

Public Summary of Training Content	
Version of the Summary:	Document version number: V2
Last Update:	7/8/26
1. General information	
1.1 Provider Identification	
Provider's name and contact details:	Meta Platforms Ireland Limited (MPIL)
Authorized representative name and contact details:	Support Contact: https://ai.meta.com/help
1.2. Model identification	
Versioned model name(s):	Muse Spark
Model dependencies:	N/A
Date of placement of the model on the Union market:	4/8/26
1.3. Modalities, overall training data size and other characteristics	
Modality Select the modalities present in the training data, to the extent that they are identifiable	Training data size For each selected modality, select the range within which the estimated total training data size for that modality falls. Dynamic datasets may be excluded from the estimation.
Text ☒	More than 10 trillions tokens This dataset comprises multimodal content sourced from publicly available data, data provided by third parties and information from Meta's products and services, curated and enriched by external vendor networks and Meta personnel.
Perception (Image & Video) ☒	More than 10 trillions tokens This dataset comprises multimodal content sourced from publicly available data, data provided by third parties and information from Meta's products and services, curated and enriched by external vendor networks and Meta personnel.
Audio (Excluding audio that is part of video, as this should be reported under the "video" modality instead. Furthermore, the Commission understands the modality of 'audio' to include 'speech'.) ☒	More than 10 trillions tokens This dataset comprises multimodal content sourced from publicly available data, data provided by third parties and information from Meta's products and services, curated and enriched by external vendor networks and Meta personnel.
Other ☐	N/A
1.3.1 Other Characteristics	
Latest date of data: acquisition/collection for model training:	Up to June 2026
Description of the linguistic characteristics of the overall training data:	Multiple languages and geographies, including EU official languages.
Other relevant characteristics of the overall training data:	N/A
Additional comments (optional):	For more about the information we use for AI at Meta, where it comes from and how it works, see: https://transparency.meta.com/features/ai-at-meta-training-data/
2. List of data sources	
2.1. Publicly available datasets	
Have you used publicly available datasets to train the model?	Yes
If yes, specify the modality(ies) of the content covered by the datasets concerned:	☒ Text ☒ Image ☒ Video ☒ Audio ☐ Other (Please Specify):
List of large publicly available datasets:	Training Data: A mix of publicly available, licensed data and information from Meta's products and services, including publicly shared posts from Instagram and Facebook. Learn more in our Privacy Center.
General description of other publicly available datasets not listed above:	The model has been trained on multi-modal and multi-lingual (including EU language) data from a range of publicly available datasets.
Additional comments (optional):	N/A
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
If yes, specify the modality(ies) of the content covered by the datasets concerned:	☒ Text ☒ Image ☒ Video ☒ Audio ☐ Other (Please Specify):
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
If publicly known, list private datasets obtained from other third parties:	N/A
General description of non-publicly known private datasets obtained from third parties:	This dataset comprises multimodal content sourced from publicly available data, data provided by third parties and information from Meta's products and services, curated and enriched by external vendor networks and Meta personnel.
Additional comments (optional):	N/A
2.3. Data crawled and scraped from online sources	
Were crawlers used by the provider or on behalf of?	Yes
If yes, specify crawler name(s)/identifier(s):	It is Meta's policy to provide information on Meta's web crawlers through Meta's developer center (https://developers.facebook.com/docs/sharing/webmasters/web-crawlers/).
Purposes of the crawler(s)	
General description of crawler behaviour:	Meta crawlers are designed to respect standard web protocols and ethical data collection practices.
Period of data collection:	Up to June 2026

Public Summary of Training Content	
Comprehensive description of the type of content and online sources crawled:	See the developer center resource linked above for more information about Meta's web crawlers.
Type of modality covered:	☒ Text ☒ Image ☒ Video ☒ Audio ☐ Other (Please Specify):
Summary of the most relevant domain names crawled:	See the developer center resource linked above for more information about Meta's web crawlers.
Additional comments (optional):	N/A
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
Was data collected from user interactions with the provider's other services or products used to train the model?	Yes
If yes, provide a general description of the provider's services or products that were used to collect the user data:	Data from 1st party services like Facebook and Instagram.
Type of modality covered	☒ Text ☒ Image ☒ Video ☒ Audio ☐ Other (Please Specify):
Additional comments (optional):	N/A
2.5. Synthetic data	
Was synthetic AI-generated data created by the provider or on their behalf to train the model?	Yes
If yes, modality of the synthetic data:	☒ Text ☒ Image ☒ Video ☒ Audio ☐ Other (Please Specify):
If yes, specify the general-purpose AI model(s) used to generate the synthetic data if available on the market:	Synthetic data used in training the model was generated by a collection of generative AI models, to generate training data such as ideal responses, golden training datasets and high quality captions for images. For more information on how Meta uses synthetic data for AI at Meta see: https://transparency.meta.com/features/ai-at-meta-training-data/
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:	See above
Additional comments (optional):	N/A
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?	No
If yes, provide a narrative description of these data sources and the data	N/A
Additional comments (optional):	N/A
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?	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	It is Meta's policy to provide information on Meta's web crawlers through Meta's developer center (https://developers.facebook.com/docs/sharing/webmasters/web-crawlers/). Meta's developer page and information relating to crawlers employed by Meta may be updated from time to time.
Additional comments (optional):	N/A
3.2. Removal of illegal content	
General description of measures taken:	Data used for training was subject to several curation methodologies such as cleaning, filtering, summarization and ratings. Meta takes proactive measures aimed at preventing the inclusion of illegal content in generative AI training datasets.
3.2 Other information (optional)	
Other relevant information about data processing (optional):	N/A