Public Summary of Training Content for Grok 4.5

Version of the Summary: V1
Last update: 8 July 2026

1. General information

1.1. Provider identification

Provider name and contact details: 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): Grok 4.5

Model dependencies: Not applicable

Date of placement of the model on the Union market: 14 July 2026

1.3 Modalities, overall training data size and other characteristics

Modality	Training data size	Types of content
☒ Text	☐ Less than 1 billion tokens ☐ 1billion to 10 trillions tokens ☒ More than 10 trillions tokens	Grok 4.5 was pre-trained using a carefully designated data recipe that incorporates a diverse corpus of publicly available information, including, for example, certain scientific text, legal and official documents, social media posts, and source code. The training data also encompasses content produced by third parties, data from users and contractors, as well as internally generated data. Training data was curated and filtered for quality and safety.
☒ Image	☐ Less than 1 million images ☐ 1Million to1 billion images ☒ More than 1 billion images	Grok 4.5 was trained on a large and diverse set of images from publicly available sources, including, for example, certain photographs, visual art works, infographics, and social media images. The training data may also encompass images produced by third parties, from users and contractors, as well as internally generated content. Training data was curated and filtered for quality and safety.
☒ Audio[1]	☐ Less than 10 000 hours ☐ 10 000 to1 million hours	Grok 4.5 was trained on a diverse set of publicly available audio data, including, for example, certain clips and audio

[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'.

	☒ More than 1 million hours	from social media videos. The training data may also encompass audio content produced by third parties, from users and contractors, as well as internally generated content. Training data was curated and filtered for quality and safety.
☒ Video	☐ Less than 10 000 hours ☐ 10 000 to1 million hours ☒ More than 1 million hours	Grok 4.5 was trained on publicly available audiovisual content, including, for example, certain clips and social media videos. The training data may also encompass audiovisual content produced by third parties, from users and contractors, as well as internally generated content. Training data was curated and filtered for quality and safety.
☐ Other	Not applicable	Not applicable

Latest date of data acquisition/collection for model training:	The data used to train Grok 4.5 includes different datasets from varying time periods, with data collected no later than June 2026. The model is also continuously refined on new data after this date.
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 text-output model with strong multimodal understanding. It draws from a wide range of written materials, text-bearing images, and audiovisual content. The corpus is intentionally broad, aiming for extensive coverage across topics, geographies, languages, and formats.
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:	The primary source of Grok 4.5's training is data consisting of text, images, audio content and audiovisual content 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:	Not applicable
General description of non-publicly known private datasets obtained from third parties	xAI partnered with third parties to access data spanning a diverse set of domains and contexts to improve Grok 4.5, consisting of text, images, audio and/or video.
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):	*xAI uses a web crawler to crawl publicly available internet pages to help improve language understanding, reasoning, and overall capabilities.*
General description of crawler behaviour:	*xAI uses a web crawler to discover and scan publicly available internet pages and rescan for updates and analysis.*
Period of data collection:	From 01/2024 to 06/2026
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, audio, video, and associated metadata.
Type of modality covered:	☒ Text ☒ Image ☒ Video ☒ Audio ☐ Other

Summary of the most relevant domain names crawled:	The most relevant crawled domains include academic/research repositories, patent and technical databases, legal/government resources, document-sharing platforms, community sites, and region-specific portals.
Additional comments (optional):	Not applicable

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, xAI may use interactions with Grok to help train and improve our models, including Grok 4.5. 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):	xAI 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:	xAI generates synthetic data using its own general-purpose AI models, including Grok.
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:	xAI uses a combination of internal models to generate synthetic data for targeted training objectives.
Additional comments (optional):	Not applicable

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:	xAI and third parties create data spanning a diverse set of domains and contexts to help our models

	improve on a wide variety of tasks.
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:	xAI respects opt-out signals. For example, xAI does not train on X data from X users who have opted out of training.
Additional comments (optional):	

3.2 Removal of illegal content

General description of measures taken:	xAI 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):	