```
---
title: "Models Overview"
slug: "docs/models"

hidden: false
description: "Cohere has a variety of models that cover many different use cases. If you
need more customization, you can train a model to tune it to your specific use case."
image: "../../assets/images/672b039-cohere_docs_preview_image_1200x630_copy.jpg"
keywords: "large language models, generative AI models"

createdAt: "Thu Apr 20 2023 18:10:10 GMT+0000 (Coordinated Universal Time)"
updatedAt: "Mon Jun 10 2024 15:57:07 GMT+0000 (Coordinated Universal Time)"
---


Cohere has a variety of models that cover many different use cases. If you need more
customization, you can [train a model](/docs/fine-tuning) to tune it to your specific use
case.

Cohere models are currently available on the following platforms:

- [Cohere’s proprietary platform](https://dashboard.cohere.com/playground/chat)
- [Amazon SageMaker](https://aws.amazon.com/marketplace/seller-profile?id=87af0c85-6cf9-
4ed8-bee0-b40ce65167e0)
- [Amazon Bedrock](https://us-west-2.console.aws.amazon.com/bedrock/home?region=us-west-
2#/providers?model=cohere.command-r-plus-v1:0)
- [Microsoft Azure](https://ai.azure.com/explore/models/?tid=694fed05-7f6d-4ab2-8c38-
9afb438eab6f&selectedCollection=cohere)
- [Oracle GenAI Service](https://www.oracle.com/artificial-intelligence/generative-
ai/generative-ai-service/)

At the end of each major sections below, you'll find technical details about how to call a
given model on a particular platform.

## What can These Models Be Used For?

In this section, we'll provide some high-level context on Cohere's offerings, and what the
strengths of each are.

- The Command family of models includes [Command](https://cohere.com/models/command?
_gl=1*15hfaqm*_ga*MTAxNTg1NTM1MS4xNjk1MjMwODQw*_ga_CRGS116RZS*MTcxNzYwMzYxMy4zNTEuMS4xNzE3Nj
AzNjUxLjIyLjAuMA..), [Command R](/docs/command-r), and [Command R+](/docs/command-r-plus).
Together, they are the text-generation LLMs powering conversational agents, summarization,
copywriting, and similar use cases. They work through the [Chat](/reference/chat) endpoint,
which can be used with or without [retrieval augmented generation](/docs/retrieval-
augmented-generation-rag) (RAG).
- [Rerank](https://cohere.com/blog/rerank/?
_gl=1*1t6ls4x*_ga*MTAxNTg1NTM1MS4xNjk1MjMwODQw*_ga_CRGS116RZS*MTcxNzYwMzYxMy4zNTEuMS4xNzE3Nj
AzNjUxLjIyLjAuMA..) is the fastest way to inject the intelligence of a language model into
an existing search system. It can be accessed via the [Rerank](/reference/rerank-1)
endpoint.
- [Embed](https://cohere.com/models/embed?
_gl=1*1t6ls4x*_ga*MTAxNTg1NTM1MS4xNjk1MjMwODQw*_ga_CRGS116RZS*MTcxNzYwMzYxMy4zNTEuMS4xNzE3Nj
AzNjUxLjIyLjAuMA..) improves the accuracy of search, classification, clustering, and RAG
results. It also powers the [Embed](/reference/embed) and [Classify](/reference/classify)
endpoints.

## Command

Command is Cohere's default generation model that takes a user instruction (or command) and
generates text following the instruction. Our Command models also have conversational
capabilities which means that they are well-suited for chat applications.

| Model Name               | Description
```

```
| Context Length | Maximum Output Tokens | Endpoints
|
|--------------------------|-----------------------------------------------------------------
-----------------------------------------------------------------------------------------------
-----------------------------------------------------------------------------------------------
-----------------------------------------------------------------------------------------------
-------------------------|----------------|-----------------------|-------------------------
--------------------------------------------------------------------|
| `command-r-plus-08-2024` | `command-r-plus-08-2024` is an update of the Command R+ model,
delivered in August 2024. Find more information [here]
(https://docs.cohere.com/changelog/command-gets-refreshed)
| 128k           | 4k                    | [Chat](/reference/chat)
|
| `command-r-plus-04-2024` | Command R+ is an instruction-following conversational model
that performs language tasks at a higher quality, more reliably, and with a longer context
than previous models. It is best suited for complex RAG workflows and multi-step tool use.
| 128k           | 4k                    | [Chat](/reference/chat)
|
| `command-r-plus`         | `command-r-plus` is an alias for `command-r-plus-04-2024`, so
if you use `command-r-plus` in the API, that's the model you're pointing to.
| 128k           | 4k                    | [Chat](/reference/chat)
|
| `command-r-08-2024`      | `command-r-08-2024` is an update of the Command R model,
delivered in August 2024. Find more information [here]
(https://docs.cohere.com/changelog/command-gets-refreshed)
| 128k           | 4k                    | [Chat](/reference/chat)
|
| `command-r-03-2024`      | Command R is an instruction-following conversational model that
performs language tasks at a higher quality, more reliably, and with a longer context than
previous models. It can be used for complex workflows like code generation, retrieval
augmented generation (RAG), tool use, and agents.
| 128k           | 4k                    | [Chat](/reference/chat)
|
| `command-r`              | `command-r` is an alias for `command-r-03-2024`, so if you use
`command-r` in the API, that's the model you're pointing to.
| 128k           | 4k                    | [Chat](/reference/chat)
|
|                          |
|                | 4k                    |
|
| `command`                | An instruction-following conversational model that performs
language tasks with high quality, more reliably and with a longer context than our base
generative models.
| 4k             | 4k                    | [Chat](/reference/chat),  <br/>[Summarize]
(/reference/summarize)                                                                  |
| `command-nightly`        | To reduce the time between major releases, we put out nightly
versions of command models. For `command`, that is `command-nightly`.  <br/><br/>Be advised
that `command-nightly` is the latest, most experimental, and (possibly) unstable version of
its default counterpart. Nightly releases are updated regularly, without warning, and are
not recommended for production use. | 128k           | 128k           | [Chat]
(/reference/chat)
|
| `command-light`          | A smaller, faster version of `command`. Almost as capable, but
a lot faster.
| 4k             | 4k                    | [Chat](/reference/chat),  <br/>[Summarize]
(/reference/summarize-2)                                                                |
| `command-light-nightly`  | To reduce the time between major releases, we put out nightly
versions of command models. For `command-light`, that is `command-light-nightly`.  <br/>
<br/>Be advised that `command-light-nightly` is the latest, most experimental, and
(possibly) unstable version of its default counterpart. Nightly releases are updated
regularly, without warning, and are not recommended for production use. | 4k             |
4k                    | [Chat](/reference/chat)
|
| `c4ai-aya-23-35b`                     | The 35B version of the [Aya 23 model]
(https://huggingface.co/CohereForAI/aya-23-35B). Pairs a highly performant pre-trained
```

```
Command family of models with the [Aya Collection]
(https://huggingface.co/datasets/CohereForAI/aya_collection). Serves 23 languages.
| 8k                 | 8k                      | [Chat](/reference/chat)
|
| `c4ai-aya-23-8b`                   | The 8B version of the [Aya 23 model]
(https://huggingface.co/CohereForAI/aya-23-8B). Pairs a highly performant pre-trained
Command family of models with the [Aya Collection]
(https://huggingface.co/datasets/CohereForAI/aya_collection). Serves 23 languages.
| 8k                  | 8k                      | [Chat](/reference/chat)
|
```

Using Command Models on Different Platforms

In this table, we provide some important context for using Cohere Command models on Amazon Bedrock, Amazon SageMaker, and more.

```
| Model Name              | Amazon Bedrock Model ID         | Amazon SageMaker      | Azure
AI Studio Model ID | Oracle OCI Generative AI Service |
| :---------------------- | :------------------------------ | :-------------------- | :-----
------------------ | :------------------------------ |
| `command-r-plus`        | `cohere.command-r-plus-v1:0`    | Unique per deployment | Unique
per deployment    | `cohere.command-r-plus v1.2`     |
| `command-r`             | `cohere.command-r-v1:0`         | Unique per deployment | Unique
per deployment    | `cohere.command-r-16k v1.2`      |
| `command`               | `cohere.command-text-v14`       | N/A                   | N/A
| `cohere.command v15.6`           |
| `command-nightly`       | N/A                             | N/A                   | N/A
| N/A                              |
| `command-light`         | `cohere.command-light-text-v14` | N/A                   | N/A
| `cohere.command-light v15.6`     |
| `command-light-nightly` | N/A                             | N/A                   | N/A
| N/A                              |
|                         |                                 |                       |
|                                  |
```

Embed

These models can be used to generate embeddings from text or classify it based on various parameters. Embeddings can be used for estimating semantic similarity between two sentences, choosing a sentence which is most likely to follow another sentence, or categorizing user feedback, while outputs from the Classify endpoint can be used for any classification or analysis task. The Representation model comes with a variety of helper functions, such as for detecting the language of an input.

```
| Model Name                    | Description
| Dimensions | Context Length | Similarity Metric   | Endpoints
|
|-------------------------------|-------------------------------------------------------------
-------------------------------------------------------------------------------------------
---------------------|------------|----------------|--------------------|----------------
-------------------------------------------------------------------------------------------
|
| `embed-english-v3.0`          | A model that allows for text to be classified or turned
into embeddings. English only.
| 1024       | 512            | Cosine Similarity   | [Embed](/reference/embed),  <br/>
[Embed Jobs](/reference/embed-jobs)               |
| `embed-english-light-v3.0`    | A smaller, faster version of `embed-english-v3.0`. Almost
as capable, but a lot faster. English only.
| 384        | 512            | Cosine Similarity   | [Embed](/reference/embed),  <br/>
[Embed Jobs](/reference/embed-jobs)               |
| `embed-multilingual-v3.0`     | Provides multilingual classification and embedding
support. [See supported languages here.](/docs/supported-languages)
| 1024       | 512            | Cosine Similarity   | [Embed](/reference/embed), [Embed
Jobs](/reference/embed-jobs)                      |
```

| `embed-multilingual-light-v3.0` | A smaller, faster version of `embed-multilingual-v3.0`. Almost as capable, but a lot faster. Supports multiple languages.
| 384 | 512 | Cosine Similarity | [Embed](/reference/embed),
[Embed Jobs](/reference/embed-jobs) |
| | | | | |
| `embed-english-v2.0` | Our older embeddings model that allows for text to be classified or turned into embeddings. English only
| 4096 | 512 | Cosine Similarity | [Classify](/reference/classify), [Embed](/reference/embed) |
| `embed-english-light-v2.0` | A smaller, faster version of embed-english-v2.0. Almost as capable, but a lot faster. English only.
| 1024 | 512 | Cosine Similarity | [Classify](/reference/classify), [Embed](/reference/embed) |
| `embed-multilingual-v2.0` | Provides multilingual classification and embedding support. [See supported languages here.](/docs/supported-languages)
| 768 | 256 | Dot Product Similarity | [Classify](/reference/classify), [Embed](/reference/embed) |

In this table we've listed older `v2.0` models alongside the newer `v3.0` models, but we recommend you use the `v3.0` versions.

Using Embed Models on Different Platforms

In this table, we provide some important context for using Cohere Embed models on Amazon Bedrock, Amazon SageMaker, and more.

Model Name	Amazon Bedrock Model ID	Amazon SageMaker	Azure AI Studio Model ID	Oracle OCI Generative AI Service
`embed-english-v3.0`	`cohere.embed-english-v3`	Unique per deployment	Unique per deployment	`cohere.embed-english-v3.0`
`embed-english-light-v3.0`	N/A	N/A	N/A	`cohere.embed-english-light-v3.0`
`embed-multilingual-v3.0`	`cohere.embed-multilingual-v3`	Unique per deployment	Unique per deployment	`cohere.embed-multilingual-v3.0`
`embed-multilingual-light-v3.0`	N/A	N/A	N/A	`cohere.embed-multilingual-light-v3.0`
`embed-english-v2.0`	N/A	N/A	N/A	N/A
`embed-english-light-v2.0`	N/A	N/A	N/A	`cohere.embed-english-light-v2.0`
`embed-multilingual-v2.0`	N/A	N/A	N/A	N/A

Rerank

The Rerank model can improve created models by re-organizing their results based on certain parameters. This can be used to improve search algorithms.

| Model Name | Description
Context Length	Endpoints
`rerank-english-v3.0`	A model that allows for re-ranking English Language documents and semi-structured data (JSON). This model has a context length of 4096 tokens.
4k	[Rerank](/reference/rerank)
`rerank-multilingual-v3.0`	A model for documents and semi-structure data (JSON) that are not in English. Supports the same languages as embed-multilingual-v3.0. This model has a context length of 4096 tokens.	4k	[Rerank](/reference/rerank)

```
|                                |
|                |                                                              |
| `rerank-english-v2.0`      | A model that allows for re-ranking English language
documents.
| 512            | [Rerank](/reference/rerank) |
| `rerank-multilingual-v2.0` | A model for documents that are not in English. Supports the
same languages as `embed-multilingual-v3.0`.
| 512            | [Rerank](/reference/rerank) |

### Using Rerank Models on Different Platforms

In this table, we provide some important context for using Cohere Rerank models on Amazon
Bedrock, SageMaker, and more.

| Model Name                 | Amazon Bedrock Model ID | Amazon SageMaker      | Azure AI
Studio Model ID | Oracle OCI Generative AI Service |
| :------------------------- | :---------------------- | :-------------------- | :----------
------------ | :------------------------------- |
| `rerank-english-v3.0`      | Not yet available       | Unique per deployment | Not yet
available        | N/A                                |
| `rerank-multilingual-v3.0` | Not yet available       | Unique per deployment | Not yet
available        | N/A                                |
| `rerank-english-v2.0`      | N/A                     | N/A                   | N/A
| N/A                                |
| `rerank-multilingual-v2.0` | N/A                     | N/A                   | N/A
| N/A                                |

<br />

<Note>
Rerank accepts full strings rather than tokens, so the token limit works a little
differently. Rerank will automatically chunk documents longer than 510 tokens, and there is
therefore no explicit limit to how long a document can be when using rerank. See our [best
practice guide](/docs/reranking-best-practices) for more info about formatting documents for
the Rerank endpoint.
</Note>
```