AFM 3 Cloud Model Documentation Form

Date this document was last updated: September 23, 2026
Document version number: v1.1

General Information

Legal name for the model provider:	Apple Inc.
Model name:	AFM 3 Cloud
Release date:	September 14, 2026
Union market release date:	September 14, 2026
Model dependencies:	N/A. The model does not have any dependencies.

Model properties

Model architecture:	AFM 3 Cloud is a general-purpose multimodal model based on a Parallel-Track Mixture-of-Experts (PT-MoE) architecture.
Input modalities:	Text; max. 32K tokens Images; max. 32K tokens Video; max. 32K tokens
Output modalities:	Text; max. 32K tokens
Total model size: *The range within which the total number of parameters falls.*	500B-1T

Methods of distribution and licenses

Distribution channels:	AFM 3 Cloud is available to Apple developers through the Foundation Models framework API, for use in their apps distributed on the App Store in regions where Apple Intelligence is available. The model is not made directly available to end users. AFM 3 Cloud is a server-based model accessed through the Foundation Models framework API from an app with a Private Cloud Compute entitlement. The entitlement is available to developers who meet defined eligibility criteria.
License:	AFM 3 Cloud is made available to Apple developers via APIs governed by proprietary licenses.

Use

Acceptable Use Policy:	https://developer.apple.com/apple-intelligence/acceptable-use-requirements-for-the-foundation-models-framework
Intended uses:	The uses that are prohibited by Apple are listed in the Acceptable use requirements at https://developer.apple.com/apple-intelligence/acceptable-use-requirements-for-the-foundation-models-framework/.
Type and nature of AI systems in which the general-purpose AI model can be integrated:	AFM 3 Cloud can be integrated into the following types of AI systems: • Conversational assistants requiring extended multi-turn dialogue • Document analysis and summarisation systems handling long-form or complex textual content • Decision support systems leveraging extended reasoning capabilities (light, moderate, or deep reasoning levels) • Human-AI collaboration tools where on-device processing is insufficient and server-side intelligence is required AFM 3 Cloud should not be integrated into AI systems that: • Operate without a network connection, which is required to access Private Cloud Compute • Run on devices that do not support Apple Intelligence or do not meet the eligibility requirements to access Private Cloud Compute
Technical means for model integration:	The following documentation describes the technical means required to integrate AFM 3 Cloud into an AI system: https://developer.apple.com/documentation/foundationmodels The model preserves user privacy and does not require developers to manage authentication or API keys, as access is governed through Apple's platform and iCloud+ subscription framework.
Required hardware:	Use of Apple silicon with the required software (described in the entry “Required Software” below) or Nvidia B200s.
Required software:	Use of AFM 3 Cloud requires iOS 27, iPadOS 27, macOS 27, watchOS 27, or visionOS 27 or later.

Information on the data used for training, testing, and validation

Data type/modality:	Text Images Video
Data provenance:	Web Crawling Private non-publicly available datasets obtained from third parties Publicly available datasets Data collected through other means Synthetic data that is not publicly accessible (when created directly by or on behalf of the provider) [Other] Data Apple obtained from user studies and human-labeled datasets by Apple employees and annotation vendors.

Data curation methodologies:	Raw data used for training is transformed through a multi-stage pipeline comprising cleaning, quality scoring, synthetic augmentation, and normalization before being combined into the final training mixture. Web and text sources are first filtered for compliance, including respect of publisher opt-outs from foundation-model training prior to ingestion. We take steps to prevent personal data from being included in our models, such as by applying filters to remove personally identifiable information like social security and credit card numbers that are publicly available on the internet. Rather than applying binary keep/discard filtering, the pipeline employs LLM-based rubric scoring to grade web text into relative quality tiers. These scores inform mixture weighting rather than outright exclusion: higher-quality content is up-weighted, while lower-quality material is retained at reduced weight to preserve overall coverage and diversity. Where naturally occurring data is scarce or insufficiently challenging, synthetic data is generated to supplement training and improve performance in targeted domains, such as mathematics and reasoning. Synthetic data generation is treated as a core acquisition source alongside web crawling, licensed data, open-source datasets, and dedicated data-collection studies. Source formats are normalized to ensure consistent model consumption; for example, structured markup is applied to code. Carried-over sources from previous training runs are periodically refreshed to maintain currency. All cleaned, scored, and synthetic components are ultimately combined into a single training mixture, with weights rebalanced across data pillars and languages to reflect updated priorities. Dedicated budget is allocated to high-value domains such as mathematics, code, and documents, while broad web coverage is retained to support overall diversity.