

Data Summary for microsoft__MAI-DS-R1

1. General information

1.0.1 Version of the Summary: 1.0

1.0.2 Last update: 10-Dec-2025

1.1 Model Developer Identification

1.1.1 Model Developer name and contact details: Microsoft Corporation
at One Microsoft Way, Redmond, WA 98052. Tel: 425-882-8080

1.2 Model Identification

1.2.1 Versioned model name(s): MAI-DS-R1

1.2.2 Model release date: April 2025

1.3 Overall training data size and characteristics

1.3.1 Size of dataset and characteristics

1.3.1.A Text training data size: Not applicable. Text data is not part of the training data

1.3.1.B Text training data content: Not applicable

1.3.1.C Image training data size: Not applicable. Images are not part of the training data

1.3.1.D Image training data content: Not applicable

1.3.1.E Audio training data size: Not applicable. Audio data is not part of the training data

1.3.1.F Audio training data content: Not applicable

1.3.1.G Video training data size: Not applicable. Video data is not part of the training data

1.3.1.H Video training data content: Not applicable

1.3.1.I Other training data size: The model was post-trained using 110k Safety and Non-Compliance examples from the Tulu 3 SFT dataset and ~350k multilingual examples internally developed capturing various topics with reported biases

1.3.1.J Other training data content: Safety and Non-Compliance examples from Tulu 3 SFT and internally developed multilingual examples addressing various topics with reported biases

1.3.2 Latest date of data acquisition/collection for model training:
March 2025

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

1.3.4 Date the training dataset was first used to train the model:
March 2025

1.3.5 Rationale or purpose of data selection: The post-training data combined safety and non-compliance examples and a large multilingual set addressing reported biases to improve responsiveness on previously blocked topics and reduce harmful or unsafe outputs while preserving reasoning capabilities

2. List of data sources

2.1 Publicly available datasets

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

2.2 Private non-publicly available datasets obtained from third parties

2.2.1 Datasets commercially licensed by rights holders or their representatives

2.2.1.A Have you concluded transactional commercial licensing agreement(s) with rights holder(s) or with their representatives? Not applicable

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? No

2.3 Personal Information

2.3.1 Was personal data used to train the model? Microsoft follows all relevant laws and regulations pertaining to personal information

2.4 Synthetic data

2.4.1 Was any synthetic AI-generated data used to train the model?
Yes

3. Data processing aspects

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

3.1.1 Does this dataset include any data protected by copyright, trademark, or patent? Microsoft follows all required regulations and laws for processing data protected by copyright, trademark, or patent

3.2 Other information

3.2.1 Does the dataset include information about consumer groups without revealing individual consumer identities? Microsoft follows all required regulations and laws for protecting consumer identities

3.2.2 Was the dataset cleaned or modified before model training? Yes