

# Data Summary for Magma 8B

## 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):** Magma-8B

**1.2.2 Model release date:** 19-Feb-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:** Less than 1 billion tokens

**1.3.1.B Text training data content:** Image captions, Conversational Dialogs, Text instructions for tasks.

**1.3.1.C Image training data size:** 1 billion to 10 trillion tokens

**1.3.1.D Image training data content:** Training included multimodal image datasets and UI screenshots for grounding and navigation such as ShareGPT4V, LLaVA-1.5 instruction data, InfoGraphicVQA, ChartQA, FigureQA, TQA, ScienceQA, SeeClick and Vision2UI; images cover photography, charts, figures, documents, infographics, and interface elements

**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:** Less than 1 billion tokens

**1.3.1.H Video training data content:** Instructional and egocentric videos used for agentic pretraining and temporal grounding, including Epic-Kitchens, Ego4D, Something-Something v2 and other instructional clips; videos were segmented and filtered, and used to derive Trace-of-Mark trajectories for action planning

**1.3.1.I Other training data size:** Robotics data comprising approximately 9.4 million image-language-action triplets from around 326,000 trajectories within Open-X-Embodiment mixtures

**1.3.1.J Other training data content:** Robotics manipulation datasets from Open-X-Embodiment used for vision-language-action learning, including 7-DoF gripper states and visual traces to support action prediction

**1.3.2 Latest date of data acquisition/collection for model training:**  
11-Jan-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:**  
8-Jan-2024

**1.3.5 Rationale or purpose of data selection:** Datasets were selected to cover multimodal understanding and agentic capabilities across digital and physical environments. UI datasets provide actionable elements for grounding and navigation; instructional videos supply rich temporal dynamics for action planning; robotics datasets provide action trajectories for manipulation; and multimodal image instruction data maintains general visual-language competence. This mix supports spatial-temporal reasoning, grounding, and planning

## **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?** No

### **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