Public Summary of Training Data Content for Seed 2.0 Pro

Version of the Summary:	1.0
Last update:	31 July 2026

1. General information

1.1 Provider identification

Provider name and contact details:	ByteDance Nexus AI Pte. Ltd (**ByteDance**)
Authorised representative name and contact details:	Mikros Information Technology Ireland Limited

1.2 Model identification

Versioned model name(s):	Seed 2.0-Pro (Model name on Byteplus: Dola-Seed-2.0-Pro)
Model dependencies:	N/A
Date of placement of the model on the Union market:	March 2026

1.3 Modalities, overall training data size and other characteristics

Modality	Training data size	Types of content
Text	More than 10 trillion tokens	The model was pre-trained using a large-scale, diverse collection of data spanning a wide-range of domains and modalities, including datasets

		containing information represented as written language.
Image	More than 1 billion images	The model was pre-trained using a large-scale, diverse collection of data spanning a wide-range of domains and modalities, including datasets containing static and visual information (such as photographs and diagrams).
Audio	N/A	N/A
Video	10,000 to 1 million hours	The model was pre-trained using a large-scale, diverse collection of data spanning a wide-range of domains and modalities, including datasets containing videoclips.
Other	N/A	N/A

Latest date of data acquisition/collection for model training:	February 2026
Description of the linguistic characteristics of the overall training data:	The model's training data contains a diverse range of global languages, including languages from within Europe.
Other relevant characteristics of the overall training data:	The model was pre-trained using a large-scale, diverse collection of data spanning a wide-range of domains and modalities, including datasets containing information represented as written language, static and visual information (such as photographs and diagrams) and video clips. Seed 2.0 Pro's training dataset was carefully curated to maximize diversity, representativeness, and robustness, while appropriate filtering mechanisms are applied to filter for quality and safety in relation to low-quality and harmful content.
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
If yes, specify the modality(ies) of the content covered by the datasets concerned:	Text, image and video
List of large publicly available datasets:	We have used publicly available datasets across various sectors and stages of the model development pipeline. These publicly available datasets are subject to pre-processing such as data cleaning, deduplication, filtering, selection, etc to exclude unsafe or harmful content.
General description of other publicly available datasets not listed above:	See the description of linguistic characteristics and other relevant characteristics of overall training data in Section 1.
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 and image

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 yes, specify the modality(ies) of the content covered by the datasets concerned:	Text and image
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	We have obtained datasets from third-parties on a licensed basis. Such datasets primarily comprise content across text and image modalities.
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):	Bytespider
Purposes of the crawler(s):	In the context of ByteDance's AI model development, Bytespider is used to collect publicly accessible data, including text and image data to support training, testing, and validation of the AI model.
General description of crawler behaviour:	ByteDance's crawler is configured to avoid overloading websites and to operate in a manner consistent with ethical web scraping practices. Our crawler is designed to respect robots.txt instructions, and not to circumvent control measures such as paywalls or to access password-protected content.
Period of data collection:	Approximately January 2022 to June 2024.
Comprehensive description of the type of content and online sources crawled:	As above, we may use our crawler to collect publicly accessible data. The crawled content primarily consists of text and image and includes a variety of content types including academic, mathematical, code-related and general-purpose content.
Type of modality covered:	Text and image
Summary of the most relevant domain names crawled:	Crawled content includes a variety of publicly accessible online materials across various sectors, such as education, government, technology and research sectors.

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?	No
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:	ByteDance's models are trained using a proprietary mix of datasets, which tend to be large-scale and diverse. Such industry-standard datasets typically include a mixture of publicly available documents as described in this disclosure. Our datasets may also include data obtained in accordance with the relevant terms and conditions, privacy policy and pursuant to relevant user-controls as applicable.
Type of modality covered:	See above
Additional comments (optional):	We use data filtering processes to reduce personal information from training data, and 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
If yes, modality of the synthetic data:	Text and image
If yes, specify the general-purpose AI model(s) used to generate the synthetic data if available on the market:	Synthetic data was generated using a range of general-purpose AI models, including vision language models, image generation models and large language 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:	We may use fine-tuned versions of internal models to generate synthetic data.
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?	Yes
If yes, provide a narrative description of these data sources and the data:	We worked with our vendors to create labelled datasets so as to improve the model's capability on certain tasks, such as reasoning, coding and knowledge tasks.
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:	ByteDance's crawler is configured to avoid overloading websites and to operate in a manner consistent with ethical web scraping practices. Our crawler is designed to respect robots.txt instructions, and not to circumvent control measures such as paywalls or to access password-protected content.
Additional comments (optional):	N/A

3.2 Removal of illegal content

General description of measures taken:	As mentioned, we applied preprocessing and filtering methods, including filtering models focused on unsafe or harmful content.

3.3 Other information (optional)

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