Public Summary of Training Content for Hy3

Version of the Summary: V2

Last update: August 20, 2026

1. General information

1.1. Provider identification

Provider name and contact details: OriGen Tech Pte. Ltd., 10 Anson Road, #21-07, International Plaza, Singapore

Authorised representative name and contact details: Tencent International Service Europe B.V., Buitenveldertselaan 1-5, 1082 VA Amsterdam, the Netherlands

1.2. Model identification

Versioned model name(s): Hy3

Model dependencies: N/A

Date of placement of the model on the Union market: July 6, 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	Hy3 was trained on a large-scale, multilingual mixture of publicly available data, data accessed through partnerships, synthetic data, and human-generated text, including general web content, reference materials, technical documentation, source code, and other text, curated and filtered for quality and safety.
☐ 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

Latest date of data acquisition/collection for model training: The data used to train Hy3 includes different datasets from varying time periods, with data collected no later than March 2026.

Description of the linguistic characteristics of the overall training data:	Multilingual, with strong English and Chinese coverage and substantial representation across all EU official languages and other languages from around the world.
Other relevant characteristics of the overall training data:	The overall training corpus is designed for a text-output model. It aims to provide broad topical, geographic, linguistic, and format coverage.
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:	Training data includes filtered text from Common Crawl, with selection criteria such as deduplication and quality screening to filter the dataset for training.
General description of other publicly available datasets not listed above:	Other publicly available datasets include broad, multi-domain text datasets made available by third parties through public repositories, online platforms, and specialized websites, including reference materials, scientific and technical content, and source code. These datasets are global in scope, multilingual, and subject to preprocessing such as quality filtering, deduplication, and safety filtering before training. We use advanced data filtering processes to reduce personal information from training data.
Additional comments (optional):	From public web datasets, we take steps to identify and apply relevant rights-reservation and opt-out signals, including robots.txt signals for Sogou Web Spider and Sogou News Spider, where those signals are available for domains listed in the datasets.

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	Datasets obtained from third parties cover subject areas including educational materials, scientific reasoning, and mathematics, and consist primarily of textbooks, problem sets, etc. The content is predominantly in English and Chinese.
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 ☐ No
If yes, specify crawler name(s)/identifier(s):	Sogou Web Spider, and Sogou News Spider. For details please refer to https://www.sogou.com/docs/help/webmasters.htm#07
Purposes of the crawler(s):	Sogou Web Spider and Sogou News Spider are used to discover and scan websites, find information for building Sogou's search indexes, perform other product specific crawls, and for analysis.
General description of crawler behaviour:	Sogou Web Spider and Sogou News Spider are designed to respect robots.txt instructions. Sogou Web Spider and Sogou News Spider are not designed to circumvent captchas or paywalls or to access password-protected content.
Period of data collection:	Training data for Hy3 was collected from approximately 2023. The knowledge cut-off date for Hy3 is 28 February 2026. We may periodically update models after this date.
Comprehensive description of the type of content and online sources crawled:	Crawled data includes a broad range of publicly available online material, such as publicly available websites across educational, government, legal, and research sectors. This includes a wide variety of media types and languages.
Type of modality covered:	☒ Text ☐ Image ☐ Video ☐ Audio

	☐ Other
Summary of the most relevant domain names crawled:	The most relevant domains crawled include publicly available websites across educational, government, legal, and research sectors comprising a wide variety of media types and languages. Sogou Web Spider/Sogou News Spider is used to discover and scan websites, find information for building Sogou's search indexes, perform other product specific crawls, and for analysis.
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?	☐ 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):	N/A

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:	We generate synthetic data using a collection of generative AI models, including Hy2. Hy2 is a foundation large language model provided by Tencent. Synthetic data generated by these models include examples for instruction following, reasoning, coding, writing, role-playing, knowledge, safety, and evaluation.
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 ☒ No
If yes, provide a narrative description of these data sources and the data:	N/A
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:	We implement measures to respect rights reservations and opt-out signals relevant to text and data mining. For web data used for training. Sogou Web Spider and Sogou News Spider are designed to respect reservation and opt-out protocols, including robots.txt instructions for Sogou Web Spider and Sogou News Spider, as well as HTTP headers. For datasets obtained from third parties, we rely on those providers' implementation of rights-reservation and opt-out protocols and solutions.
Additional comments (optional):	N/A

3.2 Removal of illegal content

General description of measures taken:	Hy3 applies preprocessing and screening measures intended to avoid or remove illegal content under Union law (including child sexual abuse material, terrorist content) from training data. Measures may include automated filtering, keyword-based rules, hash-matching, and model-based classifiers to help identify and exclude unlawful material.

3.3. Other information (optional)

Other relevant information about data processing (optional):	To the extent the training data contains personal information, we use advanced data filtering processes to reduce the presence of such personal information.