Public Summary of Training Content for Imagine Image 2.0

Version of the Summary: V1
Last update: 6 August 2026

1. General information

1.1. Provider identification

Provider name and contact details:	SpaceXAI LLC (formerly XAI LLC), 1450 Page Mill Rd, Palo Alto, CA 94304
Authorised representative name and contact details:	EDSR, Valukoja 8/2, 2nd floor, Tallinn, Estonia, EE-11415

1.2. Model identification

Versioned model name(s):	Imagine Image 2.0
Model dependencies:	N/A
Date of placement of the model on the Union market:	7 August 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	The text tokens used during training consist primarily of captions associated with the training images. These captions describe the visual content of each image and provide the textual supervision needed to align language with image generation and editing. The captions were generated or refined using Grok 4.2. As a result, the text-token component of the training data is predominantly synthesized text designed to accurately reflect and preserve the content, attributes, and context of the corresponding images. Both the images and their associated captions were curated and processed to improve quality, relevance, diversity, and suitability for image-generation and image-editing tasks. Depending on the source and intended use, this processing may include filtering, deduplication, quality assessment, caption generation or recaptioning, and safety-related

		screening.
☒ Image	☐ Less than 1 million images ☐ 1Million to1 billion images ☒ More than 1 billion images	Imagine Image 2.0 was trained using a combination of publicly available data, data generated internally by SpaceXAI, and data that SpaceXAI acquired or licensed from third-party providers. These third-party sources include providers of licensed visual and multimedia content, creator-sourced content platforms, and providers of curated or custom data sets.. The training data was curated and processed to improve data quality, relevance, diversity, and suitability for image generation and editing tasks. Depending on the data source and intended use, this process may include filtering, deduplication, quality assessment, captioning or recaptioning, and safety-related screening.
☐ Audio	☐ Less than 10 000 hours ☐ 10 000 to1 million hours ☐ More than 1 million hours	N/A
☐ Video	☐ Less than 10 000 hours ☐ 10 000 to1 million hours ☐ More than 1 million hours	N/A
☐ Other	N/A	N/A

Latest date of data acquisition/collection for model training:	The data used to train Imagine Image 2.0 includes different datasets from varying time periods, with data collected no later than June 2026.
Description of the linguistic characteristics of the overall training data:	Multilingual, with strong English coverage and substantial representation across EU official languages and other languages from around the world.
Other relevant characteristics of the overall training data:	Grok's training corpus is designed for a highly capable image-output model with strong multimodal understanding.
Additional comments (optional):	

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 ☐ Other
List of large publicly available datasets:	Public sources of Grok Imagine Image 2.0 training data consist of images from publicly available sources from the Internet.
General description of other publicly available datasets not listed above:	
Additional comments (optional):	

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 ☐ Other

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
If yes, specify the modality(ies) of the content covered by the datasets concerned:	☒ Text ☒ Image ☐ Video ☐ Audio ☐ Other
If publicly known, list private datasets obtained from other third parties:	
General description of non-publicly known private datasets obtained from third parties	Private datasets used to train Grok Imagine Image 2.0 include licensed visual content, creator-sourced content, and curated or custom datasets developed for AI-training purposes. Dataset may include images, associated text, captions, labels, and other metadata.
Additional comments (optional):	

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):	xAI Web Crawler
Purposes of the crawler(s):	SpaceXAI uses a web crawler to crawl certain publicly available internet pages to help improve language understanding,

	reasoning, and image coverage.
General description of crawler behaviour:	SpaceXAI uses a web crawler to discover and scan certain publicly available internet pages and rescan for updates and analysis.
Period of data collection:	From January 2024 to June 2026
Comprehensive description of the type of content and online sources crawled:	Content encompasses a wide variety of publicly available internet data including educational, scientific, technical, government, institutional, and general-interest sources. These materials may consist of text, images, and associated metadata.
Type of modality covered:	☒ Text ☒ Image ☐ Video ☐ Audio ☐ Other
Summary of the most relevant domain names crawled:	The domains accounting for the largest share of content crawled and used to train Imagine Image 2.0 may comprise a mix of publicly available information, including content hosting and content sharing services, like the X social media platform, educational repositories, government and institutional resources, community and document-sharing sites and region-specific portals. These domains are global in scope and multilingual. Before use in training, the data is subject to preprocessing measures, including quality filtering, deduplication, safety filtering, and processes designed to reduce the presence of personal information in the training data.
Additional comments (optional):	SpaceXAI 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.

2.4 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
If yes, provide a general description of the provider's services or products that were used to collect the user data:	Subject to privacy settings, controls, user requests, opt-outs, and our policies, SpaceXAI may use interactions with Grok to help train and improve our models, including Grok 4.2 and Grok Imagine. For more information on how we use data to improve model performance, please visit: https://x.ai/privacy
Type of modality covered:	☒ Text ☒ Image ☐ Video ☐ Audio ☐ Other
Additional comments (optional):	SpaceXAI uses advanced data filtering processes to reduce personal information from training data. We also use advanced data processes to reduce the amount of personal data in our training data.

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 ☐Other
If yes, specify the general-purpose AI model(s) used to generate the synthetic data if available on the market:	SpaceXAI generates synthetic data using its own general-purpose AI models, including Grok 4.2, Grok 4.5 and Grok Imagine.
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:	SpaceXAI uses other models to generate synthetic data for targeted training objectives, including augmenting data in domains, languages, tasks, or formats where specialized training data is comparatively scarce. These models may be used to generate examples for instruction following, reasoning, coding, multimodal understanding, safety, and evaluation.
Additional comments (optional):	

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:	
Additional comments (optional):	

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:	SpaceXAI implements measures to respect rights reservations and opt-out signals relevant to data. For example, SpaceXAI does not train on X data from X users who have opted out of training. For web data used for training, SpaceXAI respects instructions that identified crawled content should not be used to train SpaceXAI's generative foundation models.

Additional comments (optional):

3.2 Removal of illegal content

General description of measures taken:	SpaceXAI applies preprocessing and screening measures designed to avoid or remove content that may be illegal from our training data. These measures include filtering, keyword-based rules and model-based classifiers.

3.3. Other information (optional)

Other relevant information about data processing (optional):