

Public Summary of Training Content

Version of the Summary: *v.1*

Last update: *31/07/2025*

1. General information

1.1. Provider identification

Provider name and contact details: *Mistral AI
15 rue des Halles, 75001 Paris FRANCE*

Authorised representative name and contact details: *NA.*

1.2. Model identification

Versioned model name(s): *This Public Summary of Training Content applies to Mistral Small 4.*

Model dependencies: *NA.*

Date of placement of the model on the Union market: *March 16, 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	<i>The text dataset is a large-scale, multilingual text dataset, comprising highly general content originating from publicly available text datasets and user data, as well as highly specialized and technical datasets, both synthetically generated and human-curated by third-party providers.</i>
☒ Image	☐ Less than 1 million images ☒ 1Million to1 billion images ☐ More than 1 billion images	<i>The image dataset is composed of a multimodal content, including images, sourced from publicly available datasets, third-party providers, and user-provided content, covering diagrams, objects, and combined text-image datasets.</i>
☐ Audio	☐ Less than 10 000 hours ☐ 10 000 to1 million hours ☐ More than 1 million hours	<i>NA.</i>
☐ Video	☐ Less than 10 000 hours ☐ 10 000 to1 million hours ☐ More than 1 million hours	<i>NA.</i>

☐ Other	Specify the modality and for each one indicate approximate size and unit of measurement	NA.
--------------------------------	---	-----

Latest date of data acquisition/collection for model training:

The training data is composed of several datasets, with varying periods of data collection. The latest date of data collection was July 2025.

Description of the linguistic characteristics of the overall training data:

The training data is multilingual in its coverage, ensuring broad representation of official EU languages.

Other relevant characteristics of the overall training data:

The overall training data is curated and optimized to improve the model's multimodal, multilingual, and text-generation capabilities.

Additional comments (optional):

NA.

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:

The datasets used to train the model include Common Crawl.

General description of other publicly available datasets not listed above:

Additionally, Mistral AI used a combination of other publicly-available datasets to train Mistral Small 4.

These datasets covered text and image modalities, and included: broad, general-reference datasets; highly specialized, academic datasets or datasets addressing complex tasks such as STEM, coding and/or reasoning tasks; and government-produced, administrative, or legal datasets.

Additional comments (optional):

NA.

2.2 Private non-publicly available datasets obtained from third parties

2.2.1. Datasets commercially licensed by rightsholders or their representatives

☐ Yes ☐ No ☒ Other

Have you concluded transactional commercial licensing agreement(s) with rightsholder(s) or with their representatives?

Mistral AI concludes data access agreements with rights holders or their representatives for access to non-publicly available datasets.

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:

NA.

General description of non-publicly known private datasets obtained from third parties

Mistral AI works with a variety of third-party providers for access to synthetically generated and human-curated datasets. These datasets are obtained and curated using industry-standards and accessed under commercial or partnership agreements with the relevant contractual guarantees.

Additional comments (optional):

NA.

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):

NA.

Purposes of the crawler(s):

Crawlers were used to collect publicly available sources on the internet.

General description of crawler behaviour:

Our crawlers are designed to respect robots.txt, extract information from lawfully accessible publicly available sources, and to not circumvent technological measures.

Period of data collection:

Up to July 2025.

Comprehensive description of the type of content and online sources crawled:

Crawlers obtained a wide range of publicly-accessible content types, including text, images, multimodal content, code and metadata from general-knowledge domain names, and more specialised websites.

Type of modality covered:

☒ Text ☒ Image ☐ Video ☐ Audio
☐ Other If so, please specify...

Summary of the most relevant domain names crawled:

The most relevant domain names crawled included generalist websites, highly-specialised resources such (academic and technical repositories) and government-maintained,

administrative or legal domains (such as patent portals or other document-hosting sites).

Additional comments (optional):

NA.

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:

Mistral Small 4 was trained on user interaction data, subject to user opt-out from training and strict privacy safeguards. For more information see our [Privacy Policy](#).

Type of modality covered:

☒ Text ☒ Image ☐ Video ☐ Audio
☐ Other *If so, please specify...*

Additional comments (optional):

NA.

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:

Mistral Small 4 was trained using synthetic data generated by other models such as internal Mistral AI models. This data was combined with additional synthetic datasets from third-party providers.

Additional comments (optional):

NA.

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:

NA.

Additional comments (optional):

NA.

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:

Mistral AI enters into contractual arrangements with third parties that respect opt-out protocols; Mistral AI uses crawlers that are designed to collect from lawfully accessible sources, respect robots.txt instructions, and not to circumvent technological measures. For more information, please see our website docs.mistral.ai.

Additional comments (optional):

NA.

3.2 Removal of illegal content

General description of measures taken: *Mistral AI implements a multi-layered approach to avoid or remove illegal content from its training data. At the data acquisition or collection stage, Mistral AI sources high-quality, publicly available datasets which are subject to industry-recognized mechanisms - such as blacklisting, robots.txt, and automated classification - to exclude illegal and harmful content.*

At the data processing stage, Mistral AI applies an additional combination of model-specific and model-agnostic measures, including proprietary filters, deduplication, and safety classifiers to further identify and remove any residual illegal content.

3.3. Other information (optional)

Other relevant information about data processing (optional): *NA.*