

Public Summary of Training Content

This document is Canva's public summary of the content used to train Craft-1 2K, prepared in accordance with Article 53(1)(d) of the EU AI Act.

Document version:	Version 1
Last update:	22 October 2026

1. General information

1.1. Provider identification

Provider name and contact details:	Canva Pty Ltd with registered address at 110 Kippax St, Surry Hills, NSW 2010, Australia
Authorised representative name and contact details:	European Data Protection Office (EDPO), 60 Merrion Square South, Dublin 2, D02 HE24, Ireland (info@edpo.com)

1.2. Model identification

Versioned model name(s):	Craft-1 2K
Model dependencies:	Not applicable
Date of placement of the model on the EU market:	22 October 2026

1.3 Modalities, overall training data size and other characteristics

Modality	Training data size	Types of content
Text	☐ Less than 1 billion tokens ☐ 1 billion to 10 trillion tokens ☒ More than 10 trillion tokens	The training corpus for Craft-1 2K includes text associated with images (such as captions), structured JSON layouts paired with images, human-created annotations and labels, synthetic text, and prompts and preference data used during reinforcement learning.
Image	☐ Less than 1 million images ☐ 1 million to 1 billion images ☒ More than 1 billion images	The training corpus for Craft-1 2K includes a diverse range of image data, including photographs, illustrations, visual art, designs (such as posters, slides, charts and diagrams), video-derived visual content, and synthetic visual content.

Audio	Not applicable	
Video	Not applicable	
Other	Not applicable	

Latest date of data acquisition/collection for model training:	The data used to train Craft-1 2K includes datasets from different time periods, with the most recent data collected in October 2026. The model may undergo additional training and/or fine-tuning.	
Description of the linguistic characteristics of the overall training data:	More than 50 languages have been detected in text embedded in image data, including EU official languages and other languages from around the world.	
Other relevant characteristics of the overall training data:	Craft-1 2K is an image generation model trained on a broad corpus of image data and text associated with images. The variety of subjects, styles, formats and languages helps the model understand prompts and visual context to generate a wide range of images.	
Additional comments (optional):	Not applicable	

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 <i>If so, please specify...</i>
List of <u>large</u> publicly available datasets:	The training data includes publicly available data from Common Crawl and image-text datasets hosted on public repositories, such as Hugging Face.
General description of other publicly available datasets not listed above:	Other publicly available data includes broad, mixed-domain image datasets, rendered arXiv documents, and other publicly accessible image collections.
Additional comments (optional):	Not applicable

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	☒ Text ☒ Image ☐ Video ☐ Audio

concerned:	☐ Other <i>If so, please specify...</i>
------------	--

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 <i>If so, please specify...</i>
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	The training data includes datasets obtained from third parties, covering a range of subjects and visual styles.
Additional comments (optional):	Not applicable

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):	ImageBot/1.0
Purposes of the crawler(s):	The crawler collects publicly accessible images from the web for training purposes.
General description of crawler behaviour:	The crawler is designed to respect machine-readable opt outs, including robots.txt instructions and AI-specific signals. It only accesses publicly available data and is not designed to circumvent paywalls, captchas or other technical measures that control access to content. It provides an opt-out mechanism for website owners and excludes certain domains and regions. Images are filtered for known child sexual abuse material (CSAM) before anything is stored.
Period of data collection:	Approximately May 2026 to September 2026
Comprehensive description of the type of content and online sources crawled:	The crawler is exposed to a wide variety of publicly available data from a range of online sources, including images and associated text descriptions.
Type of modality covered:	☒ Text ☒ Image ☐ Video ☐ Audio ☐ Other <i>If so, please specify...</i>
Summary of the most relevant domain names crawled:	The crawled images come from publicly accessible websites across the open web. The most common top-level domains are .com, .org and .net.

Additional comments (optional):	Not applicable
---------------------------------	----------------

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:	Where permitted and consistent with users' privacy controls, user content and interactions with Canva may be used to train and improve our models, including during supervised fine-tuning.
Type of modality covered:	☒ Text ☒ Image ☐ Video ☐ Audio ☐ Other <i>If so, please specify...</i>
Additional comments (optional):	Not applicable

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 <i>If so, please specify...</i>
If yes, specify the general-purpose AI model(s) used to generate the synthetic data if available on the market:	Synthetic data is used to support specific training objectives. This includes synthetic screenshots of app and website interfaces, created by populating templates with generated text, product images and other visual elements. We also use synthetic text generated by large language models for image captioning and prompt rewriting.
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 is generated by Canva models.
Additional comments (optional):	Not applicable

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:	The training data includes Canva-owned image data, including Canva templates.

Additional comments (optional):	Not applicable
---------------------------------	----------------

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 respect rights reservations expressed through machine-readable opt-outs, including robots.txt instructions and AI-specific signals. Our crawlers are designed to only access publicly available data and are not designed to circumvent paywalls, captchas or other technical measures that control access to content. They also provide an opt-out mechanism for website owners and exclude certain domains and regions.
Additional comments (optional):	Not applicable

3.2 Removal of illegal content

General description of measures taken:	We take reasonable measures to remove illegal content from training data before storage, including filtering out known child sexual abuse material (CSAM) and other disallowed content.
--	---

3.3. Other information (optional)

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