

AI Platforms for Modern Development and Deployment

A visual presentation of
Leading AI Studios
and
Inference Services



A paper by Marian Veteanu

Building apps using foundational models

Building apps using foundational models involves leveraging pre-trained models that can be adapted for various tasks such as text generation, summarization, and image recognition.

Step 1. Select a Foundational Model

Choose a foundational model suitable for your application's needs. Popular foundational models include GPT-4, BERT, CLIP, T5, and DALL·E, which you can access on platforms like Hugging Face, Azure AI, or OpenAI Platform.

Step 2. Understand the Use Case and Apply Prompt Engineering

Prompt engineering is the process of carefully crafting inputs (prompts) to the model to obtain desired outputs. This is especially crucial for language models like GPT, where the prompt directly affects the quality and relevance of the generated text.

Step 3. Fine-Tune the Model (Optional)

In some cases, the foundational model will need to be fine-tuned on a domain-specific dataset to achieve better performance for your particular use case. Fine-tuning involves further training the model on labeled data specific to your task, which helps it adjust to the nuances of your application.

Step 4. Use Retrieval-Augmented Generation (RAG)

RAG involves retrieving relevant documents or data from external sources to improve the model's responses. This is especially useful when working with models like GPT-3 that may not have specific knowledge of newer or proprietary information.

You store the additional information in a vector database (such as Pinecone, FAISS, etc.), which you retrieve and provide to the model to augment its response generation, enhancing the relevance and accuracy of the output.

Step 5. Deploy Model

After fine-tuning and integrating RAG, deploy the model to production using platforms like Azure AI, AWS SageMaker, Google Vertex AI or others. You can set up APIs to serve the model and make it available for inference in real-time applications.

Step 6. Integrate with Application

Connect your deployed model to your application via APIs.

```
const response = await fetch(`https://...`, {
  method: 'POST',
  headers: {
    'Content-Type': 'application/json',
    'api-key': apiKey
  },
  body: JSON.stringify({
    messages: [{ role: 'user', content: "Who is Marian Veteanu?" }]
  })
});
```

Step 7. Monitor and Optimize

Use monitoring tools to track model performance (e.g., latency, accuracy) and detect model drift. Continuous monitoring helps identify when retraining or further fine-tuning is necessary.

All major AI cloud platforms offers you APIs to monitor a model to track data quality.

```
const url = `https://management.azure.com/subscriptions/...`;
const response = await fetch(url, {...});

const data = await response.json(),
driftStatus = data.properties.driftMetrics.dataDriftDetected,
driftDetails = data.properties.driftMetrics.driftMetricsDetails;

if (driftStatus) {
  ...
}
```

Step 8. Continuous Improvement (MLOps)

Automate retraining, testing, and deployment processes through MLOps practices. This ensures that your model stays relevant as data evolves. Platforms like Azure Machine Learning, Google Vertex AI, and AWS SageMaker provide MLOps tools for this purpose.

Azure AI Studio

<https://ai.azure.com/>

Azure AI Studio is a cloud platform provided by Microsoft Azure (akin to an IDE) that unifies various AI and ML services into one environment, enabling users to build, train, fine-tune, and deploy AI models.

It offers tools for experimenting with pre-built AI models from Azure Cognitive Services, as well as custom models developed using Azure Machine Learning. It is a great generative AI application builder with support for prompt engineering, RAG, agent building, and low-code or no-code development.

Similar services:

- Amazon SageMaker
- Google Vertex AI



- Home
- Get started
- Model catalog
- Model benchmarks
- Azure OpenAI
- AI Services



Azure AI Studio

Innovate with AI

Develop and deploy custom copilots at scale, in a safe, secure, and responsible way

Explore cutting-edge models

Explore and experiment with hundreds of large AI models to find the right one for your scenario.

[Browse the model catalog](#)

Start using Azure OpenAI

Get access to Azure OpenAI to start using its wide range of prebuilt and curated models.

[Get started with Azure OpenAI](#)

Work in code with the SDK

Get going with our SDKs, and clone samples to start building in code

[View available SDKs and documentation](#)

Jump to top tasks and tools

Get started with Azure AI Studio

Learn how you can start developing generative AI applications with next-gen models and tools, via web app or directly in code with the SDK.

[Get started in Azure AI Studio](#)

Assistants API available on Azure OpenAI service

Develop power agent-like experiences with built-in state and thread management, knowledge retrieval, and tools including code interpreter and function calling.

[Sign in to build your AI Assistant](#)

Experiment with prompts in the playground

See how different foundation models respond to user input by trying out different prompts, adjusting parameters, and even grounding on your own data.

[Sign in to try some prompts in the playground](#)

Deploy large language models (LLMs)

Deploy an LLM or prompt flow and make its API available for use to an application, website, or other production environment.

[Sign in to view deployments](#)

Explore cutting-edge models



Llama-3.1-405B-Instruct

Meet Meta's latest groundbreaking models; Llama 3.2 11B Vision Instruct and 90B Vision Instruct models for image reasoning.

[View models](#)

gpt-4o-mini

Azure OpenAI Service is powered by a diverse set of models with different capabilities, such as GPT-4o mini, GPT-4o, DALL-E 3 and Whisper.

[View models](#)

Phi-3-mini-4k-instruct

New Phi-3 series of models offers the powerful performance and flexibility of small language models for every computing use cases.

[View models](#)

Cohere

Cohere's suite of Enterprise AI models is available now including Command R, Command R+, Embed, and Rerank

[View models](#)

Explore a variety of large and small models to suit your scenario: [Browse the model catalog](#)

Infuse your solutions with AI Services



Speech

Enhance customer experiences through speech to text, text to speech, and speech translation features.

[View all Speech capabilities](#)

Language + Translator

Analyze, summarize and translate using LLM-powered natural language processing capabilities.

[View all Language + Translator capabilities](#)

Vision + Document

Discover information and insights from documents, images and video with OCR and multi-modal AI.

[View all Vision + Document capabilities](#)

Content Safety

Detect harmful, offensive, or inappropriate user-generated or AI-generated content in your app—including text, image, and multi-modal APIs.

[View all Content Safety capabilities](#)

Find the right model to build your custom AI solution

Show filters

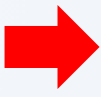
All filters Collections Industry Deployment options Inference tasks Fine-tuning tasks

Licenses

Search

Models 1794

- Home
- Get started
- Model catalog**
- Model benchmarks
- Prompt catalog
- Azure OpenAI
- AI Services
- Management
 - All resources
 - Model quota
 - VM quota



gpt-4o-realtime-preview Audio generation	openai-whisper-large-v3 Speech recognition	openai-whisper-large Speech recognition
gpt-4 Chat completion	gpt-35-turbo Chat completion	o1-preview Chat completion
o1-mini Chat completion	gpt-4o-mini Chat completion	gpt-4o Chat completion
gpt-4-32k Chat completion	gpt-35-turbo-instruct Chat completion	gpt-35-turbo-16k Chat completion
dall-e-3 Text to image	dall-e-2 Text to image	whisper Speech recognition
tts-hd Text to speech	tts Text to speech	text-embedding-3-small Embeddings
text-embedding-3-large Embeddings	Phi-3.5-mini-instruct Chat completion	Phi-3.5-MoE-instruct Chat completion
Phi-3-mini-4k-instruct Chat completion	Phi-3-medium-4k-instruct Chat completion	Phi-3-mini-128k-instruct Chat completion
Phi-3-medium-128k-instruct Chat completion	Phi-3-small-8k-instruct Chat completion	Phi-3-small-128k-instruct Chat completion
Phi-3.5-vision-instruct Chat completion	Phi-3-vision-128k-instruct Chat completion	Meta-Llama-3.1-8B-Instruct Chat completion
Meta-Llama-3.1-8B Text generation	Meta-Llama-3.1-70B-Instr... Chat completion	Meta-Llama-3.1-70B Text generation
Meta-Llama-3-8B-Instruct Chat completion	Meta-Llama-3-8B Text generation	Meta-Llama-3-70B-Instruct Chat completion
Meta-Llama-3-70B Text generation	Llama-3.2-1B Text generation	Llama-3.2-3B Text generation
Llama-Guard-3-8B Chat completion	Llama-3.2-3B-Instruct Chat completion	Llama-3.2-1B-Instruct Chat completion



- Azure AI Studio / Prompt catalog
- Home
- Get started
- Model catalog
- Model benchmarks
- Prompt catalog**
- Azure OpenAI
- AI Services
- Management
- All resources
- Model quota
- VM quota

Prompt catalog



Browse prompt samples for common use cases

Choose a sample prompt to see how it works or as a starting point for your project. Then customize it for your scenario and evaluate how it performs before integrating into your app.



Prompt Samples

[View filters](#)

Prompts

Applied filters

- Generate User Questions On A P...**
To submit your application to evaluation, you ca...
Chat completions
- Travel Assistant**
Provide the travel information.
Summarization
- Social Media Post Analysis**
Analyze short videos to generate tags, content s...
Summarization
- Property Listing Video**
Provide a summary and highlights of the property
Summarization
- Insurance Report**
Car Insurance Damage report writing
Summarization
- Advertising Summary**
Summarize an advertisement and identify issues...
Summarization
- Real Estate Agents Assistant**
Provide overview descriptions for houses.
Summarization
- Listing Assistant**
Generate enticing vacation rental listings from i...
Summarization
- Image Tagging Assistant**
Identify and list prevalent tags associated with t...
Summarization
- Image Description Assistant**
Image Content Description Prompt
Summarization
- Defect Detector**
Find defects on a test image based a reference i...
Summarization
- Apple Cycle Analyst**
Examine the test image to determine the specifi...
Summarization
- Summarize Issue Resolution Fro...**
Summarize issue resolution from conversation
Summarization
- Shakespearean Writing Assistant**
Shakespearean Writing Assistant Prompt
Chat completions
- Product Description And Conten...**
Content creation
- Personalized Marketing**
Content creation
- Parse Unstructured Data**
Parse unstructured data
Reasoning & insights
- Parent Teacher Conference Invit...**
Content creation
- Natural Language To Sql**
Natural language to SQL
Natural language to code
- Natural Language To Python**
Natural language to Python
Natural language to code
- Marketing Writing Assistant**
Marketing Writing Assistant Prompt
Chat completions
- Lesson Plan By Grade Level And ...**
Recommendation
- Json Formatter Assistant**
JSON Formatter Assistant
Chat completions
- Hiking Recommendations Chatbot**
Hiking Recommendations Chatbot
Chat completions
- Generate Product Description**
Generate product description
Content creation
- Generate Blog**
Generate blog
Content creation
- Generate A Job Description**
Generate a job description
Content creation

- AI Services
- Speech
- Vision + Document**
- Language + Translator
- Content Safety



Vision + Document



Give your apps the ability to read text, analyze images, process documents and detect faces with technology like optical character recognition (OCR) and machine learning.

Integrate vision with generative AI

Document field extraction Preview

Extract fields from documents and forms using a custom generative extraction model.

[Try demo](#)

View all other vision capabilities

- Document**
- Face
- Image



Prebuilt models for specific documents

Invoices

Extract invoice ID, customer details, vendor details, ship to, bill to, total tax, subtotal, line items and more.

[Try demo](#)

Receipts

Extract time and date of the transaction, merchant information, amounts of taxes, totals and more.

[Try demo](#)

Identity documents

Extract expiration name, dates, machine readable zone, and more from passports and US driver licenses.

[Try demo](#)

Health insurance cards

Extract information such as, medical network, member name, and deductible.

[Try demo](#)

US Tax forms

Extract key information, etc. from US Tax forms. Supported US Tax forms are W-2, 1040, 1098, and 1099.

[Try demo](#)

US Mortgage forms

Extract key information from US Mortgage forms. Supported US Mortgage forms are 1003, 1008, and closing disclosure.

[Try demo](#)

Marriage certificates

Extract marriage certificate information, such as, couple's names, date of birth/marriage, occupations, addresses, nationality, parents, and officiant.

[Try demo](#)

Credit cards

Extract information such as, card number, cardholder name, due date, and bank information.

[Try demo](#)

Contracts

Extract title, signatory parties' information (including names, reference names, and address), and more from a contract.

[Try demo](#)

General document analysis models

Read

Extract printed and handwritten text from images and documents.

[Try demo](#)

Layout

Extract tables, check boxes, and text from forms and documents.

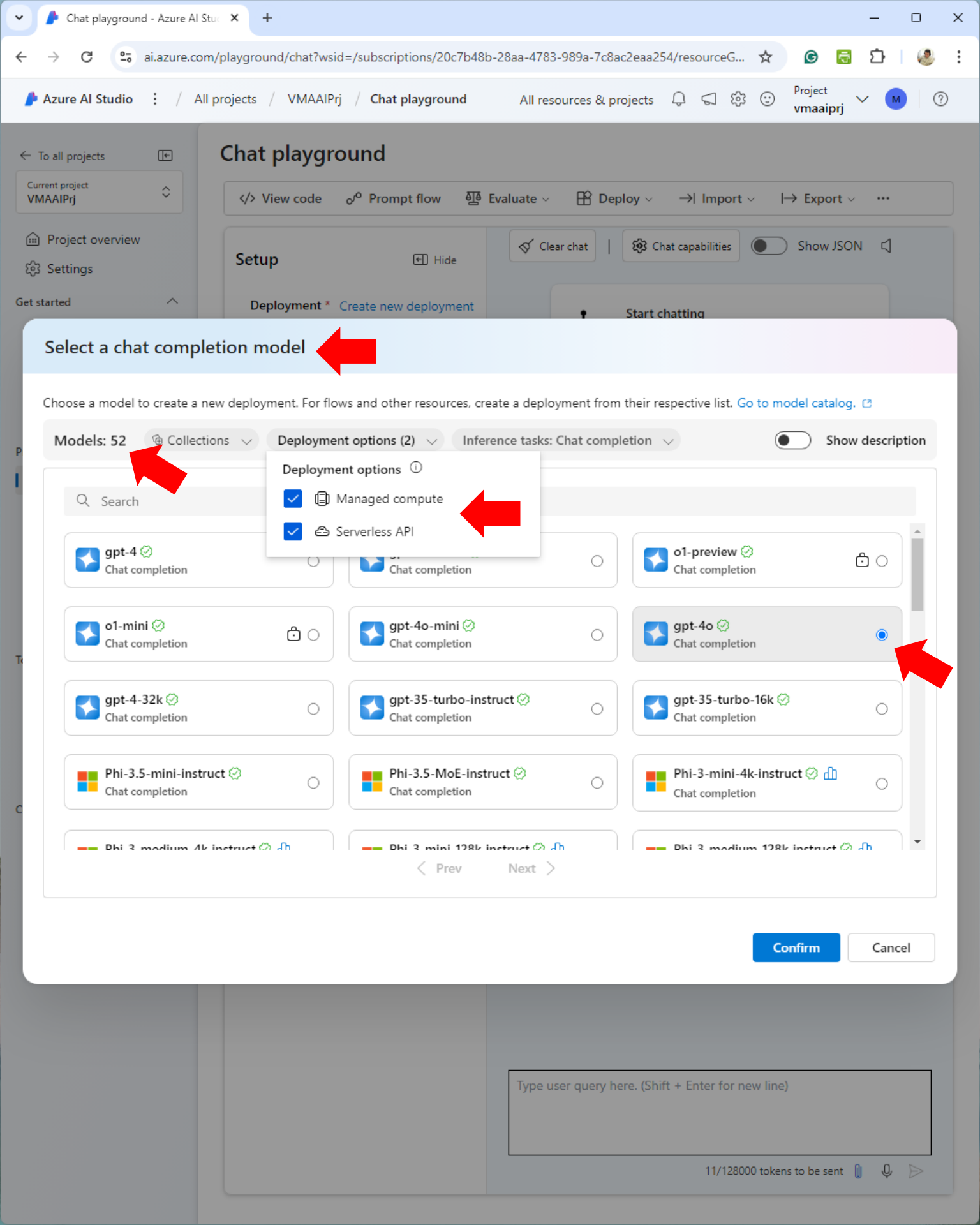
[Try demo](#)

Build document field extraction models from your own data

Document field extraction Preview

Document field extraction - neural and

Document classification model



Select a chat completion model

Choose a model to create a new deployment. For flows and other resources, create a deployment from their respective list. [Go to model catalog.](#)

Models: 52

Collections

Deployment options (2)

Inference tasks: Chat completion

Show description

Search

Deployment options

- Managed compute
- Serverless API

gpt-4 Chat completion

o1-mini Chat completion

gpt-4-32k Chat completion

Phi-3.5-mini-instruct Chat completion

Phi-3-medium-4k-instruct Chat completion

gpt-4o Chat completion

gpt-4o-mini Chat completion

gpt-35-turbo-instruct Chat completion

Phi-3.5-MoE-instruct Chat completion

Phi-3-mini-128k-instruct Chat completion

o1-preview Chat completion

gpt-4o Chat completion

gpt-35-turbo-16k Chat completion

Phi-3-mini-4k-instruct Chat completion

Phi-3-medium-128k-instruct Chat completion

Prev Next

Confirm

Cancel

Type user query here. (Shift + Enter for new line)

11/128000 tokens to be sent

- To all projects
- Current project VMAAIPrj
- Project overview
- Settings
- Get started
 - Model catalog
 - Model benchmarks
 - Prompt catalog
 - AI Services
- Project playground
 - Chat**
 - Assistants PREVIEW
 - Real-time audio PREVIEW
 - Images
 - Completions
 - Speech PREVIEW
- Tools
 - Code PREVIEW
 - Prompt flow
 - Evaluation PREVIEW
 - Fine-tuning PREVIEW
- Components
 - Data
 - Indexes
 - Deployments
 - Content filters

Chat playground

View code Prompt flow Evaluate Deploy Import Export

Setup

Deployment * [Create new deployment](#)
gpt-4o-mini (version:2024-07-...)

System message

Add your data PREVIEW

Parameters

Give the model instructions and context i ↺

You are an AI assistant that helps people find information.

Save

+ Add section

Clear chat Chat capabilities Show JSON

Start chatting

The chat playground can now see, hear, and speak. Select the microphone in the chat window and start speaking to prompt the model without manually entering text. You can also hear the model's output by selecting the speaker icon.

Type user query here. (Shift + Enter for new line)

Google Vertex AI Studio

<https://console.cloud.google.com/vertex-ai/studio>

Google Vertex AI Studio is a cloud-based development environment within Google Cloud for building and using generative AI. You can select from 150+ foundation models from Google's "Model Garden".

It provides support for RAG and model tuning, and other compelling features.

Similar services:

- AWS SageMaker
- Azure AI Studio



- Google Cloud
- CodeGuppy
- Search (/) for resources, docs, products, and more
- Search
- Vertex AI
- TOOLS
 - Dashboard
 - Model Garden
 - Pipelines
- NOTEBOOKS
 - Colab Enterprise
 - Workbench
- VERTEX AI STUDIO
 - Overview
 - Freeform
 - Chat
 - Vision
 - Translation
 - Speech
 - Prompt gallery
 - Prompt management
 - Tuning
- BUILD WITH GEN AI
 - Extensions
- DATA
 - Feature Store
 - Datasets
 - Labeling tasks
- MODEL DEVELOPMENT
 - Training
 - Experiments
 - Metadata
- DEPLOY AND USE
 - Model Registry
 - Online prediction
 - Batch predictions
 - Monitoring
 - Vertex Search



Vertex AI Studio

Test, tune and deploy enterprise-ready generative AI

- TRY A TUTORIAL
- DOCUMENTATION
- API REFERENCE

Generate with Gemini

[OPEN FREEFORM](#) [OPEN CHAT](#)

- Translate text →
- Generate images →
- Convert or synthesize speech →

Prompt management

Saved prompts and tools to make them better.

Tuning

Customize a model to your task

Evaluation

Understand how a model performs on your data

Prompt gallery [SEE ALL](#)

Generate code from comments

Generate Java code from natural-language comments

Summarize Video

Summarize video, extract key information, and organize them in structured output.

French text sample

French text sample

JavaScript physics simulation

Modifying and explaining a JavaScript marble simulation.

Skin care questions

Use only the provided sources to answer questions without citations.

Logo Detection

Extract general Logos' appearance in the video with timestamps.

Python algorithm

Generate codes for algorithm using context clues.

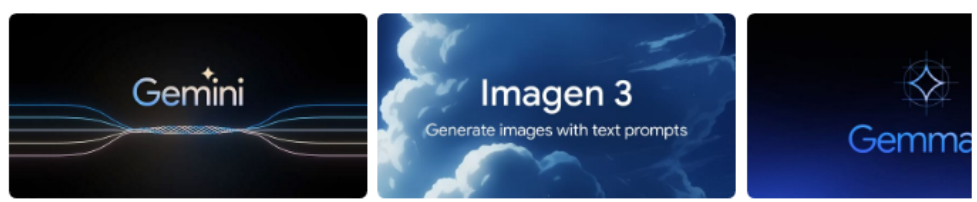
Company chatbot

Create a chatbot for customers with basic company information.

Modalities	
Language	64
Vision	87
Tabular	7
Document	6
Speech	2
Video	6
Tasks	
Generation	72
Classification	64
Detection	43
Extraction	27
Recognition	25
Translation	21
Embedding	7
Segmentation	10
Retrieval	2
Open vocabulary detection	2
Open vocabulary segmentation	2
Tracking	1
Forecasting	5
Automatic speech recognition	1
Providers	
Google	88
Salesforce	3
Meta	18
Stability.ai	4
Mistral AI	5
Anthropic	4

Search models

Browse, customize, and deploy machine learning models with Model Garden. Choose from models created by Google and other providers.



Sort by: [Trending](#) [Newest](#) [Last Update](#)

Foundation models → SHOW ALL (91)

Pre-trained multi-task models that can be further tuned or customized for specific tasks.

Gemini 1.5 Pro

Created from the ground up to be multimodal (text, images, videos) and to scale across a wide range of tasks

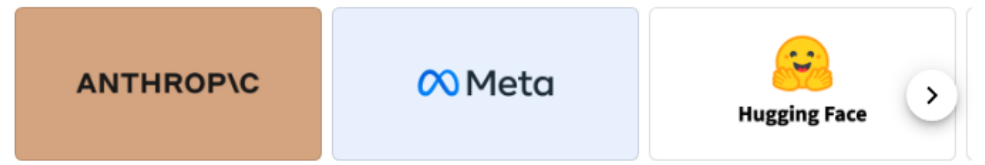
Gemini 1.5 Flash

The best performing Gemini model with features for a wide range of tasks

Gemini 1.0 Pro

Designed to balance quality, performance, and cost for tasks such as content generation, editing, summarization, and classification

Featured partners



Open models on Hugging Face → SHOW MORE

Deploy some of the most popular open source models from Hugging Face to Vertex AI.













google/gemma-2-9b-it google/gemma-2-9b-it	google/gemma-2-27b-it google/gemma-2-27b-it	black-llama-3.1-70b-instruct black-llama-3.1-70b-instruct
meta-llama/Llama-3.1-8B-Instruct meta-llama/meta-llama-3.1-8b-instruct	Qwen/Qwen2.5-72B-Instruct qwen/qwen2.5-72b-instruct	black-llama-3.1-70b-instruct black-llama-3.1-70b-instruct
ibm-granite/granite-3.0-8b-instruct ibm-granite/granite-3.0-8b-instruct	Qwen/Qwen2.5-7B-Instruct qwen/qwen2.5-7b-instruct	ibm-granite/granite-3.0-8b-instruct ibm-granite/granite-3.0-8b-instruct
google/gemma-2-27b-it google/gemma-2-27b-it	meta-llama/Llama-3.1-70B-Instruct meta-llama/Llama-3.1-70B-Instruct	google/gemma-2-27b-it google/gemma-2-27b-it

Open models on Hugging Face



[→ SHOW MORE](#)

Deploy some of the most popular open source models from Hugging Face to Vertex AI.





 google/gemma-2-9b-it google/gemma-2-9b-it	 google/gemma-2-27b-it google/gemma-2-27b-it	 black- black-fo
 meta-llama/Llama-3.1-8B-Instruct meta-llama/meta-llama-3.1-8b-instruct	 Qwen/Qwen2.5-72B-Instruct qwen/qwen2.5-72b-instruct	 black- black-fo
 ibm-granite/granite-3.0-8b-instruct ibm-granite/granite-3.0-8b-instruct	 Qwen/Qwen2.5-7B-Instruct qwen/qwen2.5-7b-instruct	 ibm-gr ibm-gra
 google/gemma-2-2b-jpn-it google/gemma-2-2b-jpn-it	 meta-llama/Llama-3.1-70B-Instruct meta-llama/meta-llama-3.1-70b-instruct	 google google/

Fine-tunable models



[→ SHOW ALL \(57\)](#)

Models that data scientists can further fine-tune through a custom notebook or pipeline.





 Gemini 1.0 Pro Designed to balance quality, performance, and cost for tasks such as content generation, editing, summarization, and classification	 Gemma 2 Lightweight, state-of-the-art open models from Google, built from the same research and technology used to create the Gemini models	 Llama 3.1 Explore and build with Llama models on Vertex AI. 
--	--	--

Task-specific solutions



[→ SHOW ALL \(37\)](#)

Most of these pre-built models are ready to use off the shelf, and many can be customized using your own data.

 Video Speech Transcription Useful for transcribing the speech in video.	 Video Text Detection Useful for detecting visible text in video.	 BiomedCLIP Zero-shot image classification the BiomedCLIP biomedical vision- 
---	---	--

- Vertex AI
- TOOLS
 - Dashboard
 - Model Garden
 - Pipelines
- NOTEBOOKS
 - Colab Enterprise
 - Workbench
- VERTEX AI STUDIO
 - Overview
 - Freeform
 - Chat
 - Vision
 - Translation
 - Speech
 - Prompt gallery**
 - Prompt management
 - Tuning
- BUILD WITH GEN AI
 - Extensions
- DATA
 - Feature Store
 - Datasets
 - Labeling tasks
- MODEL DEVELOPMENT
 - Training
 - Experiments
 - Metadata
- DEPLOY AND USE
 - Model Registry
 - Online prediction
 - Batch predictions
 - Monitoring
 - Vector Search
- MANAGE
 - Ray on Vertex AI

Prompt gallery



Browse prompts across media types and models to help you get started.

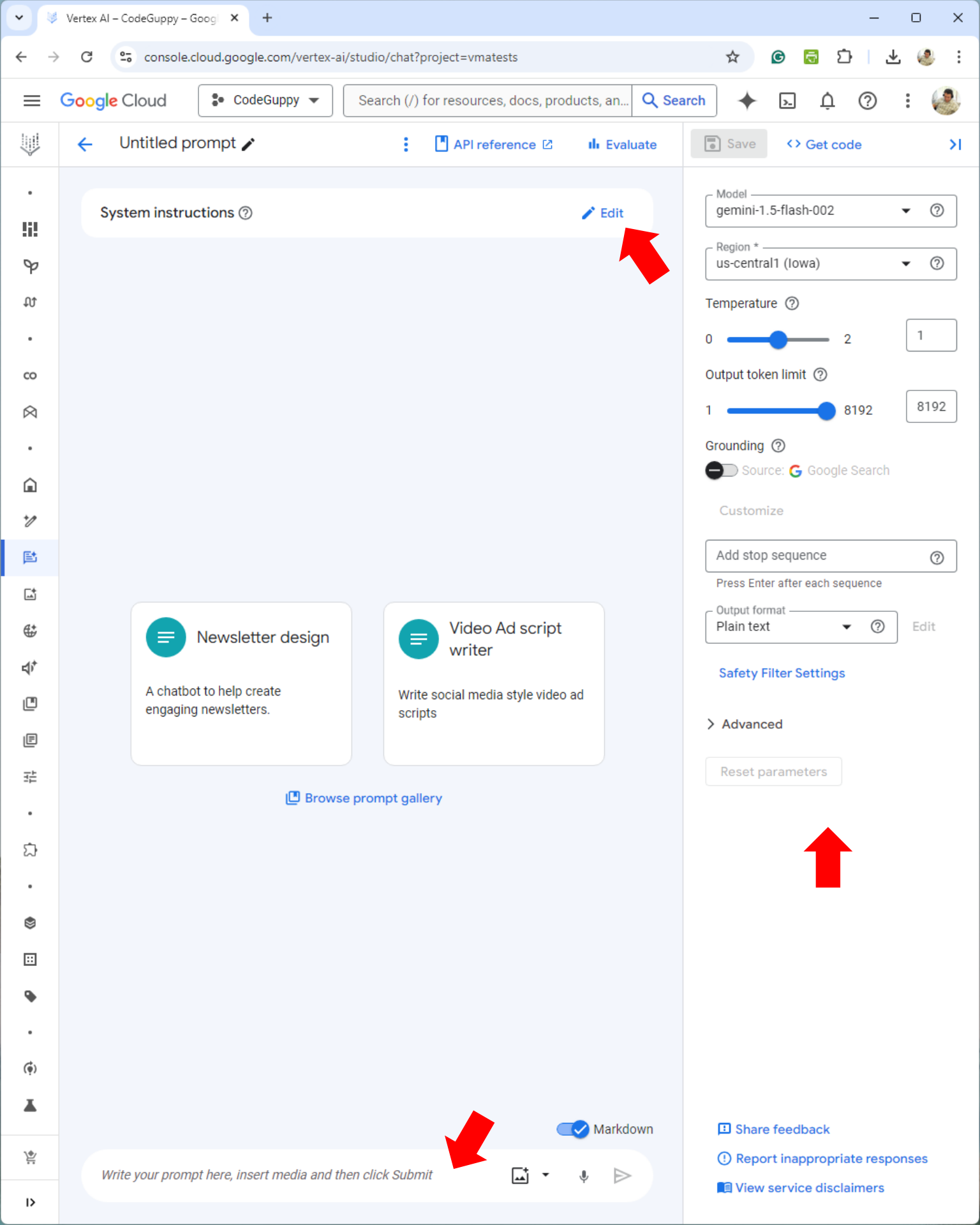
Search sample prompts

Tasks Features Prompt types

Audio Document Image Text Video

CLEAR ALL FILTERS

Ad copy from video Write a creative ad copy based on a video.	Advertising Campaign The AI is tasked to create advertising campaigns for its clients.	Airline reviews The prompt asks the model to write a summary based on customer reviews of an airline company called GoWhereYouLike.
Animal Information Chatbot The animal assistant chatbot answers questions about animals.	Audio diarization Segment an audio record by speaker labels.	Audio Summarization Summarize an audio file
Audio summary on clean energy Summarize a piece of audio recording.	Audio transcription Generate the transcription for a piece of audio recording.	Audio/video Q&A Audio/video Q&A
Beach vacation The prompt asks the model to write a summary based on customer reviews of a beach in California.	Blog post creator Create a blog post	Book Publishing and Editing Take a verbose, subjective excerpt and distill it into a concise, objective list of facts
Business Development Writing Extract relevant information from the user input that can be used in business development initiatives.	Chatbot recommendations for courses A chatbot suggests courses for a performing arts program.	Classify headlines Label news headlines with custom topics using examples.
Code optimization explanation Optimize and explain C++ code, focusing on time complexity.	Company chatbot Create a chatbot for customers with basic company information.	Company Financial Analysis Company Financial Analysis
Completing Go functions	Culinary Dish Classification	Customer Service Assistance



System instructions [Edit](#)

Newsletter design
A chatbot to help create engaging newsletters.

Video Ad script writer
Write social media style video ad scripts

[Browse prompt gallery](#)

Model: gemini-1.5-flash-002

Region: us-central1 (Iowa)

Temperature: 1

Output token limit: 8192

Grounding: Source: Google Search

Customize

Add stop sequence: Press Enter after each sequence

Output format: Plain text

[Safety Filter Settings](#)

Advanced

Reset parameters

- [Share feedback](#)
- [Report inappropriate responses](#)
- [View service disclaimers](#)

Write your prompt here, insert media and then click Submit Markdown

AWS SageMaker Studio

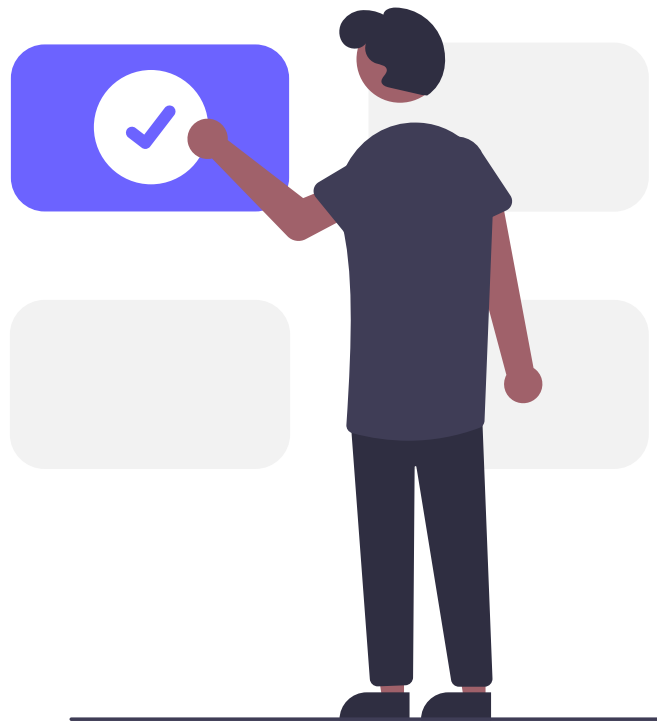
<https://aws.amazon.com/sagemaker/studio/>

AWS SageMaker Studio is a fully integrated development environment (IDE) within AWS SageMaker that streamlines the entire machine learning lifecycle, from data preparation to model training, deployment, and monitoring.

SageMaker Studio supports Retrieval-Augmented Generation (RAG), prompt engineering and it offers access to a catalog of pre-built models. This makes it a powerful platform for building, experimenting, and scaling machine learning solutions with robust support for model management, versioning, and monitoring.

Similar services:

- Azure AI Studio
- Google Vertex AI Studio



- Amazon SageMaker
- Getting started
- Applications and IDEs
- Admin configurations
- SageMaker dashboard
- Search
- JumpStart
 - Foundation models
 - Computer vision models
 - Natural language processing models
- Governance
- HyperPod Clusters
- Ground Truth
- Processing
- Training
- Inference
- Augmented AI
- AWS Marketplace
- Tutorials
- Documentation

Foundation models



Foundation models are pre-trained on large amounts of data so you can perform a wide range of tasks such as article summarization and text, image, or video generation. This page has a select number of foundation models available from JumpStart. Please visit JumpStart in Studio to view all available models. ([documentation](#))

Foundation models

Stable Diffusion XL 1.0
By Stability AI | Ver 20230726

PROFESSIONAL: COMPARED TO PREVIOUS VERSIONS, SDXL 1.0...

The official foundation model for image generation from Stability. Deploy this optimized instance and serve generative AI within minutes.

[View model](#)

Meta Llama 2 7B Chat
By Meta | Ver 1.0.0

CHAT OPTIMIZED, TEXT GENERATION, LLAMA 2

7B dialogue use case optimized variant of Llama 2 models. Llama 2 is an auto-regressive language model that uses an optimized transformer architecture. Llam...

[View model](#)

Meta Llama 2 70B Chat
By Meta | Ver 1.0.0

CHAT OPTIMIZED, TEXT GENERATION, LLAMA 2

70B dialogue use case optimized variant of Llama 2 models. Llama 2 is an auto-regressive language model that uses an optimized transformer architecture. Llam...

[View model](#)

AI21 Jurassic-2 Ultra
By AI21 Labs | Ver 2.2.004

RECOGNIZED AMONG STANFORD'S TOP-TIER LLM EVALUATIONS, JURASSIC-2...

Top-of-the-line LLM for complex text generation tasks that require the highest quality output.

[View model](#)

Cohere Generate Model - Command
By Cohere | Ver v1.1.6

TEXT GENERATION, GENERATIVE AI, CONTENT GENERATION, AI TEXT WRITE...

Powered by a large language model use Cohere Generate for tasks like copywriting, named entity recognition, paraphrasing or

LightOn Mini-instruct 40B
By LightOn | Ver v1.0

TEXT GENERATION, KEYWORD EXTRACTION, INFORMATION...

40B multilingual and instruction-tuned language model.



- Amazon SageMaker
- Getting started
- Applications and IDEs
- Admin configurations
- SageMaker dashboard
- Search
- JumpStart
 - Foundation models
 - Computer vision models
 - Natural language processing models
- Governance
 - Model dashboard **NEW**
 - Model cards **NEW**
- HyperPod Clusters
- Ground Truth
- Processing
- Training
- Inference
- Augmented AI
- AWS Marketplace
 - Model packages
 - Algorithms
 - AWS Data Exchange
 - All products
- Tutorials
- Documentation

Amazon SageMaker > AWS Marketplace

Search AWS Marketplace



Need help creating a custom machine learning solution? [Connect with an AWS IQ expert](#) to train a custom model in Amazon SageMaker.

Featured model packages [See all products](#)

NAVINFO **Face and License Plate Anonymizer**
By NavInfo Europe B.V. | Ver 4.0.0
Free Trial
Detection and blurring of faces and license plates.

[View product](#)

mxnet **GluonCV YOLOv3 Object Detector**
By Amazon Web Services | Ver 1.1
★★★★★ 2
YOLOv3 is a powerful network for fast and accurate object detection, powered by GluonCV.

[View product](#)

VITECHLab **PPE Detector for Worker Safety**
By VITech Lab | Ver 3.0
Free Trial
Image recognition and classification model for real-time PPE compliance monitoring in industrial environments.

[View product](#)

Mphasis **Mphasis DeepInsights Text Summarizer**
By Mphasis | Ver 3.2
Mphasis DeepInsights Text Summarizer helps in summarizing text documents.

[View product](#)

SageMaker Studio > Jumpstart > SageMakerPublicHub

Applications (6)

- JupyterLab
- RStudio
- Canvas
- Code Editor
- Studio CL...
- MLflow

Home

Running instances

Data

Auto ML

Experiments

Jobs

Pipelines

Models

JumpStart

Deployments

Collapse Menu

All public models

Discover all popular pre-trained models offered by SageMaker

Providers 20

text

Models 535

- Meta-Llama-3.1-405B by Meta
Text Generation
- Meta-Llama-3.1-405B-Instruct by Meta
Text Generation
- Meta Llama 3.2 11B Vision by Meta
Vision Language
- Meta Llama 3.2 11B Vision Instruct by Meta
Vision Language
- Meta Llama 3.2 90B Vision by Meta
Vision Language
- Meta Llama 3.2 90B Vision Instruct by Meta
Vision Language
- Meta Llama 3.2 1B by Meta
Text Generation
- Meta Llama 3.2 1B Instruct by Meta
Text Generation
- Meta Llama 3.2 3B by Meta



Search in this guide

Contact Us

English

Return to the Console

AWS > Documentation > Amazon SageMaker > Developer Guide

Feedback Preferences

Amazon SageMaker Developer Guide

Recently added to this guide

Deploy a Model in Studio
October 16, 2024

Share model group in Studio
October 16, 2024

View shared model groups in Studio
October 16, 2024

View all

- ▶ What is Amazon SageMaker?
- ▶ Setting up SageMaker
- ▶ Automated ML, no-code, or low-code
- ▶ Machine learning environments offered by Amazon SageMaker
- ▶ Data labeling with a human-in-the-loop
- ▶ Prepare data
- ▶ Processing jobs
- ▶ Create, store, and share features
- ▶ Model training
- ▼ Deploy models for inference
 - Model Deployment
 - ▶ Options for deploying models and getting inferences
 - Model creation with ModelBuilder
 - ▶ Model inference optimization
 - Options for evaluating your model
 - ▶ Inference Recommender
 - ▶ Real-time inference
 - ▼ Serverless Inference
 - ▶ Serverless endpoint operations
 - Alarms and logs
 - ▶ Automatically scale Provisioned Concurrency for a serverless endpoint

Deploy models with Amazon SageMaker Serverless Inference

PDF | RSS



Amazon SageMaker Serverless Inference is a purpose-built inference option that enables you to deploy and scale ML models without configuring or managing any of the underlying infrastructure. On-demand Serverless Inference is ideal for workloads which have idle periods between traffic spurts and can tolerate cold starts. Serverless endpoints automatically launch compute resources and scale them in and out depending on traffic, eliminating the need to choose instance types or manage scaling policies. This takes away the undifferentiated heavy lifting of selecting and managing servers. Serverless Inference integrates with AWS Lambda to offer you high availability, built-in fault tolerance and automatic scaling. With a pay-per-use model, Serverless Inference is a cost-effective option if you have an infrequent or unpredictable traffic pattern. During times when there are no requests, Serverless Inference scales your endpoint down to 0, helping you to minimize your costs. For more information about pricing for on-demand Serverless Inference, see [Amazon SageMaker Pricing](#).

Optionally, you can also use Provisioned Concurrency with Serverless Inference. Serverless Inference with provisioned concurrency is a cost-effective option when you have predictable bursts in your traffic. Provisioned Concurrency allows you to deploy models on serverless endpoints with predictable performance, and high scalability by keeping your endpoints warm. SageMaker ensures that for the number of Provisioned Concurrency that you allocate, the compute resources are initialized and ready to respond within milliseconds. For Serverless Inference with Provisioned Concurrency, you pay for the compute capacity used to process inference requests, billed by the millisecond, and the amount of data processed. You also pay for Provisioned Concurrency usage, based on the memory configured, duration provisioned, and the amount of concurrency enabled. For more information about pricing for Serverless Inference with Provisioned Concurrency, see [Amazon SageMaker Pricing](#).

You can integrate Serverless Inference with your MLOps Pipelines to streamline your ML workflow, and you can use a serverless endpoint to host a model registered with [Model Registry](#).

Serverless Inference is generally available in 21 AWS Regions: US East (N. Virginia), US East (Ohio), US West (N. California), US West (Oregon), Africa (Cape Town), Asia Pacific (Hong Kong), Asia Pacific (Mumbai), Asia Pacific (Tokyo), Asia Pacific (Seoul), Asia Pacific (Osaka), Asia Pacific (Singapore), Asia Pacific (Sydney), Canada (Central), Europe (Frankfurt), Europe (Ireland), Europe (London), Europe (Paris), Europe (Stockholm), Europe (Milan), Middle East (Bahrain), South America (São Paulo). For more information about Amazon



OpenAI Platform

<https://platform.openai.com/>

OpenAI Platform is perhaps the most well-known cloud platform of APIs that allows developers to integrate gen-AI models into their applications. The platform provides access to LLMs like GPT-4, which are used for tasks such as text generation, language translation, code generation, and more.

The platform includes a Playground where users can experiment with various models interactively, tweaking prompts and parameters in real time to see how models respond. Other core features include prompt engineering and RAG.

OpenAI offers various services such as text generation, image creation (e.g., DALL·E), code assistance (e.g., Codex), and more.



Search CTRL K

- GET STARTED
- Overview
- Quickstart
- Models
- Changelog
- Terms and policies
- CAPABILITIES
- Text generation
- Image generation
- Vision
- Audio generation
- Text to speech
- Speech to text
- Vector embeddings
- Moderation
- Reasoning
- GUIDES
- Function calling
- Structured outputs
- Evaluations
- Fine-tuning
- Distillation
- Realtime API
- Batch API
- ASSISTANTS
- Overview
- Quickstart
- Deep dive
- Tools
- What's new?
- Migration guide
- CHATGPT
- Actions
- Release notes
- BEST PRACTICES
- Prompt engineering
- Production best practices

OpenAI developer platform

Developer quickstart

Set up your environment and make your first API request in minutes

5 min

```
node.js
1 import OpenAI from "openai";
2 const openai = new OpenAI();
3 const completion = await openai.chat.completions.create({
4   model: "gpt-4o",
5   messages: [
6     {"role": "user", "content": "write a haiku about ai"}
7   ]
8 });
```

Meet the models

Pricing Explore all

GPT-4o

Our high-intelligence flagship model for complex, multi-step tasks

Text and image input, text output

128k context length

Smarter model, higher price per token

GPT-4o mini

Our affordable and intelligent small model for fast, lightweight tasks

Text and image input, text output

128k context length

Faster model, lower price per token

o1-preview and o1-mini Beta

A new series of reasoning models for solving hard problems

Text input, text output

128k context length

Higher latency, uses tokens to think



Start building

Distillation

Evaluate and fine-tune models using production logs

Realtime

Build low-latency multimodal experiences

Structured Outputs

Ensure model responses adhere to your supplied JSON schema

```
{
  "name": "math_response",
  "strict": true,
  "schema": {
    "type": "object",
    "properties": {
      "steps": {
        "type": "array",
```

Explore our guides

Prompt engineering
Get better results from LLMs

Production best practices
Transition from prototype to production

Safety best practices
Make sure your application is safe

Latency optimization
Improve latency across multiple use cases

Optimizing LLM accuracy


- PLAYGROUND
- Chat
- Realtime
- Assistants
- TTS
- Completions



Chat

Clear History Code Compare

System instructions Generate

You are a helpful assistant...



Your presets

Model: gpt-4o

Functions: + Add

Response format: text

Model configuration

Temperature: 1.00

Maximum length: 2048

Stop sequences: Enter sequence and press Tab

Top P: 1.00

Frequency penalty: 0.00

Presence penalty: 0.00



- Cookbook
- Forum
- Help

Enter user message...

User 

Add Run Ctrl+↵

Save as preset

HuggingFace

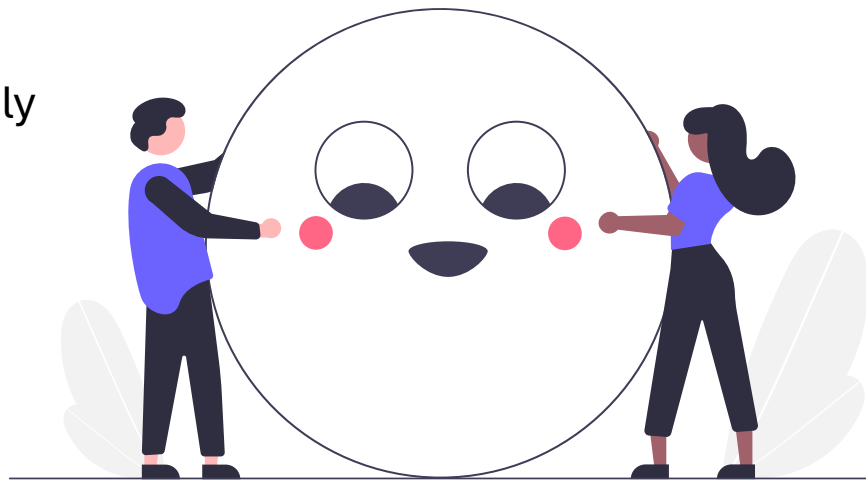
<https://huggingface.co/>

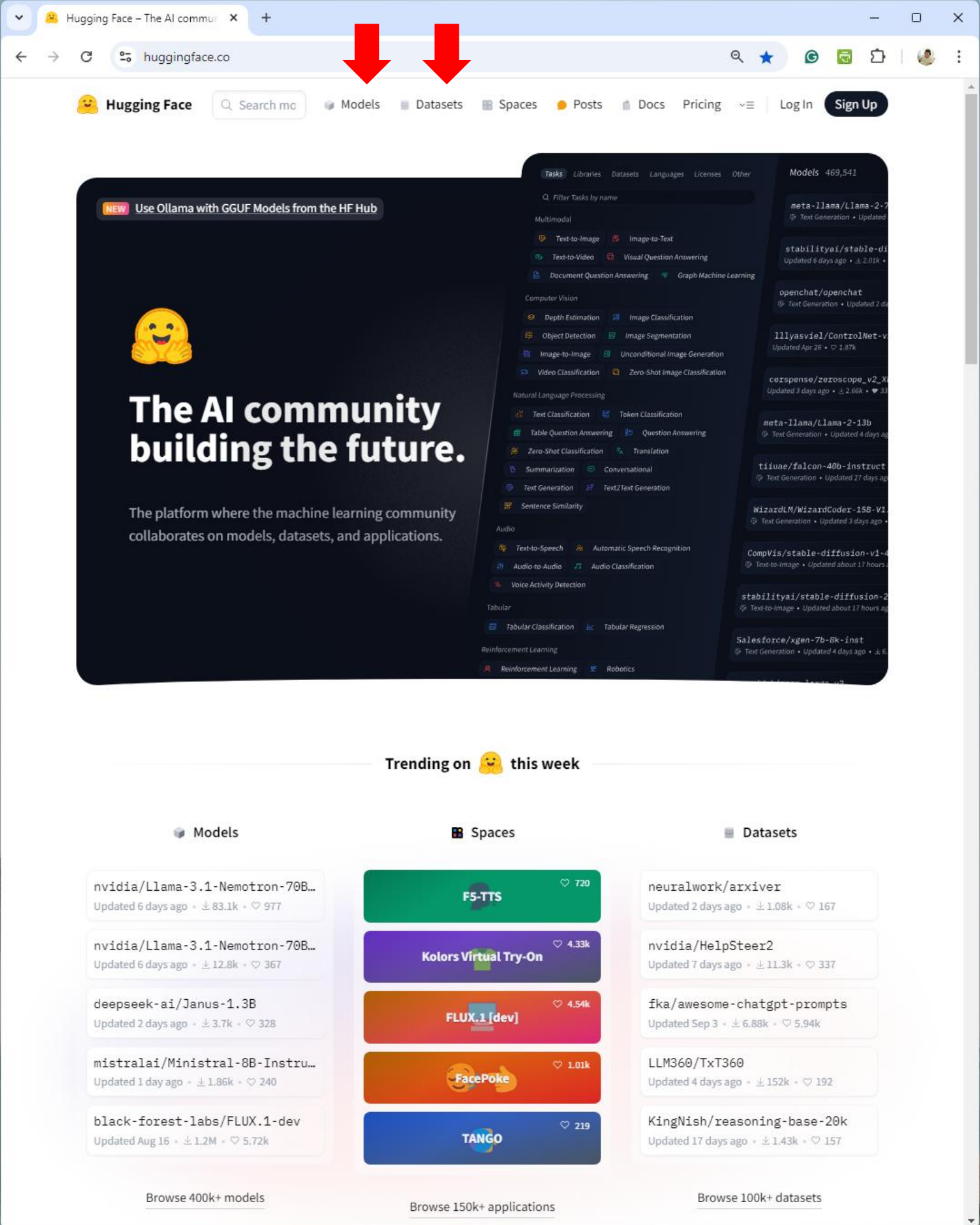
Hugging Face is an open-source platform that has gained widespread popularity for its extensive library of pre-trained models and tools for NLP, computer vision, audio, and multimodal tasks.

The platform's flagship Transformers library enables developers to access thousands of pre-trained models for tasks like text generation, translation, classification, and more, with support for popular deep learning frameworks like PyTorch and TensorFlow.

Hugging Face is widely adopted in the AI and machine learning communities, both in academia and industry, with its tools being used by companies like Microsoft, Google, and Facebook, as well as numerous research institutions.

If you need a model for a specific scenario, probably HuggingFace has it!





The AI community building the future.

The platform where the machine learning community collaborates on models, datasets, and applications.

Trending on 🤗 this week

Models

nvidia/Llama-3.1-Nemotron-70B...
Updated 6 days ago • ⬇️ 83.1k • ❤️ 977

nvidia/Llama-3.1-Nemotron-70B...
Updated 6 days ago • ⬇️ 12.8k • ❤️ 367

deepseek-ai/Janus-1.3B
Updated 2 days ago • ⬇️ 3.7k • ❤️ 328

mistralai/Ministral-8B-Instru...
Updated 1 day ago • ⬇️ 1.86k • ❤️ 240

black-forest-labs/FLUX.1-dev
Updated Aug 16 • ⬇️ 1.2M • ❤️ 5.72k

Spaces

F5-TTS
❤️ 720

Kolors Virtual Try-On
❤️ 4.33k

FLUX.1 [dev]
❤️ 4.54k

FacePoke
❤️ 1.01k

TANGO
❤️ 219

Datasets

neuralwork/arxiv
Updated 2 days ago • ⬇️ 1.08k • ❤️ 167

nvidia/HelpSteer2
Updated 7 days ago • ⬇️ 11.3k • ❤️ 337

fka/awesome-chatgpt-prompts
Updated Sep 3 • ⬇️ 6.88k • ❤️ 5.94k

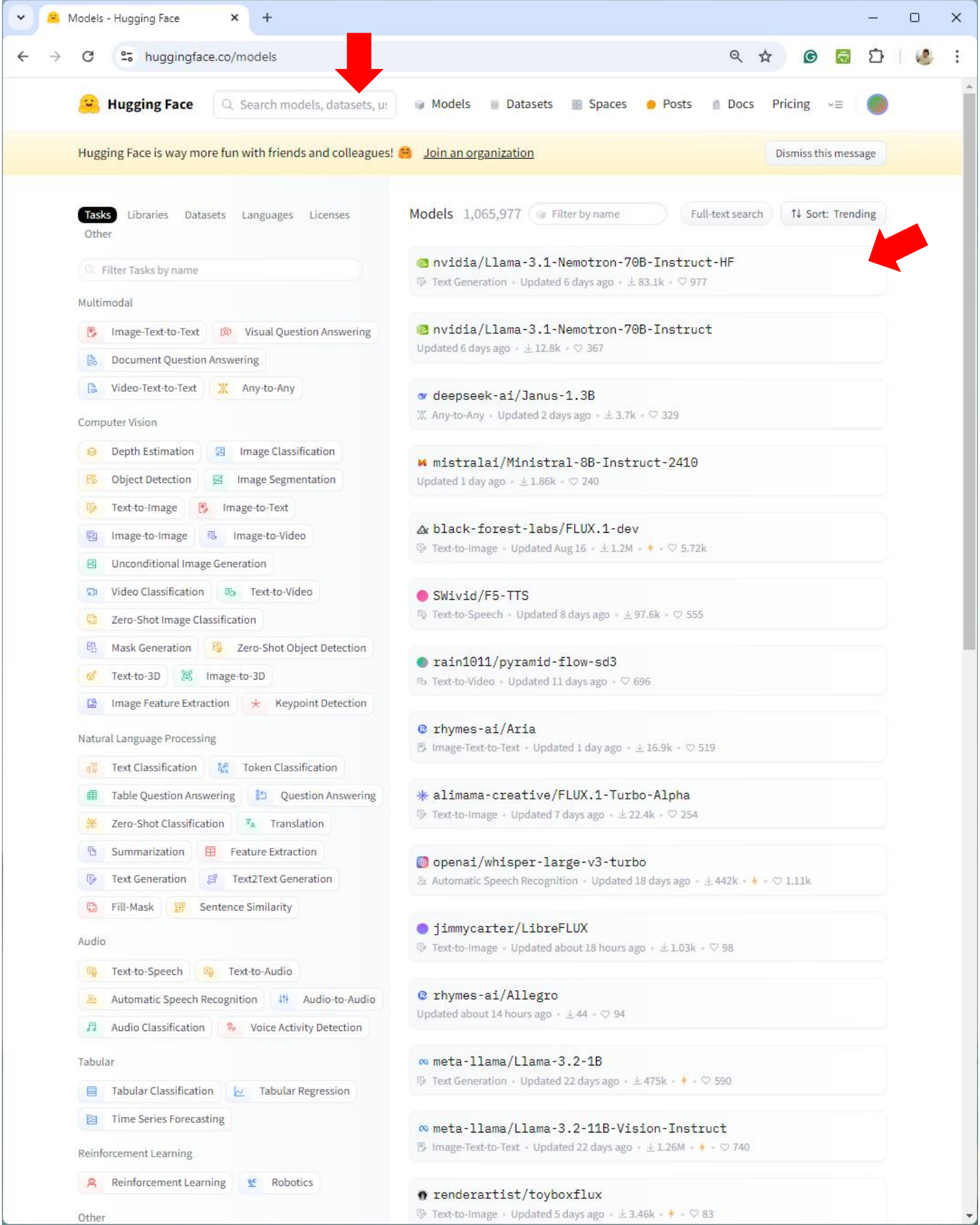
LLM360/TxT360
Updated 4 days ago • ⬇️ 152k • ❤️ 192

KingNish/reasoning-base-20k
Updated 17 days ago • ⬇️ 1.43k • ❤️ 157

[Browse 400+ models](#)

[Browse 150+ applications](#)

[Browse 100k+ datasets](#)



Hugging Face is way more fun with friends and colleagues! Join an organization Dismiss this message

meta-llama / Llama-3.2-1B like 590



Text Generation Transformers Safetensors PyTorch 8 languages llama facebook meta llama-3
text-generation-inference Inference Endpoints arxiv:2204.05149 License: llama3.2

Model card Files Community 26

Train Deploy Use this model



You need to agree to share your contact information to access this model

The information you provide will be collected, stored, processed and shared in accordance with the [Meta Privacy Policy](#).

LLAMA 3.2 COMMUNITY LICENSE AGREEMENT

Llama 3.2 Version Release Date: September 25, 2024

"Agreement" means the terms and conditions for use, reproduction, distribution and modification of the Llama Materials set forth herein.

"Documentation" means the specifications, manuals and documentation accompanying Llama 3.2 distributed by Meta at <https://llama.meta.com/doc/overview...>

Expand to review

Expand to review and access



Safetensors Model size 1.24B params Tensor type BF16

Inference API Warm

Text Generation Examples

My name is Merve and my favorite

Compute ctrl+Enter 0.0

View Code Maximize

Model Information

The Meta Llama 3.2 collection of multilingual large language models (LLMs) is a collection of pretrained and instruction-tuned generative models in 1B and 3B sizes (text in/text out). The Llama 3.2 instruction-tuned text only models are optimized for multilingual dialogue use cases, including agentic retrieval and summarization tasks. They outperform many of the available open source and closed chat models on common industry benchmarks.

Model Developer: Meta

Model Architecture: Llama 3.2 is an auto-regressive language model that uses an optimized transformer architecture. The tuned versions use supervised fine-tuning (SFT) and reinforcement learning with human feedback (RLHF) to align with human preferences for helpfulness and safety.

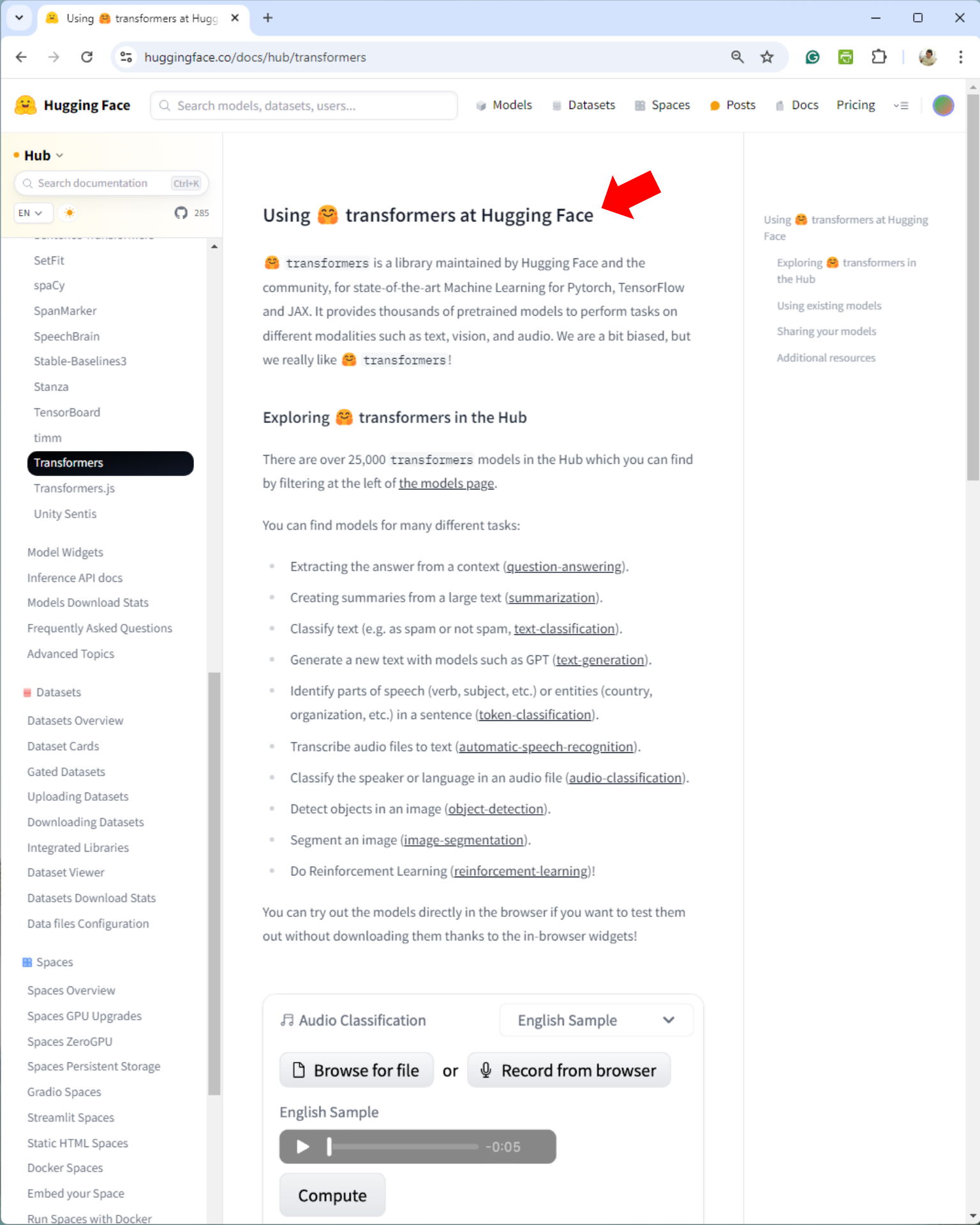
Model tree for meta-llama/Llama-3.2-1B

Adapters	31 models
Finetunes	46 models
Quantizations	35 models

Spaces using meta-llama/Llama-3.2-1B 100

- Vishal3152/meta-llama-Llama-3.2-1B
- Minorutanaka14052005/meta-llama-Llama-3.2-1Bk
- Gihani111/meta-llama-Llama-3.2-1B
- DraxJC/Tecnicas2024
- Meow88/meta-llama-Llama-3.2-1B
- emirkaanozdemr/Trendyol-Hackathon
- solnone/meta-llama-Llama-3.2-1B
- Chilliming/meta-llama-Llama-3.2-1B
- tayyabmalik4/llama_model
- ObindiG/moto
- abhilashn2006/EmailGenie

Training	Params	Input	Output	Context
----------	--------	-------	--------	---------



Hub

- SetFit
- spaCy
- SpanMarker
- SpeechBrain
- Stable-Baselines3
- Stanza
- TensorBoard
- timm
- Transformers**
- Transformers.js
- Unity Sentis
- Model Widgets
- Inference API docs
- Models Download Stats
- Frequently Asked Questions
- Advanced Topics

- Datasets**
- Datasets Overview
- Dataset Cards
- Gated Datasets
- Uploading Datasets
- Downloading Datasets
- Integrated Libraries
- Dataset Viewer
- Datasets Download Stats
- Data files Configuration

- Spaces**
- Spaces Overview
- Spaces GPU Upgrades
- Spaces ZeroGPU
- Spaces Persistent Storage
- Gradio Spaces
- Streamlit Spaces
- Static HTML Spaces
- Docker Spaces
- Embed your Space
- Run Spaces with Docker

Using transformers at Hugging Face

transformers is a library maintained by Hugging Face and the community, for state-of-the-art Machine Learning for Pytorch, TensorFlow and JAX. It provides thousands of pretrained models to perform tasks on different modalities such as text, vision, and audio. We are a bit biased, but we really like transformers!

Exploring transformers in the Hub

There are over 25,000 transformers models in the Hub which you can find by filtering at the left of [the models page](#).

You can find models for many different tasks:

- Extracting the answer from a context ([question-answering](#)).
- Creating summaries from a large text ([summarization](#)).
- Classify text (e.g. as spam or not spam, [text-classification](#)).
- Generate a new text with models such as GPT ([text-generation](#)).
- Identify parts of speech (verb, subject, etc.) or entities (country, organization, etc.) in a sentence ([token-classification](#)).
- Transcribe audio files to text ([automatic-speech-recognition](#)).
- Classify the speaker or language in an audio file ([audio-classification](#)).
- Detect objects in an image ([object-detection](#)).
- Segment an image ([image-segmentation](#)).
- Do Reinforcement Learning ([reinforcement-learning](#))!

You can try out the models directly in the browser if you want to test them out without downloading them thanks to the in-browser widgets!

🔊 Audio Classification English Sample

or

English Sample

- Using transformers at Hugging Face
- Exploring transformers in the Hub
- Using existing models
- Sharing your models
- Additional resources

IBM Watsonx.ai

<https://watsonx.ai/>

If you're a IBM cloud customer, you should check IBM watsonx.ai.

It is IBM's AI platform designed to build, deploy, and scale AI models with a focus on foundation models and generative AI.

Watsonx.ai supports large-scale AI, including NLP, computer vision, and other machine learning tasks, and is particularly geared toward enterprise-grade applications.

In terms of competition, IBM watsonx.ai competes with platforms like Google Cloud Vertex AI, Microsoft Azure AI, and Amazon SageMaker.



- Overview
- IBM models
- Benefits
- Foundation model library**
- Embedding model library
- Client stories
- Model partnerships
- News and resources
- IP Protection for AI Models
- Next steps

Foundation model library

Select a generative foundation model that best fits your needs. After you have a short list of models for your use case, systematically test the models by using prompt engineering techniques to see which ones consistently return the desired results.

[See more watsonx pricing information](#) →

Model name	Provider	Use cases	Context length	Price USD/1 million tokens
granite-3-8b-instruct New Featured model	IBM	Supports questions and answers (Q&A), summarization, classification, generation, extraction, RAG, and coding tasks.	4096	0.20
granite-3-2b-instruct New Featured model	IBM	Supports questions and answers (Q&A), summarization, classification, generation, extraction, RAG, and coding tasks.	4096	0.10
granite-guardian-3-8b New Featured model	IBM	Supports detection of HAP/PII, jailbreaking, bias, violence, and other harmful content.	4096	0.20
granite-guardian-3-2b New Featured model	IBM	Supports detection of HAP/PII, jailbreaking, bias, violence, and other harmful content.	4096	0.10
granite-20b-multilingual	IBM	Supports Q&A, summarization, classification, generation, extraction, translation and RAG tasks in French, German, Portuguese, Spanish and English.	8192	0.60



Replicate

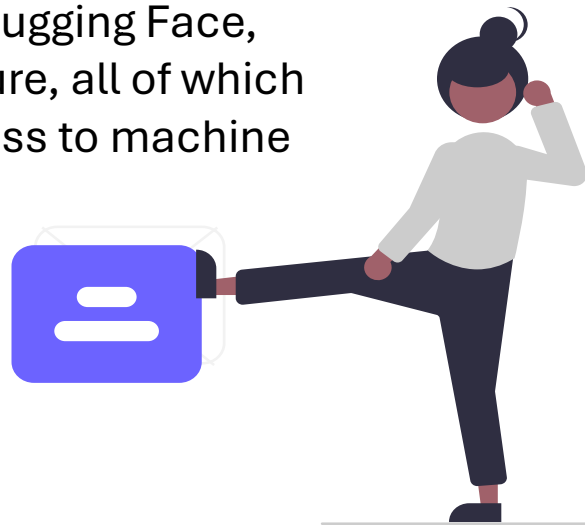
<https://replicate.com/>

If you want a simple 3rd party platform, then replicate.com is a platform that simplifies the process of deploying and running machine learning models through a cloud-based API.

It allows developers to access a variety of open-source pre-trained models for tasks such as image generation, video editing, text understanding, and more.

Replicate.com can be effectively used for AI inference, enabling developers to leverage pre-trained models or deploy their custom machine learning models for real-time inference via a cloud API. Once a model is deployed on Replicate, it can be accessed through API calls that handle input data processing and provide instant results.

Replicate competes with services like Hugging Face, and the ones from AWS, Google and Azure, all of which offer model hosting and API-based access to machine learning capabilities.






Run AI with an API.



Run and fine-tune open-source models. Deploy custom models at scale. All with one line of code.


With Replicate you can ↴

- Generate images
- Generate text**
- Caption images
- Generate music
- Generate speech
- Fine tune models
- Restore images



google-deepmind / gemma-2b-it
2B instruct version of Google's Gemma model

89K runs [View on Replicate](#)



yorickvp / llava-13b
Visual instruction tuning towards large language and vision models with GPT-4...

18M runs [View on Replicate](#)

Write me a poem about Machine Learning. **Run model**

Python JavaScript cURL

```
import replicate

output = replicate.run(
    "google-deepmind/gemma-2b-it:dff94eaf770e1fc211e425"
    input={
        "prompt": "Write me a poem about Machine Learning"
    }
)

print(output)
```

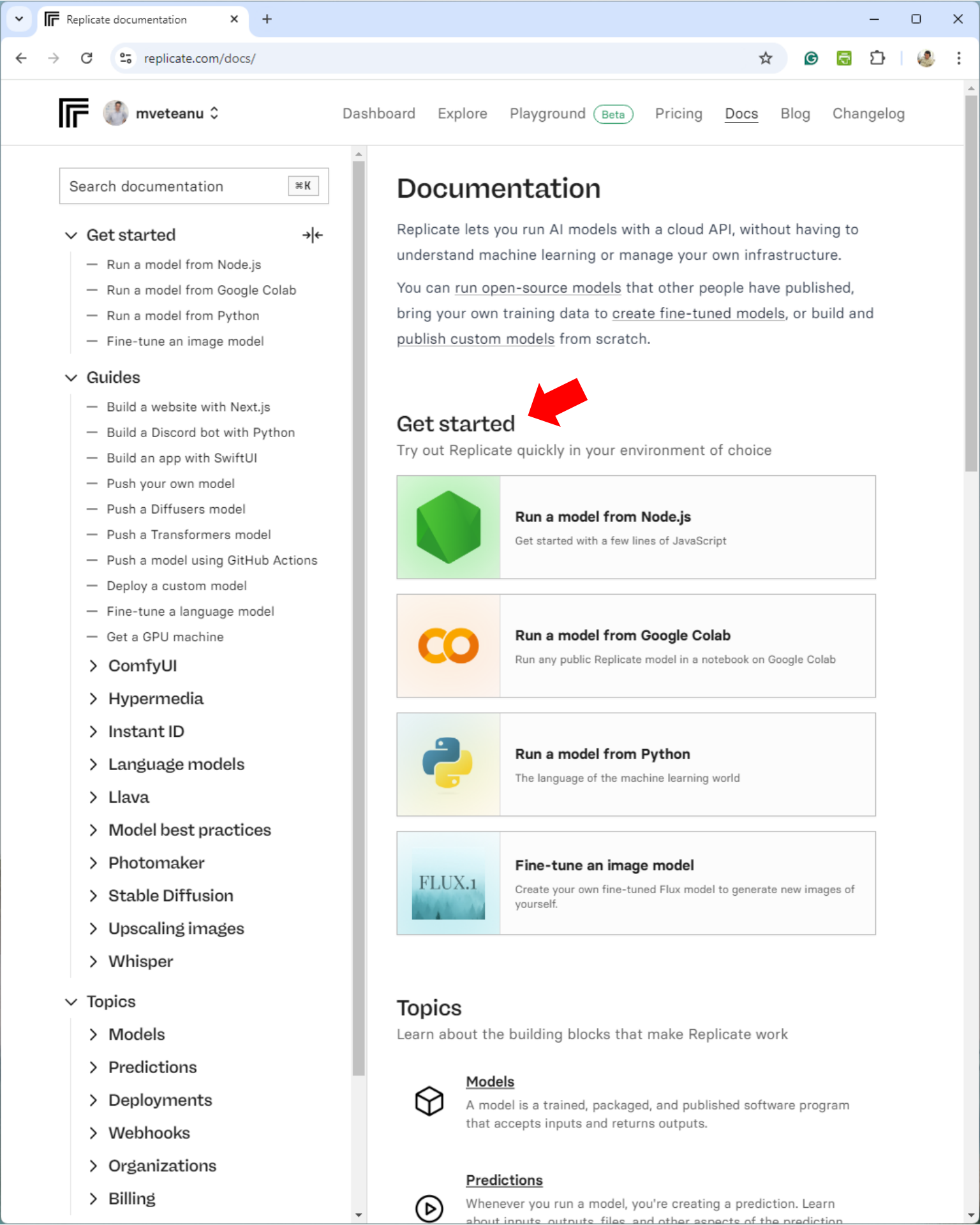
Run **google-deepmind/gemma-2b-it** with an API

Machines learn, they never sleep,
Algorithms spin, a pattern to keep.
From data's vast and ever-growing sea,
They weave insights, a knowledge to believe.

Supervised learning, a guiding hand,
Training data, shaping the sand.
Unsupervised, finding hidden gems,
Clustering data, revealing their fears.

Reinforcement, a learning spree,
Learning from mistakes, a path to believe.
Deep learning, a hidden black box,
Extracting patterns, where once there was no shock.

The future holds, a world transformed,
With machines learning, we are transformed.



Search documentation 🔍

Get started →

- Run a model from Node.js
- Run a model from Google Colab
- Run a model from Python
- Fine-tune an image model

Guides

- Build a website with Next.js
- Build a Discord bot with Python
- Build an app with SwiftUI
- Push your own model
- Push a Diffusers model
- Push a Transformers model
- Push a model using GitHub Actions
- Deploy a custom model
- Fine-tune a language model
- Get a GPU machine

- > ComfyUI
- > Hypermedia
- > Instant ID
- > Language models
- > Llava
- > Model best practices
- > Photomaker
- > Stable Diffusion
- > Upscaling images
- > Whisper

Topics

- > Models
- > Predictions
- > Deployments
- > Webhooks
- > Organizations
- > Billing

Documentation





Replicate lets you run AI models with a cloud API, without having to understand machine learning or manage your own infrastructure.

You can [run open-source models](#) that other people have published, bring your own training data to [create fine-tuned models](#), or build and [publish custom models](#) from scratch.

Get started





Try out Replicate quickly in your environment of choice

	<p>Run a model from Node.js</p> <p>Get started with a few lines of JavaScript</p>
	<p>Run a model from Google Colab</p> <p>Run any public Replicate model in a notebook on Google Colab</p>
	<p>Run a model from Python</p> <p>The language of the machine learning world</p>
	<p>Fine-tune an image model</p> <p>Create your own fine-tuned Flux model to generate new images of yourself.</p>

Topics

Learn about the building blocks that make Replicate work

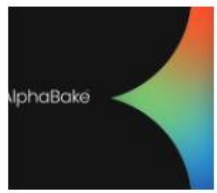
 **Models**
A model is a trained, packaged, and published software program that accepts inputs and returns outputs.

 **Predictions**
Whenever you run a model, you're creating a prediction. Learn about [inputs](#), [outputs](#), [files](#), and other aspects of the prediction.

Latest models



stability-ai / stable-diffusion-3.5-large
A text-to-image model that generates high-resolution images with fine details. It supports various artistic styles and produces diverse outputs from the same prompt, thanks to Query-Key Normalization.
Updated 2 hours ago 🔥 1.8K runs



nitishtri3d / alphabake-vton
Virtual Tryon by AlphaBake , upload your face image and see the garment on yourself.
Updated 6 hours ago 🔥 50 runs



decorx-ai / augment-experiments
Updated 7 hours ago 🔥 7.8K runs



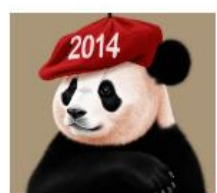
specialcarbide tools / sc...
Updated 16 hours ago 🔥 21 runs



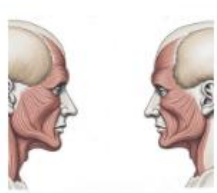
mayoita / max-mayoita3
Updated 1 day ago 🔥 12 runs



maxvanderwerf / crocs
Updated 1 day, 3 hours ago 🔥 19 runs



ailingangel / fubao-1
Updated 1 day, 3 hours ago 🔥 9 runs



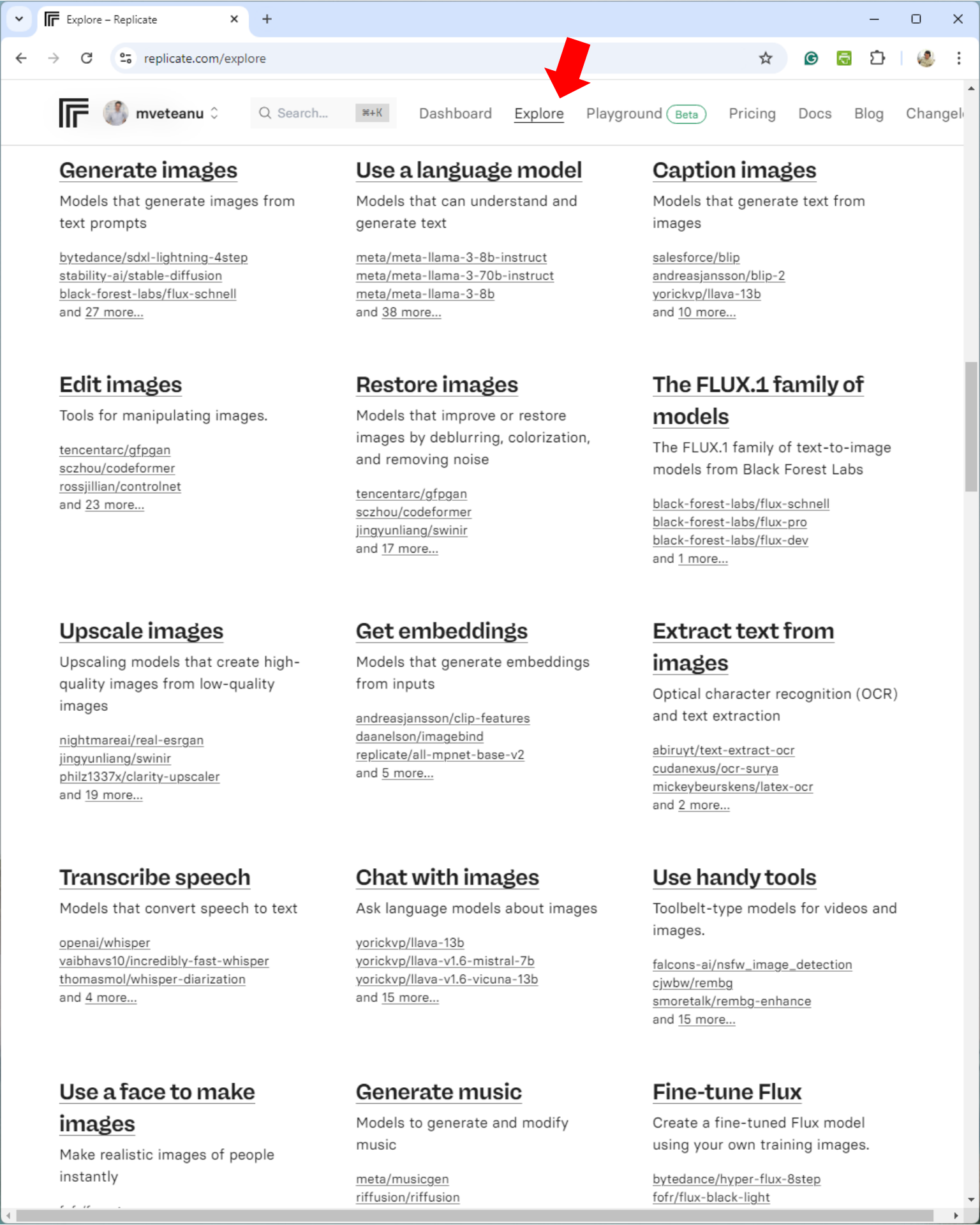
sliday / anatomy
Somewhat precise anatomical images trained on the 1918 edition of Gray's Anatomy medical textbook.
Updated 1 day, 4 hours ago 🔥 69 runs



yasseryera / yasser
Updated 1 day, 4 hours ago 🔥 258 runs



mmaluchnick / sarah
Updated 1 day, 7 hours ago 🔥 29 runs



Generate images

Models that generate images from text prompts

- [bytedance/sdxl-lightning-4step](#)
- [stability-ai/stable-diffusion](#)
- [black-forest-labs/flux-schnell](#)
- and [27 more...](#)

Use a language model

Models that can understand and generate text

- [meta/meta-llama-3-8b-instruct](#)
- [meta/meta-llama-3-70b-instruct](#)
- [meta/meta-llama-3-8b](#)
- and [38 more...](#)

Caption images

Models that generate text from images

- [salesforce/blip](#)
- [andreasjansson/blip-2](#)
- [yorickvp/llava-13b](#)
- and [10 more...](#)

Edit images

Tools for manipulating images.

- [tencentarc/gfpgan](#)
- [sczhou/codeformer](#)
- [rossjillian/controlnet](#)
- and [23 more...](#)

Restore images

Models that improve or restore images by deblurring, colorization, and removing noise

- [tencentarc/gfpgan](#)
- [sczhou/codeformer](#)
- [jingyunliang/swinir](#)
- and [17 more...](#)

The FLUX.1 family of models

The FLUX.1 family of text-to-image models from Black Forest Labs

- [black-forest-labs/flux-schnell](#)
- [black-forest-labs/flux-pro](#)
- [black-forest-labs/flux-dev](#)
- and [1 more...](#)

Upscale images

Upscaling models that create high-quality images from low-quality images

- [nightmareai/real-esrgan](#)
- [jingyunliang/swinir](#)
- [philz1337x/clarity-upscaler](#)
- and [19 more...](#)

Get embeddings

Models that generate embeddings from inputs

- [andreasjansson/clip-features](#)
- [daanelson/imagebind](#)
- [replicate/all-mpnet-base-v2](#)
- and [5 more...](#)

Extract text from images

Optical character recognition (OCR) and text extraction

- [abiruyt/text-extract-ocr](#)
- [cudanexus/ocr-surya](#)
- [mickeybeurskens/latex-ocr](#)
- and [2 more...](#)

Transcribe speech

Models that convert speech to text

- [openai/whisper](#)
- [vaibhavs10/incredibly-fast-whisper](#)
- [thomasmol/whisper-diarization](#)
- and [4 more...](#)

Chat with images

Ask language models about images

- [yorickvp/llava-13b](#)
- [yorickvp/llava-v1.6-mistral-7b](#)
- [yorickvp/llava-v1.6-vicuna-13b](#)
- and [15 more...](#)

Use handy tools

Toolbelt-type models for videos and images.

- [falcons-ai/nswf_image_detection](#)
- [cjwbw/rembg](#)
- [smoretalk/rembg-enhance](#)
- and [15 more...](#)

Use a face to make images

Make realistic images of people instantly

Generate music

Models to generate and modify music

- [meta/musicgen](#)
- [riffusion/riffusion](#)

Fine-tune Flux

Create a fine-tuned Flux model using your own training images.

- [bytedance/hyper-flux-8step](#)
- [fofr/flux-black-light](#)

Miscellaneous

Other services for AI inference, Model Serving and MLOps

baseten:	https://www.baseten.co/
octoai:	https://octo.ai/
modal:	https://modal.com/
fireworks ai:	https://fireworks.ai/
deepinfra:	https://deepinfra.com/
Banana.dev:	https://www.banana.dev/
DataRobot:	https://www.datarobot.com/
RunPod:	https://www.runpod.io/
LambdaLabs:	https://lambdalabs.com/
Mistral AI:	https://mistral.ai/



Buzzwords used in this brochure

- 1. Generative AI:** AI that creates new content like text, images, or music from input data (e.g., GPT, DALL·E).
- 2. AI Hallucinations:** When AI generates incorrect or nonsensical information not based on reality or input.
- 3. Foundation Model:** Large, pre-trained models used as a base for various AI tasks, fine-tuned for specific applications.
- 4. LLM (Large Language Model):** AI models that process and generate natural language (e.g., GPT, BERT).
- 5. Transformer:** A neural network that uses self-attention for parallel processing, key for models like GPT and BERT.
- 6. Tokenization:** Breaking text into smaller units (tokens) for model input.
- 7. Fine-tuning:** Adapting a pre-trained model to specific tasks with additional training on a new dataset.
- 8. Prompt Engineering:** Crafting prompts to guide AI models in generating the desired output.
- 9. RAG (Retrieval-Augmented Generation):** Combines retrieving relevant data with generation models to improve outputs.
- 10. Vector Database:** Stores and queries vector embeddings for tasks like similarity search and recommendation.
- 11. Embeddings:** Numerical representations of data used for tasks like search and classification.
- 12. Multimodal Models:** Models that process multiple data types, like text, images, and audio.
- 13. Inference:** Using a trained model to make predictions on new input data.
- 14. MLOps:** Practices for deploying and managing machine learning models in production.

Marian Veteanu

Technology Architect and Product Leader

Looking to see how I can
add value to your organization?

Message me!

<https://www.linkedin.com/in/mveteanu/>
<https://x.com/mveteanu>

