The Public Summary of Training Content for the GLM-5 Family

Version of the Summary: *1.0*
Last update: *21/08/2026*

1. General information

1.1. Provider identification

Provider name and contact details: *JINGSHENG HENGXING TECHNOLOGY PTE. LTD.*

Authorised representative name and contact details: *N/A*

1.2. Model identification

Versioned model name(s): *The GLM-5 family — GLM-5, GLM-5.1, GLM-5.2 and GLM-5.3*

Model dependencies: *None.*

Date of placement of the model on the Union market: *Public release date: GLM-5 — February 2026; GLM-5.1 — April 2026; GLM-5.2 — June 2026; GLM-5.3 — August 2026.*

1.3 Modalities, overall training data size and other characteristics

This Section requires general information about the overall training data after pre-processing and before the training of the model.

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 Alternatively, specify the approximate size in a different measurement unit: *N/A.*	*Web documents, reference works, textual content including scientific text, source code, mathematical and scientific texts, multilingual text, synthetic reasoning data, long-context and labelled data used during reinforcement learning.*
☐ Image	☐ Less than 1 million images ☐ 1Million to1 billion images ☐ More than 1 billion images	*N/A*
☐ Audio[1]	☐ Less than 10 000 hours ☐ 10 000 to1 million hours	*N/A*

[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	
☐ 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 training data were collected approximately no later than June 2026.*

Description of the linguistic characteristics of the overall training data: *The training corpus is multilingual. Data collection was conducted without intentionally excluding any specific region.*

Other relevant characteristics of the overall training data: *The corpus is designed for a text-output general-purpose model.*

Additional comments (optional): *N/A*

2. List of data sources

This Section requires information about specific sources of data used to train the general-purpose AI model. In this section "dataset" should be understood as a single, pre-packaged collection of data. The filtering and pre-processing of data collected from the same pre-packaged collection should not be considered a new dataset to be disclosed separately in the sections below. If a particular dataset can be assigned to more than one of the categories below, providers should select the most relevant category and only report the dataset in that category, except in the case of synthetic data (see Section 2.5).

2.1. Publicly available datasets

This Section requires information about datasets that were used to train the model and which have been compiled by a third party, are made available publicly for free, and are readily downloadable as a whole or in predefined chunks, such as datasets and collections available on public repositories and online platforms, specialised websites, or snapshots of common crawl. The public availability of the datasets for free does not mean that the content at issue is necessarily free of rights since it may be subject to licensing arrangements or conditions of use (e.g., certain free and/or open licenses may determine the scope of the uses, including prohibiting uses relating to model training).

A dataset is considered to be "large" if the total data size for any one of the modalities contained in the dataset exceeds 3% of the size of all publicly available datasets for that modality used for training. The size of the dataset should be based on its size after pre-processing (for example filtering), and without splitting the dataset to prevent reporting circumvention.

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 *If so, please specify…*

List of large publicly available datasets: *Publicly available repositories, such as Common Crawl, open multilingual corpora, public platform datasets and publicly accessible source code datasets.*

General description of other *See the description of training data from Section 1.*

publicly available datasets not listed above:

Additional comments (optional): *N/A*

2.2 Private non-publicly available datasets obtained from third parties

This Section requires information about private non-publicly available datasets of third parties that are not publicly available and not disclosed under Section 2.1. These include:

1) *datasets for which transactional commercial licensing agreements were concluded between the provider and the rightsholders or their representatives, including by collective management organisations and legitimate content aggregators who have the right to collectively license works on behalf of rightsholders (Section 2.2.1);*
2) *other private datasets obtained through data intermediaries, non-publicly available databases and datasets of third parties for which transactional commercial licenses have <u>not</u> been concluded with rightsholders or their representatives (Section 2.2.2).*

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 *If so, please specify…*

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 so, please specify…*

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 comprehensive datasets from third parties pursuant to compliance commitments given by such third parties. Such datasets primarily comprise textual content.*

Additional comments (optional): *N/A*

2.3 Data crawled and scraped from online sources

This Section requires information about crawled, scraped data, or otherwise compiled from online sources directly by the provider of the model or on their behalf (i.e. excluding publicly available datasets already compiled by third parties and made available on platforms such as common crawl that are covered under Section 2.1).

Were crawlers used by the provider or on behalf of? ☒ Yes ☐ No

If yes, specify crawler name(s)/identifier(s):	*GLMBot*
Purposes of the crawler(s):	*The crawler collects publicly available online content, aiming to assemble a broad and diverse textual dataset that may be used to support model training, testing and improvement of processing capabilities.*
General description of crawler behaviour:	*The crawler is designed to respect robots.txt directives, to avoid overloading websites, and to refrain from circumventing access-control measures such as paywalls, CAPTCHAs, or password protection.*
Period of data collection:	*Up to June 2026.*
Comprehensive description of the type of content and online sources crawled:	*Predominantly multilingual text-based web content — reference sites, technical documentation, code-related pages, and mathematical/scientific resources.*
Type of modality covered:	☒ Text ☐ Image ☐ Video ☐ Audio ☐ Other *If so, please specify…*
Summary of the most relevant domain names crawled:	*.com, .org, .net and region-specific TLDs; code-hosting, documentation, reference, and math/science sites.*
Additional comments (optional):	*N/A*

2.4 User data

This Section requires information about user data collected by all services and products of the provider, including through mail services, social media platforms, content platforms or interaction with the providers' AI models and/or systems. This does not cover data licensed by users based on commercial transactional agreements described in Section 2.2.1., or customer data to fine-tune models for specific purposes.

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 *If so, please specify…*
Additional comments (optional):	*User-interaction data is used for training only where users have chosen to opt in. In these cases, we may, in accordance with our terms of service, privacy policy, and applicable law, use such content for training after applying appropriate privacy protection measures.*

2.5 Synthetic data

This Section requires information about synthetic data created by or on behalf of the provider for training the model directly on the outputs of another AI model, in particular through model distillation or model alignment (e.g. AI feedback through reinforcement learning). This does not include the use of AI models to

clean or enrich data (e.g. AI-generated metadata to enrich or modify a dataset, such as creating depth maps or text descriptions of images). In case this concerns publicly available datasets as described in Section 2.1, these should be reported in that Section of the Template. In case this concerns synthetic datasets created by third parties on behalf of the provider, these should be reported in this Section of the Template instead of in Section 2.2.2.

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 so, please specify…*
If yes, specify the general-purpose AI model(s) used to generate the synthetic data if available on the market:	*Synthetic data was generated predominantly using GLM's own models, supplemented by our earlier open-source 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:	*Internal fine-tuned GLM models were used to generate supplementary synthetic data to augment data coverage across diverse domains and tasks.*
Additional comments (optional):	*N/A*

2.6 Other sources of data

This Section requires information about data that does not fall under any of the categories in the previous Sections, for example data collected from offline sources, self-digitised media (e.g., digitised analog text context, images), datasets labelled by humans commissioned by the provider, or human generated data through reinforcement learning.

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:	*A portion of the data is created, labelled, and audited by professional teams for certain domains (e.g., alignment, evaluation, and domain-specific reasoning).*
Additional comments (optional):	*N/A*

3. Data processing aspects

3.1. Respect of reservation of rights from text and data mining exception or limitation

This Section concerns measures implemented by the provider to identify and comply with the reservation of rights from the text and data mining (TDM) exception or limitation expressed pursuant to Article 4(3) of Directive (EU) 2019/790, as outlined in the copyright policy put in place by the provider in accordance with Article 53(1)(c) AI Act.

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 before and during data collection to respect applicable rights and 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 and is not designed to circumvent CAPTCHAs, paywalls, or password-protected content.*
Additional comments (optional):	*N/A*

3.2 Removal of illegal content

This Section concerns measures taken to avoid or remove illegal content under Union law from the training data (such as blacklists, keywords, and model-based classifiers), without requiring disclosure of specific details about the provider's internal business practices or trade secrets. Such measures are advisable if the training data is likely to include illegal or unlawful content under Union law, in particular child sexual abuse material and terrorist content and the non-authorised use of material protected by intellectual property rights. Such measures do not include data selection practices, for example to increase the capability of the model.

7

General description of measures taken:	*We apply multi-layer preprocessing and screening to detect and remove content that is illegal from the training data. These measures may combine automated detection techniques, (including keyword filtering, rules-based filtering, model-based classification, hash/blacklist matching) with deduplication and quality controls to support compliance and data integrity.*

3.3. Other information (optional)

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