Data Summary for microsoft_phi-4

1. General information

1.0.1 Version of the Summary: 1.0

1.0.2 Last update: 24-Nov-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): phi-4

1.2.2 Model release date: 12-Dec-2024

1.3 Overall training data size and characteristics

1.3.1 Size of dataset and characteristics

1.3.1.A Text training data size: 1 billion to 10 trillion tokens

1.3.1.B Text training data content: Training data is an extension of the data used for Phi-3 and includes a wide variety of sources from:

1. Publicly available documents filtered rigorously for quality, selected high-quality educational data, and code.
2. Newly created synthetic, “textbook-like” data for the purpose of teaching math, coding, common sense reasoning, general knowledge of the world (science, daily activities, theory of mind, etc.).
3. Acquired academic books and Q&A datasets.
4. High quality chat format supervised data covering various topics to reflect human preferences on different aspects such as instruct-following, truthfulness, honesty and helpfulness.
5. Multilingual data constitutes about 8% of our overall data.

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

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: Not applicable

1.3.1.J Other training data content: Not applicable

1.3.2 Latest date of data acquisition/collection for model training: 30-Jun-2024

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: 10/01/2024

1.3.5 Rationale or purpose of data selection: Datasets were selected to maximize high-quality reasoning and problem-solving capabilities. The mixture emphasizes synthetic, curriculum-structured data and rigorously filtered organic sources such as academic papers, licensed books, code, and Q&A to improve STEM reasoning, coding, and general knowledge while reducing noise and contamination. Targeted acquisitions and multilingual content complement synthetic data to balance reasoning strength with factual coverage

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? Yes

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? This information cannot be provided due to unavailability of the underlying data (e.g., loss, corruption, or other access limitations)

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