ANTHROP\C

Public Summary of Training Content
Claude Haiku 5.5

Version of the Summary:	Version #1
Last update:	October 7, 2026

1. General information

1.1. Provider identification

Provider name and contact details:	Anthropic Ireland, Limited 6th Floor South Bank House, Barrow Street, Dublin 4, Dublin Ireland
Authorised representative name and contact details:	Not applicable

1.2. Model identification

Versioned model name(s):	Claude Haiku 5.5 Model Card: www.anthropic.com/system-cards
Model dependencies:	Not applicable
Date of placement of the model on the Union market:	October 7, 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	Data used in model training includes an array of text types, including short and long-form texts, software code, synthetic text, prose in a variety of languages, mathematical data, and prompts and preference data used during reinforcement learning.

Image	☐ Less than 1 million images ☐ 1Million to1 billion images X More than 1 billion images	Data used in model training includes an array of image types, including photographs, interleaved text and images from websites, computer graphics, and prompts and preference data used during reinforcement learning.
Video	Not applicable	
Audio	Not applicable	
Other	Not applicable	

Latest date of data acquisition/collection for model training:	A number of different datasets, with varying publication and cut-off dates, are included in the training corpus, with some data being acquired/collected up to September 2026.
Description of the linguistic characteristics of the overall training data:	Training sources deliberately include a diverse range of global languages, both European and non-European, including those with relatively high numbers of speakers (*e.g.,* English, Chinese, French, Spanish) as well as comparably lower volumes (*e.g.,* Basque, Breton, Korean).

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
List of large publicly available datasets:	Data used in model training is derived from several publicly accessible repositories, notably including Common Crawl, a repository of web crawl data, as well as specialized datasets available through platforms like GitHub and HuggingFace.
General description of other publicly available datasets not listed above:	Data from other publicly available datasets is included in the training corpus, including mathematical data and some image and caption 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
If yes, specify the modality(ies) of the content covered by the datasets concerned:	Text, 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	Yes

obtained from providers of private databases, or data intermediaries?	
If yes, specify the modality(ies) of the content covered by the datasets concerned:	Text, Image
If publicly known, list private datasets obtained from other third parties:	Not applicable
General description of non-publicly known private datasets obtained from third parties	We obtain non-publicly known private datasets from third parties covering diverse domains and content types.

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):	ClaudeBot
Purposes of the crawler(s):	ClaudeBot collects web content that could potentially contribute to the model's training.
General description of crawler behaviour:	We aim to minimize disruption to website owners and be thoughtful about how quickly ClaudeBot crawls domains, including by respecting crawl-delay and disallow directives in robots.txt files, where appropriate, and respecting anti-circumvention technologies such as paywalls, password protection, and CAPTCHAs. More information on our crawlers and how they work can be found at: https://support.claude.com/en/articles/8896518-does-anthropic-crawl-data-from-the-web-and-how-can-site-owners-block-the-crawler.
Period of data collection:	March 2024 - September 2026
Comprehensive description of the type of content and online sources crawled:	The crawlers may be exposed to a wide variety of content and online sources, including most forms of publicly available online data.
Type of modality covered:	Text, Image
Summary of the most relevant domain names crawled:	Data used in model training which is derived from data crawled and scraped from online sources includes technical documentation, open-source software, predominantly text-based reference sites, document sharing sites, and math sites. Top-level domains such as .com, .org, and .net are included alongside sites from a range of different countries.

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
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:	To the extent permitted by Anthropic's terms of service, privacy policy, and other contracts, and in line with applicable law, if a user explicitly reports feedback or bugs to us (e.g., via thumbs and feedback buttons) or otherwise chooses to allow us to use their data, then chats and coding session data may be used in model training. More information is available at https://privacy.claude.com/en/articles/7996868-is-my-data-used-for-model-training We may also incorporate data derived from Anthropic employees' use of internal-only model versions.
Type of modality covered:	Text

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, Image
If yes, specify the general-purpose AI model(s) used to generate the synthetic data if available on the market:	Synthetic data was provided by speech to text models, large language models (LLM), and vision-language models (VLM).
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:	Some synthetic data used in training was generated by Anthropic models not available on the market.

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:	A portion of our training data comes from acquired physical texts.

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
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:	ClaudeBot respects crawl-delay and disallow directives in robots.txt files where appropriate, as well as anti-circumvention technologies such as paywalls, password protection, and CAPTCHAs. More information on our crawlers and how they work can be found at: https://support.claude.com/en/articles/8896518-does-anthropic-crawl-data-from-the-web-and-how-can-site-owners-block-the-crawler.

3.2 Removal of illegal content

General description of measures taken:	We take a number of protective measures to remove illegal content from the training corpus such as active filtering, scoring, moderation, and blocking.

3.3. Other information (optional)

Other relevant information about data processing (optional):	Not applicable