dependencies: MAI-Code-1-Flash, which is based on MAI-Thinking-1.

1.2.3 Model release date: 27 July 2026

1.2.4 Date of placement of the model on the Union market: 27 July 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 ☐ 1 billion to 10 trillion tokens ✓ More than 10 trillion tokens Alternatively, specify the approximate size in a different measurement unit:	Code-focused training content, including source code, shell scripts, software-engineering tasks, repository and file question-answering data, instruction-following data, and general technical and software-engineering content. The training data also includes synthetic software-engineering and cybersecurity tasks supporting vulnerability discovery, validation, triage, and remediation workflows.
☐ Image	☐ Less than 1 million images ☐ 1Million to 1 billion images ☐ More than 1 billion images	N/A

☐ 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 to 1 million hours ☐ More than 1 million hours	N/A
☐ Video	☐ Less than 10 000 hours ☐ 10 000 to 1 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 May 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: 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 covers a majority of EU official languages (e.g., Bulgarian, Croatian, Czech, Danish, Dutch, Estonian, Finnish, French, German, Greek, Hungarian, Italian, Latvian, Lithuanian, Polish, Portuguese, Romanian, Spanish, Swedish) and additional non-EU languages including Chinese, Russian, Japanese, Korean, Arabic, Hindi, Hebrew, Turkish, Persian, Thai, Vietnamese, and Indonesian, among others.

1.3.6 Other relevant characteristics of the overall training data:

The inherited MAI-Code-1-Flash corpus includes source code and software-engineering data, including source files, pull requests, and related code artifacts, drawn from publicly available, commercially licensed and crawled web data (inherited from the MAI-Thinking-1 base) with code-specific public

datasets. Mid-training and post-training data for MAI-Code-1-Flash differs from MAI-Thinking-1: a higher proportion of code, repository, pull request (PR), and commit data is used, and synthetic software engineering (SWE) task generation is used at scale. Post-training added cybersecurity specialization adds task data covering repository preparation, vulnerability scanning, proof-of-concept generation, and patch remediation, along with synthetic cyber tasks generated for security specialization.

1.3.7 Rationale or purpose of data selection:

These datasets were selected to first build strong software-engineering and code-reasoning capability through the base model, and then to specialize the model for defensive cybersecurity workflows. The cybersecurity and synthetic cyber-task data, combined with reinforcement learning in real-world executable security environments, were chosen because they represent the vulnerability discovery, validation, and remediation scenarios the model is intended to support within codename MDASH, which are not well represented in general-purpose datasets.

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, PDF and academic documents, math and STEM web content (English and multilingual), general web content, books, and structured knowledge bases. Global fuzzy deduplication was applied across all data sources simultaneously, and eval decontamination was applied against held-out evaluation benchmarks. Additionally, 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.

2.2 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.

2.2.2.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 December 2025

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

Crawled training data includes filtered, publicly available web content spanning everyday topics, technical and STEM domains, news, blogs, forums, community Q&A websites, educational sites, and reference sources, in English and multilingually. Code-specific scrape covers public code repository

files, pull requests, commit histories, notebooks, shell scripts, and PR localization data. Sources are safety-filtered, quality-filtered, and deduplicated.

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-Cyber-1-Flash 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 contexts (prompts and associated context only) from GitHub Copilot Free, Pro, and Pro+ users who have not opted out of having their data used for model training, starting from the date that GitHub Copilot's notice of updates to its interaction data usage policy became effective, were used as inputs for reinforcement learning rollouts during post-training and were also used in training of an internal reward model. Production log data was filtered to exclude identifying information,, images, tool-use traces, safety-sensitive content, and non-conversational content before use. The model itself generates rollout responses during training; production log responses are not used as training targets for supervised fine-tuning.

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 used a variety of publicly available general-purpose AI models (e.g., OpenAI: GPT-4o, GPT-5) to support training data preparation, including data labeling, quality classification, metadata annotation, embedding-based deduplication, and generating natural language descriptions of publicly available code.

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:

Outside of the existing MAI-Thinking-1 base, internal Microsoft AI models and judge models not available on the market were used to generate code-specific synthetic post-training datasets: Repo Commit Sampling, Repo QA, File QA, SWE Bench (Train/Hard), SWE Instruction Following, UI Tasks, Cold Start, FFT formatting data, and Mid-training RKLD synthetic data.

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 scale verifiable training signal for software engineering tasks where high-quality human-annotated data is scarce. It supports PR creation, repository and file understanding, debugging, instruction-following, and reasoning. Cold-start synthetic data prepares the model for reinforcement learning, and format fine-tuning (FFT) synthetic data standardizes output formatting. Synthetic generation was selected over human annotation because of the volume and verifiability required for code tasks.

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 Additional comments (optional): N/A

3.1.4 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