cohere

# Public Summary of Training Content for Command A+

| | |
|---|---|
| Version of the Summary: | v1 |
| Last update: | 31 July 2026 |

## 1. General information

| 1.1 Provider identification | |
|---|---|
| Provider name and contact details: | Cohere Germany GmbH<br>Contact: support@cohere.com, use subject line “AI Act Request” |
| Authorised representative name and contact details: | n/a |
| **1.2 Model identification** | |
| Versioned model name(s): | **Command A+**<br>Public documentation is available at https://docs.cohere.com/docs/command-a-plus |
| Model dependencies: | n/a |
| Date of placement of the model on the Union market: | 20 May 2026 |

## 1.3 Modalities, overall training data size and other characteristics

| Modality | Training data size | Types of content |
|---|---|---|
| ☑ Text | ☐ Less than 1 billion tokens<br>☐ 1 billion to 10 trillion tokens<br>☑ More than 10 trillion tokens | Command A+ was trained on a diverse set of text data sourced from publicly available content and proprietary datasets accessed through partnerships or developed by Cohere, including general web content, technical documentation, code, and other text across various fields such as education, government, science, and technology. |
| ☑ Image | ☐ Less than 1 million images<br>☐ 1 million to 1 billion images<br>☑ More than 1 billion images | Command A+ was trained on a diverse corpus of images sourced from publicly available content and proprietary datasets, with a focus on enterprise-relevant content such as tables, charts, pdfs, and infographics. |
| ☐ Audio | n/a | n/a |
| ☐ Video | n/a | n/a |
| ☐ Other | n/a | n/a |

### 1.3.1 Other characteristics

| | |
|---|---|
| Latest date of data acquisition/collection for model training: | Up to April 2026 |
| Description of the linguistic characteristics of the overall training data: | Multilingual, covering 48 languages, including all official European Union languages. |
| Other relevant characteristics of the overall training data: | Command A+ is optimized and trained to excel at enterprise-relevant tasks, such as enterprise reasoning, coding, tool use, multilingual tasks, and multimodal document understanding. |
| Additional comments (optional): | |

# 2. List of data sources

| 2.1 Publicly available datasets | |
|---|---|
| Have you used publicly available datasets to train the model? | ☑ Yes<br>☐ No |
| If yes, specify the modality(ies) of the content covered by the datasets concerned: | ☑ Text<br>☑ Image<br>☐ Audio<br>☐ Video<br>☐ Other |
| List of large publicly available datasets: | The training data for Command A+ includes text from Common Crawl (https://commoncrawl.org/). Cohere employs various techniques to curate data prior to using it in training to support data quality and suitability for training, including deduplication, filtering for toxic, harmful or otherwise unsuitable content, and quality filtering. |
| General description of other publicly available datasets not listed above: | Other publicly available datasets include various sources of text, and image content across fields such as education, government, science, technology, that are relevant to enterprise reasoning, coding, tool use, multilingual tasks, and multimodal document understanding. See the description provided in Section 1.3 for details. |
| Additional comments (optional): | |
| **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<br>☐ No<br>☑ Other |
| If yes, specify the modality(ies) of the content covered by the datasets concerned: | ☑ Text<br>☐ Image<br>☐ Audio<br>☐ Video<br>☐ Other |
| Additional comments (optional): | Some of the third parties identified in Section 2.2.2 license their datasets. |

| 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<br>☐ No |
| If yes, specify the modality(ies) of the content covered by the datasets concerned: | ☑ Text<br>☑ Image<br>☐ Audio<br>☐ Video<br>☐ Other |
| If publicly known, list private datasets obtained from other third parties: | None of the datasets subject to this section 2.2.2 are publicly known. |
| General description of non-publicly known private datasets obtained from third parties: | Cohere partners with various third parties to source multimodal and multilingual training data across fields such as education, government, science, technology, that are relevant to enterprise reasoning, coding, tool use, multilingual tasks, and multimodal document understanding. See the description provided in Section 1.3 for details. |
| Additional comments (optional): | |

## 2.3 Data crawled and scraped from online sources

| | |
|---|---|
| Were crawlers used by the provider or on behalf of? | ☑ Yes<br>☐ No |
| If yes, specify crawler name(s)/ identifier(s): | Cohere makes information about its web crawlers available at https://docs.cohere.com/docs/cohere-web-crawlers. |
| Purposes of the crawler(s): | Third parties may make use of crawlers in the process of developing datasets identified in Section 2.2.2. Prior to August 2025, Cohere collected certain web data using a crawler bot that is no longer in use. |
| General description of crawler behaviour: | It is Cohere's policy to require that crawlers be designed to respect robots.txt and other effective technological measures, such as technological denial or restrictions of access like paywalls or subscription models. |
| Period of data collection: | Up to April 2026 |

| | |
|---|---|
| Comprehensive description of the type of content and online sources crawled: | Multimodal training data across fields such as education, government, science, and technology, that are relevant to enterprise reasoning, coding, tool use, multilingual tasks, and multimodal document understanding. See the description provided in Section 1.3 for details. |
| Type of modality covered: | ☑ Text<br>☑ Image<br>☐ Audio<br>☐ Video<br>☐ Other |
| Summary of the most relevant domain names crawled: | The most relevant domains used to train Command A+ include publicly available content in a variety of media types and languages, including educational, government, science and technology, code, and general-purpose content. |
| Additional comments (optional): | |

## 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<br>☐ No |
| Was data collected from user interactions with the provider's other services or products used to train the model? | ☑ Yes<br>☐ No |
| If yes, provide a general description of the provider's services or products that were used to collect the user data: | In most cases, Cohere's customers use Cohere models in their own environments or in third party environments, meaning Cohere has no access to inputs submitted to its models. Where permitted by a user via user controls and Cohere's relevant terms of service, de-identified data from the use of Cohere models on Cohere-hosted environments (e.g. user inputs) may be used in limited circumstances. |
| Type of modality covered: | ☑ Text<br>☐ Image<br>☐ Audio<br>☐ Video<br>☐ Other |
| Additional comments (optional): | Additional information on user controls and our privacy practices is available at:<br>https://cohere.com/privacy<br>https://cohere.com/enterprise-data-commitments |

| 2.5 Synthetic data | |
|---|---|
| Was synthetic AI-generated data created by the provider or on their behalf to train the model? | ☑ Yes<br>☐ No |
| If yes, modality of the synthetic data: | ☑ Text<br>☑ Image<br>☐ Audio<br>☐ Video<br>☐ Other |
| If yes, specify the general-purpose AI model(s) used to generate the synthetic data if available on the market: | Cohere generated synthetic data using its own models, including Command A, and other internal 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: | We may use other AI models to generate synthetic data for various purposes, such as to generate examples, data augmentation, and data evaluation. |
| Additional comments (optional): | |
| **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<br>☐ No |
| If yes, provide a narrative description of these data sources and the data: | Cohere may work with vendors, annotators, and other third-party experts to create or improve specialized datasets. For example, we work with specialized professionals to create or improve training data that represents enterprise-relevant tasks to optimize our model's performance on such tasks. |
| Additional comments (optional): | |

# 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<br>☐ 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: | Cohere implements measures to respect applicable rights reservations and opt-out signals relevant to the TDM exception. This includes data deduplication and filtering techniques as well as crawlers that are designed to respect robots.txt and not bypass or circumvent effective technological measures, such as technological denial or restrictions of access like paywalls or subscription models. |
| Additional comments (optional): | |
| **3.2 Removal of illegal content** | |
| General description of measures taken: | Cohere employs various techniques to support data quality and suitability for training, including deduplication, filtering for toxic, harmful or otherwise unsuitable content, and quality filtering. These measures are intended to reduce the inclusion of illegal or unlawful content and improve the safety and reliability of the training corpus. Cohere is also a member of the Internet Watch Foundation, which supports efforts to identify, report and remove child sexual abuse material online. |
| **3.3 Other information (optional)** | |
| Other relevant information about data processing (optional): | |