Public Summary of Training Content for DeepSeek-V4

Version of the Summary:	V1
Last update:	4/08/2026

1. General information

1.1. Provider identification

Provider name and contact details:	Hangzhou DeepSeek Artificial Intelligence Co., Ltd.Room 1201, Building Xi 1, Galaxy International Building, No. 169 North Huancheng Road, Gongshu District, Hangzhou, Zhejiang, China.
Authorised representative name and contact details:	We have appointed Prighter Group with its local partners as our authorised representative in the European Union. Please visit: https://app.prighter.com/portal/17162496076

1.2. Model identification

Versioned model name(s):	DeepSeek-V4（including DeepSeek-V4-Pro and DeepSeek-V4-Flash. ）
Model dependencies:	None
Date of placement of the model on the Union market:	24 April 2026

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.*	**Types of content** *For each selected modality, provide a general description of the type of content that has been included in the training data.*
☒ Text	☐ Less than 1 billion tokens ☐ 1billion to 10 trillions tokens ☒ More than 10 trillions tokens	Publicly available internet information, licensed datasets, and diverse textual content including mathematical texts, source code, multilingual materials, long-form documents and text used for agentic and domain-specific training.
☐ Image	☐ Less than 1 million images ☐ 1Million to1 billion images ☐ More than 1 billion images	N/A
☐ 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	*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 DeepSeek V4 includes different datasets from varying time periods, with some data collected approximately no later than May 2025.
Description of the linguistic characteristics of the overall training data:	Multilingual, with strong coverage in Chinese and English.
Other relevant characteristics of the overall training data:	The training corpus is designed as a whole for the text output model.
Additional comments (optional):	N/A

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 training data includes text from Common Crawl and Stack Exchange. Prior to training, they are processed through quality screening, deduplication, language identification, and safety filtering.
General description of other publicly available datasets not listed above:	Other publicly available datasets include extensive multi-domain text and code corpora released by third parties through public repositories, online platforms, open data portals, and dedicated websites, including encompass long documents, scientific and technical materials, and source code. Prior to training, they are processed through quality screening, deduplication, language identification, and safety filtering. These datasets cover diverse subject areas worldwide and support multiple languages.
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 ☐ 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:	N/A
General description of non-publicly known private datasets obtained from third parties	The dataset is mainly composed of text data sourced from multiple domains.
Additional comments (optional):	N/A

2.3 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:	N/A
Type of modality covered:	☐ Text ☐ Image ☐ Video ☐ Audio ☐ Other
Additional comments (optional):	If user input is used to construct training data, we apply secure encryption, strict de-identification, and anonymization to make it cannot be linked to any specific individual. Users are also given the right to opt out. For information on how to opt out of AI training, please refer to the DeepSeek Privacy Policy.

2.4 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:	DeepSeek generated synthetic data using its own AI models, including, where applicable, publicly available earlier DeepSeek 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:	DeepSeek generates synthetic data using internal models. For instance, enriching data in areas where specialized training data is harder to obtain, whether by domain, language, task, or format. Beyond that, these models can be applied to create examples for instruction following, reasoning, coding, multi-modal understanding, safety alignment, and evaluation tasks.
Additional comments (optional):	N/A

2.5 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:	An internal team of professionals undertakes the creation, labeling, and auditing of data across certain domains.
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?	☐ 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:	DeepSeek places emphasis on compliance with laws and regulations related to intellectual property, trade secrets, and personal privacy. For publicly accessible web data, DeepSeek implements measures before and during data collection to respect applicable rights reservations and opt-out signals relevant to text and data mining; the crawler is designed to respect robots.txt instructions and other standard web protocols, where applicable, and is not designed to circumvent captchas, paywalls, or password-protected content.
Additional comments (optional):	N/A

3.2 Removal of illegal content

General description of measures taken:	DeepSeek applies preprocessing and screening measures to screen for any proprietary information protected by intellectual property rights or trade secrets, or any content involving pornography, gambling, drugs, prejudice, discrimination, etc., in order to avoid or remove illegal content from the training data. These measures may include automatic filtering, manual inspection, keyword inspection, and classification model inspection methods.

3.3. Other information (optional)

Other relevant information about data processing (optional):	N/A