Public Summary of Training Content for AFM 3 Cloud

Version of the Summary:	v1
Last update:	14 September 2026

1. General information

1.1. Provider identification

Provider name:	Apple Inc.
Authorised representative name and contact details:	Apple Distribution International Limited (ADI) Hollyhill Industrial Estate, Cork, T23 YK84, Republic of Ireland

1.2. Model identification

Versioned model name:	AFM 3 Cloud
Model dependencies:	Not applicable
Date of placement of the model on the Union market:	14 September 2026

1.3. Modalities, overall training data size and other characteristics

Modality	Training data size	Types of content
☒ Text	☐ Less than 1 billion tokens ☐ 1 billion to 10 trillions tokens ☒ More than 10 trillions tokens	Mixture of publicly available data, synthetic data, and commercially licensed data, including scientific and academic publications, source code, mathematical content, and educational materials.
☒ Image	☐ Less than 1 million images ☒ 1 million to 1 billion images ☐ More than 1 billion images	Collection of images from publicly available sources, synthetic data, and commercial licenses, including scientific figures and diagrams, mathematical notation, scene text, object detection imagery, infographics and data visualizations.
☐ Audio/ Speech	☐ Less than 10,000 hours ☐ 10,000 to 1 million hours ☐ More than 1 million hours	Not applicable. The model was not trained on audio or speech data.
☒ Video	☐ Less than 10,000 hours ☒ 10,000 to 1 million hours ☐ More than 1 million hours	Video content from publicly available sources and commercial licenses, including video clips across topics, paired with captioning and question-and-answer annotations.

Latest date of data acquisition/ collection for model training:	Up to February 2026.
Description of the linguistic characteristics of the overall training data:	The model was trained on a breadth of languages, including strong English coverage and substantial representation across EU official languages.
Other relevant characteristics of the overall training data:	The pre-training dataset is a large-scale collection of data encompassing a wide range of domains and modalities, which include publicly-available web-documents, code, images, and video. The collection aims to provide broad format, linguistic, and geographic coverage.

2. List of data sources

2.1. Publicly available datasets

Have you used publicly available datasets to train the model?	☒ Yes ☐ No
If yes, specify the modality(ies) of the content covered by the datasets concerned:	☒ Text ☒ Image ☐ Video ☐ Audio
List of large publicly available datasets:	The training data includes a mix of publicly available datasets spanning text, code, mathematical content, scientific literature, and structured documents, sourced from the open research community.
General description of other publicly available datasets not listed above:	Other publicly available datasets include collections focused on mathematical reasoning, visual understanding, document and chart comprehension, and code and software engineering.

2.2. Private non-publicly available datasets obtained from third parties

2.2.1. Datasets commercially licensed by rightsholders or their representatives

Have you concluded transactional commercial licensing agreement(s) with rightsholder(s) or with their representatives?	☒ Yes ☐ No
If yes, specify the modality(ies) of the content covered by the datasets concerned:	☒ Text ☒ Image ☒ Video ☐ Audio

2.2.2. Private datasets obtained from other third parties

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

2.3. Data crawled and scraped from online sources

Were crawlers used by the provider or on behalf of?	☒ Yes ☐ No
If yes, specify crawler name(s)/identifier(s):	Applebot and Applebot-Extended, see here.
Purposes of the crawler(s):	Apple uses crawlers to power various features. The data crawled may be used to help train Apple foundation models powering generative AI features across Apple products, including Apple Intelligence, Services, and Developer Tools. Web publishers can opt out of having their content used to train generative foundation models by disallowing Applebot-Extended in the robots.txt file.
General description of crawler behaviour:	For information on crawler behavior, see here and here.
Period of data collection:	Up to February 2026.
Comprehensive description of the type of content and online sources crawled:	Crawled content includes publicly available online material across government, educational, mathematical, scientific, and general-interest content. Sources include text, image, and video. The content spans a wide range of languages and media types.
Type of modality covered:	☒ Text ☒ Image ☒ Video ☐ Audio
Summary of the most relevant domain names crawled:	The most relevant domains crawled include publicly available websites across academic, research, patent, mathematical, and other technical repositories; educational, legal, and government sources; and document and presentation-hosting services.

2.4. User data

Was data from user interactions with the AI model (e.g. user input and prompts) used to train the model?	☐ Yes ☒ No
Was data collected from user interactions with the provider's other services or products used to train the model?	☐ Yes ☒ No

2.5. Synthetic data

Was synthetic AI-generated data created by the provider or on their behalf to train the model?	☒ Yes ☐ No
If yes, modality of the synthetic data:	☒ Text ☒ Image ☐ Video ☐ Audio
If yes, specify the general-purpose AI model(s) used to generate the synthetic data if available on the market:	Apple relied on a collection of AI models to generate synthetic data used in training the model.
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:	Apple used two internal tools not available on the market to generate synthetic training data such as text recognition data and text tasks designed to train long-context capabilities.

2.6. Other sources of data

Have data sources other than those described in Sections 2.1 to 2.5 been used to train the model?	☒ Yes ☐ No
If yes, provide a narrative description of these data sources and the data:	These data sources do not fit the categories described above. They include data Apple obtained from user studies and human-labeled datasets by Apple employees and annotation vendors.

3. Data processing aspects

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

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
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:	Applebot respects standard robots.txt directives in general search crawls that are targeted at Applebot. In addition to following all robots.txt rules and directives, Apple has a secondary user agent, Applebot-Extended, that gives web publishers additional controls over how their website content can be used by Apple. Web publishers can choose to opt out of their website content being used to train Apple's general purpose foundation models powering generative AI features across Apple products, including Apple Intelligence, Services, and Developer Tools.

3.2. Removal of illegal content

General description of measures taken:	Apple takes protective measures to avoid or remove illegal content under Union law from training data. These measures include removing content that is pornographic, sexually explicit, violent, or that violates child sexual abuse material (CSAM) laws.