

Summary of Training Content for General-Purpose AI Model

Version of the Summary: Version 1.0
Last update: July 30, 2026

1. General information

1.1. Provider identification

Provider name and contact details: Thinking Machines Lab Inc.
95 3rd Street, 2nd Floor, San Francisco, CA 94103

Authorised representative name and contact details: EDPO may be contacted at info@edpo.com

1.2. Model identification

Versioned model name(s): Inkling-Small <https://thinkingmachines.ai/model-card/inkling-small/>

Model dependencies: N/A. This model is the first version of its kind released.

Date of placement of the model on the Union market: July 30, 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	Inkling-Small was trained on a mixture of publicly available content, content acquired through partnerships, synthetic content, and generated content, including general web content, reference materials, technical documentation, source code, and other text, curated and filtered.
☒ Image	☐ Less than 1 million images ☐ 1Million to1 billion images ☒ More than 1 billion images	Inkling-Small was trained on a mixture of publicly available images, images acquired through partnerships, synthetic images, and generated images, including images from different kinds of web content, curated and filtered.
☒ Audio ¹	☐ Less than 10 000 hours ☐ 10 000 to1 million hours ☒ More than 1 million hours	Inkling-Small was trained on a mixture of publicly available audio and audio acquired through partnerships, including audio from the public web, curated and filtered.

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’.

☒ Video	☐ Less than 10 000 hours ☐ 10 000 to 1 million hours ☒ More than 1 million hours	Inkling-Small was trained on a mixture of publicly available video, video acquired through partnerships, and generated video, including video from the public web, curated and filtered.
☐ Other	<i>Specify the modality and for each one indicate approximate size and unit of measurement</i>	N/A.

Latest date of data acquisition/collection for model training:

The latest date of data collection for the model training is July 2026. Thinking Machines Lab collects and uses datasets to train and improve its generative AI services on an ongoing basis.

Description of the linguistic characteristics of the overall training data:

The training datasets used by Thinking Machines Lab are multilingual, with languages including English.

Other relevant characteristics of the overall training data:

N/A

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 *If so, please specify...*

List of large publicly available datasets:

Our large publicly available datasets include Common Crawl (<https://commoncrawl.org/>)

General description of other publicly available datasets not listed above:

Thinking Machines Lab used various publicly available image-captioning datasets. Thinking Machines Lab also used various publicly available datasets containing text, images, audio, and/or video content.

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 *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: Thinking Machines Lab has acquired text, image, audio, and video content from various third parties. The content includes items in the public domain as well as content that may be subject to intellectual property protection in some jurisdictions.

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): N/A

Purposes of the crawler(s): Crawlers were used to download content from publicly available sources from the internet for the purpose of model training.

General description of crawler behaviour: Thinking Machines Lab's policy is that crawlers should not circumvent captchas, password-protections, or other access controls, and respect robots.txt.

Period of data collection: From 2025 to 2026

Comprehensive description of the type of content and online sources crawled: Crawlers obtained a broad variety of content types, including text and images, drawn from publicly available sources. Content types include publicly available webpages, PDF documents, academic papers, and code repositories. Content includes academic, scientific, mathematical, code-related, and general-purpose content (in English and other languages).

Type of modality covered: ☒ Text ☒ Image ☐ Video ☐ Audio

Summary of the most relevant domain names crawled:

☐ Other *If so, please specify...*

The most relevant domains used to train Inkling-Small include resources and repositories spanning academic, scientific, mathematical, code-related, and general-purpose (in English and other languages) content.

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 *If so, please specify...*

N/A

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 so, please specify...*

If yes, specify the general-purpose AI model(s) used to generate the synthetic data if available on the market:

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:

Thinking Machines Lab models and third party models were used for various tasks in the synthetic data pipeline, including, for example, augmentation and evaluation, and to generate examples.

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:

We work with vendors and experts on specialty datasets, e.g. safety.

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:

Thinking Machines Lab's policy is that any crawlers and data partners should engage in legal collection and should not circumvent captchas, password-protections, or other access controls, and respect robots.txt.

Additional comments (optional):

N/A

3.2 Removal of illegal content

General description of measures taken:

Thinking Machines Lab takes steps to avoid the collection of data that is, or is likely, illegal data, and performs some filtering to remove illegal data.

3.3. Other information (optional)

Other relevant information about data processing (optional):

N/A