Data Summary for MAI Code 2

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 Code 2

1.2.2 Model dependencies: MAI 80B Base Model (Internal)

1.2.3 Model release date: 17 November 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 ☐ 1 billion 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 multilingual content, plus post-training synthetic data covering software engineering, CLI/terminal, tool use, instruction-following, knowledge work, STEM reasoning, web research, UI generation, safety, and telemetry-derived coding tasks. Datasets were safety-filtered and curated for quality.

✓ Image	☐ Less than 1 million images ☐ 1 Million to 1 billion images ✓ More than 1 billion images	Images used in pre-training and mid-training for vision-language tasks, and images accompanying post-training knowledge-work, document, and UI generation tasks.
☐ 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 March 2026 for pre-training and synthetic data generated through November 2026 during post-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: July 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 Swedish, Romanian, Polish, Dutch, and Czech,

among others. Post-training synthetic data is predominantly in English. 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 used to pre-train and mid-train the MAI 80B base model with post-training data focused on software engineering. Pre-training and mid-training data includes a high proportion of code (public repository files, pull requests, commits, notebooks, and shell scripts) and vision-language data. Post-training uses synthetic data followed by reinforcement learning, across more than 1.4 million task prompts (~166 billion trained tokens) in 16 task families, including text-only tasks and tasks with accompanying images. A portion of post-training tasks is derived from development-assistant telemetry (see Section 2.4).

Microsoft-approved safety classifiers are used to filter and remove sensitive data from 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 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 Code 2 was selected to maximize performance on agentic software engineering tasks. Post-training data was selected to maximize coverage of repository-level issue resolution, code editing and testing, CLI/terminal operation, MCP and tool use, UI generation, knowledge work, web research, developer-facing instruction-following, and STEM reasoning, while retaining broad reasoning and language understanding from pre-training. Telemetry-derived tasks were included to align training with real-world developer workflows, and safety/responsible AI tasks were included to support policy-aware behavior.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, GitLab, and other public repositories, CommonCrawl (English and multilingual), Wikipedia, Stack Exchange, and arXiv.

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, images, and interleaved image-text data pertinent to VLM-related tasks, math and STEM web content (English and multilingual), general web content, PDFs of academic journals/books, multilingual web and reference content, and structured knowledge bases. 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:

Private datasets obtained from third parties, including data intermediaries, are covered by agreements. Approximate composition by category includes books, scientific, technical, and mathematical content, including figures, formulas, and question-and-answer content, image captions, software engineering and productivity tasks, and environments created for model training. The text content is primarily in English, with multilingual coverage.

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: July 2010 to March 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 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. Code-specific data covers public code repository files, pull requests, commit histories, notebooks, and shell scripts. 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 Code 2 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 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 references 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 (30 days from notice), were used to construct software-engineering tasks (instruction-following tasks) that served as prompts for synthetic data generation during post-training. Production log data was filtered to exclude identifying information, safety-sensitive content, and non-conversational content before use. Rollout responses are generated by the model and the teacher model during training. Production log responses are not used as training targets.

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:

GPT-5.6 Sol and GPT-6 Astra (OpenAI), commercially available general-purpose AI models, were used to generate synthetic data that was used in post-training across various task families.

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 not available on the market were used to generate, verify, and grade post-training data, including task and environment generation for software engineering, CLI/terminal, tool use, knowledge work, UI, STEM, research, instruction-following, and safety tasks.

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 signals for software engineering tasks where high-quality human-annotated data is scarce. It supports repository-level issue resolution, code editing and testing, CLI and tool use, agentic task completion, knowledge work, web research, instruction-following, and reasoning, 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 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 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 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