

# Data Summary for microsoft\_\_Phi-4-multimodal-instruct

## 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):** Phi-4-multimodal-instruct

**1.2.2 Model release date:** February 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:** 1 billion to 10 trillion tokens

**1.3.1.B Text training data content:** Publicly available documents filtered for quality, selected educational data, and code; newly created synthetic, “textbook-like” data for the purpose of teaching math, coding, common sense reasoning, general knowledge of the world (e.g., science, daily activities, theory of mind, etc.); human labeled data in chat format; selected image-text interleave data; transcriptions

**1.3.1.C Image training data size:** 1 million to 1 billion images

**1.3.1.D Image training data content:** Selected image-text interleaved data, including synthetic and publicly available images, multi-image sets, and video-derived visual data, filtered for quality and relevance to reasoning tasks

**1.3.1.E Audio training data size:** More than 1 million hours

**1.3.1.F Audio training data content:** Anonymized in-house speech-text pairs with strong and weak transcriptions, selected publicly available and anonymized in-house speech data with task-specific supervision, and selected synthetic speech data supporting automatic speech recognition, translation, QA, and understanding

**1.3.1.G Video training data size:** Not applicable

**1.3.1.H Video training data content:** Not applicable. Video must be treated as a sequence of images.

**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:**  
June 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:**  
December 2024

**1.3.5 Rationale or purpose of data selection:** Data was curated to improve reasoning abilities, including math, coding, common sense, and general knowledge, while filtering publicly available documents to focus model capacity on high-quality content. Additional multimodal data supports image understanding, OCR, chart and table parsing, speech recognition and translation, and instruction following

## **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