Template for the Public Summary of Training Content for General-Purpose AI models

This template is provided by the European Commission and required to be filled in by providers of general-purpose AI models prior to their placing on the Union market in order to comply with their obligation under Article 53 (1)(d) of Regulation (EU) 2024/1689 (AI Act).
For more information and guidance see Commission's Explanatory Notice and Template for the Public Summary of Training Content for general-purpose AI models | Shaping Europe's digital future.

Version of the Summary: 1.0
Last update: 1 August 2026

1. General information

1.1. Provider identification

Provider name and contact details:	Black Forest Labs Inc., 2261 Market Street, Suite 22997, San Francisco, CA 94114, USA ("BFL")
Authorised representative name and contact details:	BFL GmbH, Ingeborg-Krummer-Schroth-Str. 18A, 4OG 79106 Freiburg, Germany

1.2. Model identification

Versioned model name(s):	FLUX 3
Model dependencies:	None
Date of placement of the model on the Union market:	16 July 2026

1.3 Modalities, overall training data size and other characteristics

This Section requires general information about the overall training data after pre-processing and before the training of the model.

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.*	**Types of content** *For each selected modality, provide a general description of the type of content that has been included in the training data.*
x Text	☐ Less than 1 billion tokens x 1billion to 10 trillions tokens ☐ More than 10 trillions tokens	The FLUX 3 text training corpus comprises primarily multi-domain, supervised fine-tuned image-text or video-text caption dataset collections curated for quality, together with deidentified, aggregated text-based information derived from user's interactions with Black Forest Labs services as further set out in Section 2.4.

x Image	☐ Less than 1 million images ☐ 1Million to1 billion images x More than 1 billion images	The FLUX 3 image training corpus comprises primarily diverse, acquired, open, safety-filtered, supervised fine-tuned and synthetic image datasets. It includes bespoke images created by vendors for specific use cases, as well as human created labelling of images.
x Audio[1]	☐ Less than 10 000 hours x10 000 to 1 million hours ☐ More than 1 million hours	The FLUX 3 audio training corpus comprises primarily diverse, acquired, open, safety-filtered, and supervised fine-tuned audio datasets. The datasets include sound effects, background noise, and speech.
x Video	☐ Less than 10 000 hours ☐ 10 000 to1 million hours x More than 1 million hours	The FLUX 3 video training corpus comprises primarily diverse, acquired, open, safety-filtered, supervised fine-tuned and synthetic video datasets. It includes bespoke videos created by vendors for specific use cases, as well as human created labelling of videos.
☐ Other	*Specify the modality and for each one indicate approximate size and unit of measurement*	N/A

Latest date of data acquisition/collection for model training:	The data used to train the model is composed of different datasets with varying publication and cutoff dates. Datasets were acquired as recently as June 2026 for model training. The model will not be trained on new data while in production, but subsequent versions may undergo distillation or fine-tuning and be released as later versions as part of the same family of models.
Description of the linguistic characteristics of the overall training data:	To the extent there are identifiable linguistic characteristics of the datasets, the coverage is multilingual, including EU official languages with strong English language representations.
Other relevant characteristics of the overall training data:	The overall training corpus is designed for a video, image and action prediction output model with multimodal understanding capabilities.
Additional comments (optional):	

2. List of data sources

[1] 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'.

This Section requires information about specific sources of data used to train the general-purpose AI model. In this section "dataset" should be understood as a single, pre-packaged collection of data. The filtering and pre-processing of data collected from the same pre-packaged collection should not be considered a new dataset to be disclosed separately in the sections below. If a particular dataset can be assigned to more than one of the categories below, providers should select the most relevant category and only report the dataset in that category, except in the case of synthetic data (see Section 2.5).

2.1. Publicly available datasets

This Section requires information about datasets that were used to train the model and which have been compiled by a third party, are made available publicly for free, and are readily downloadable as a whole or in predefined chunks, such as datasets and collections available on public repositories and online platforms, specialised websites, or snapshots of common crawl. The public availability of the datasets for free does not mean that the content at issue is necessarily free of rights since it may be subject to licensing arrangements or conditions of use (e.g., certain free and/or open licenses may determine the scope of the uses, including prohibiting uses relating to model training).

A dataset is considered to be "large" if the total data size for any one of the modalities contained in the dataset exceeds 3% of the size of all publicly available datasets for that modality used for training. The size of the dataset should be based on its size after pre-processing (for example filtering), and without splitting the dataset to prevent reporting circumvention.

Have you used publicly available datasets to train the model?	x Yes ☐ No
If yes, specify the modality(ies) of the content covered by the datasets concerned:	x Text x Image x Video ☐ Audio ☐ Other *If so, please specify…*
List of large publicly available datasets:	N/a
General description of other publicly available datasets not listed above:	Training data includes captioning, images, videos, action prediction, text-image and text-video pairs from open source and publicly available scientific research, technical and educational repositories, and specialised collections, for example, Egocentric-100K made available by Build AI published on Hugging Face under an Apache 2.0 licence. These datasets are multilingual, and subject to preprocessing such as quality filtering, deduplication, and safety filtering before training. We use data filtering processes to reduce personally identifiable information from training data.
Additional comments (optional):	BFL uses the U.S. Trade Representative (USTR) Notorious Markets for Counterfeiting and Piracy list as a signal when deciding to exclude data from certain websites that have been recognized as persistently and repeatedly infringing copyright.

2.2 Private non-publicly available datasets obtained from third parties

This Section requires information about private non-publicly available datasets of third parties that are not publicly available and not disclosed under Section 2.1. These include:

1) *datasets for which transactional commercial licensing agreements were concluded between the provider and the rightsholders or their representatives, including by collective management organisations and legitimate content aggregators who have the right to collectively license works on behalf of rightsholders (Section 2.2.1);*

2) *other private datasets obtained through data intermediaries, non-publicly available databases and datasets of third parties for which transactional commercial licenses have <u>not</u> been concluded with rightsholders or their representatives (Section 2.2.2).*

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?	x Yes , BFL has entered into data access agreements ☐ No
If yes, specify the modality(ies) of the content covered by the datasets concerned:	☐ Text x Image x Video x Audio ☐ Other *If so, 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?	x Yes, we have entered into data access agreements ☐ No
If yes, specify the modality(ies) of the content covered by the datasets concerned:	x Text x Image x Video ☐ Audio ☐ Other *If so, please specify…*
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	Data acquired from third party providers through confidential commercial agreements and data access partnerships. Data is collected consistent with applicable law.
Additional comments (optional):	N/a

2.3 Data crawled and scraped from online sources

This Section requires information about crawled, scraped data, or otherwise compiled from online sources directly by the provider of the model or on their behalf (i.e. excluding publicly available datasets already compiled by third parties and made available on platforms such as common crawl that are covered under Section 2.1).

Were crawlers used by the provider or on behalf of?	☐Yes x No

2.4 User data

This Section requires information about user data collected by all services and products of the provider, including through mail services, social media platforms, content platforms or interaction with the providers' AI models and/or systems. This does not cover data licensed by users based on commercial transactional agreements described in Section 2.2.1., or customer data to fine-tune models for specific purposes.

Was data from user interactions with the AI model (e.g. user input and prompts) used to train the model?	x Yes ☐ No

Was data collected from user interactions with the provider's other services or products used to train the model?	x Yes ☐ No
If yes, provide a general description of the provider's services or products that were used to collect the user data:	Subject to user opt-out, data from user's interactions with FLUX models accessed via an API pursuant to an agreement may have been used to improve the quality and capabilities of the FLUX 3 model. BFL uses data filtering and minimisation techniques to reduce personally identifiable information from training data.
Type of modality covered:	x Text x Image ☐ Video ☐ Audio ☐ Other *If so, please specify…*
Additional comments (optional):	

2.5 Synthetic data

This Section requires information about synthetic data created by or on behalf of the provider for training the model directly on the outputs of another AI model, in particular through model distillation or model alignment (e.g. AI feedback through reinforcement learning). This does not include the use of AI models to clean or enrich data (e.g. AI-generated metadata to enrich or modify a dataset, such as creating depth maps or text descriptions of images). In case this concerns publicly available datasets as described in Section 2.1, these should be reported in that Section of the Template. In case this concerns synthetic datasets created by third parties on behalf of the provider, these should be reported in this Section of the Template instead of in Section 2.2.2.

Was synthetic AI-generated data created by the provider or on their behalf to train the model?	x Yes ☐ No
If yes, modality of the synthetic data:	☐ Text x Image x Video ☐ Audio ☐ Other *If so, please specify…*
If yes, specify the general-purpose AI model(s) used to generate the synthetic data if available on the market:	BFL's proprietary publicly available FLUX.1 and FLUX.2 suite of models.
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:	Output from BFL proprietary, internal FLUX models were used selectively as part of the training data mix. Synthetic data is used to support various training objectives including fine tuning, safety research and capability development.
Additional comments (optional):	N/a

2.6 Other sources of data

This Section requires information about data that does not fall under any of the categories in the previous Sections, for example data collected from offline sources, self-digitised media (e.g., digitised analog text context, images), datasets labelled by humans commissioned by the provider, or human generated data through reinforcement learning.

Have data sources other than those described in Sections 2.1 to 2.5 been used to train the model?	☐Yes x No

3. Data processing aspects

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

This Section concerns measures implemented by the provider to identify and comply with the reservation of rights from the text and data mining (TDM) exception or limitation expressed pursuant to Article 4(3) of Directive (EU) 2019/790, as outlined in the copyright policy put in place by the provider in accordance with Article 53(1)(c) AI Act.

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?	x 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:	Third parties from which datasets have been obtained implement a variety of measures to comply with applicable laws.
Additional comments (optional):	N/a

3.2 Removal of illegal content

This Section concerns measures taken to avoid or remove illegal content under Union law from the training data (such as blacklists, keywords, and model-based classifiers), without requiring disclosure of specific details about the provider's internal business practices or trade secrets. Such measures are advisable if the training data is likely to include illegal or unlawful content under Union law, in particular child sexual abuse material and terrorist content and the non-authorised use of material protected by intellectual property rights. Such measures do not include data selection practices, for example to increase the capability of the model.

General description of measures taken:	BFL follows applicable laws and best practices to remove illegal or harmful content from its training corpus by way of preprocessing, deduplication and filtering. To learn more about our practices, please refer to our Responsible AI Development Policy (https://bfl.ai/legal/responsible-ai-development-policy).