Amazon Titan Models

Amazon Titan foundation models (FMs) are a family of FMs pretrained by AWS on large datasets, making them powerful, general-purpose models built to support a variety of use cases. Use them as-is or privately customize them with your own data.

Amazon Titan supports the following models for Amazon Bedrock.

- **Amazon Titan Text**
- **Amazon Titan Text Embeddings V2**
- **Amazon Titan Multimodal Embeddings G1**
- **Amazon Titan Image Generator G1**

Topics

Amazon Titan Text models

Amazon Titan text models include Amazon Titan Text G1 - Premier, Amazon Titan Text G1 - Express and Amazon Titan Text G1 - Lite.

Amazon Titan Text G1 - Premier

Amazon Titan Text G1 - Premier is a large language model for text generation. It is useful for a wide range of tasks including open-ended and context-based question answering, code generation, and summarization. This model is integrated with Amazon Bedrock Knowledge Base and Amazon Bedrock Agents. The model also supports Custom Finetuning in preview.

- **Model ID** – `amazon.titan-text-premier-v1:0`
- **Max tokens** – 32K
- **Languages** – English

- **Supported use cases** – 32k context window, open-ended text generation, brainstorming, summarizations, code generation, table creation, data formatting, paraphrasing, chain of thought, rewrite, extraction, QnA, chat, Knowledge Base support, Agents support, Model Customization (preview).
- **Inference parameters** – Temperature, Top P (defaults: Temperature = 0.7, Top P = 0.9)

AWS AI Service Card - Amazon Titan Text Premier

Amazon Titan Text G1 - Express

Amazon Titan Text G1 - Express is a large language model for text generation. It is useful for a wide range of advanced, general language tasks such as open-ended text generation and conversational chat, as well as support within Retrieval Augmented Generation (RAG). At launch, the model is optimized for English, with multilingual support for more than 30 additional languages available in preview.

- **Model ID** – `amazon.titan-text-express-v1`
- **Max tokens** – 8K
- **Languages** – English (GA), 100 additional languages (Preview)
- **Supported use cases** – Retrieval augmented generation, open-ended text generation, brainstorming, summarizations, code generation, table creation, data formatting, paraphrasing, chain of thought, rewrite, extraction, QnA, and chat.

Amazon Titan Text G1 - Lite

Amazon Titan Text G1 - Lite is a light weight efficient model, ideal for fine-tuning of English-language tasks, including like summarizations and copy writing, where customers want a smaller, more cost-effective model that is also highly customizable.

- **Model ID** – `amazon.titan-text-lite-v1`
- **Max tokens** – 4K
- **Languages** – English
- **Supported use cases** – Open-ended text generation, brainstorming, summarizations, code generation, table creation, data formatting, paraphrasing, chain of thought, rewrite, extraction, QnA, and chat.

Amazon Titan Text Model Customization

For more information on customizing Amazon Titan text models, see the following pages.

- Prepare the datasets
- Amazon Titan text model customization hyperparameters

Amazon Titan Text Prompt Engineering Guidelines

Amazon Titan text models can be used in a wide variety of applications for different use cases. Amazon Titan Text models have prompt engineering guidelines for the following applications including:

- Chatbot
- Text2SQL
- Function Calling
- RAG (Retrieval Augmented Generation)

For more information on Amazon Titan Text prompt engineering guidelines, see Amazon Titan Text Prompt Engineering Guidelines.

For general prompt engineering guidelines, see Prompt Engineering Guidelines.

AWS AI Service Card - Amazon Titan Text

AI Service Cards provide transparency and document the intended use cases and fairness considerations for our AWS AI services. AI Service Cards provide a single place to find information on the intended use cases, responsible AI design choices, best practices, and performance for a set of AI service use cases.

Amazon Titan Text Embeddings models

Amazon Titan Embeddings text models include Amazon Titan Text Embeddings v2 and Titan Text Embeddings G1 model.

Text embeddings represent meaningful vector representations of unstructured text such as documents, paragraphs, and sentences. You input a body of text and the output is a (1 x n) vector. You can use embedding vectors for a wide variety of applications.

The Amazon Titan Text Embedding v2 model (`amazon.titan-embed-text-v2:0`) can intake up to 8,192 tokens and outputs a vector of 1,024 dimensions. The model also works in 100+ different languages. The model is optimized for text retrieval tasks, but can also perform additional tasks, such as semantic similarity and clustering. Amazon Titan Embeddings text v2 also supports long documents, however, for retrieval tasks it is recommended to segment documents into logical segments (such as paragraphs or sections), per our recommendation.

Amazon Titan Embeddings models generate meaningful semantic representation of documents, paragraphs and sentences. Amazon Titan Text Embeddings takes as input a body of text and generates a n-dimensional vector. Amazon Titan Text Embeddings is offered via latency-optimized endpoint invocation [link] for faster search (recommended during the retrieval step) as well as throughput optimized batch jobs [link] for faster indexing.

The Amazon Titan Embedding Text v2 model supports the following languages: English, German, French, Spanish, Japanese, Chinese, Hindi, Arabic, Italian, Portuguese, Swedish, Korean, Hebrew, Czech, Turkish, Tagalog, Russian, Dutch, Polish, Tamil, Marathi, Malayalam, Telugu, Kannada, Vietnamese, Indonesian, Persian, Hungarian, Modern Greek (1453-), Romanian, Danish, Thai, Finnish, Slovak, Ukrainian, Norwegian, Bulgarian, Catalan, Serbian, Croatian, Lithuanian, Slovenian, Estonian, Latin, Bengali, Latvian, Malay (macrolanguage), Bosnian, Albanian, Azerbaijani, Galician, Icelandic, Georgian, Macedonian, Basque, Armenian, Nepali (macrolanguage), Urdu, Kazakh, Mongolian, Belarusian, Uzbek, Khmer, Norwegian Nynorsk, Gujarati, Burmese, Welsh, Esperanto, Sinhala, Tatar, Swahili (macrolanguage), Afrikaans, Irish, Panjabi, Kurdish, Kirghiz, Tajik, Oriya (macrolanguage), Lao, Faroese, Maltese, Somali, Luxembourgish, Amharic, Occitan (post 1500), Javanese, Hausa, Pushto, Sanskrit, Western Frisian, Malagasy, Assamese, Bashkir, Breton, Waray (Philippines), Turkmen, Corsican, Dhivehi, Cebuano, Kinyarwanda, Haitian, Yiddish, Sindhi, Zulu, Scottish Gaelic, Tibetan, Uighur, Maori, Romansh, Xhosa, Sundanese, Yoruba.

> **Note**
>
> Amazon Titan Text Embeddings v2 model and Titan Text Embeddings v1 model do not supports inference parameters such as `maxTokenCount` or `topP`.

Amazon Titan Text Embeddings V2 model

- **Model ID –** `amazon.titan-embed-text-v2:0`
- **Max input text tokens –** 8,192
- **Languages –** English (100+ languages in preview)