

[Featured](#)[Getting started](#)[Hello, world](#)[Simple web scraper](#)[Serving web endpoints](#)[Large language models \(LLMs\)](#)

## Featured examples



### Serving Diffusion models

Serve Stable Diffusion XL on Modal with a number of optimizations for blazingly fast inference.

[View on GitHub](#)

### Serverless TensorRT-LLM

Run large language models at SOTA performance by building a TRT-LLM engine on Modal.

[View on GitHub](#)

### Stable Diffusion fine-tuning with Dreambooth



### Voice chat with LLMs

Build a real-time voice chat app by combining speech-to-text, an LLM, and text-to-speech.

Fine-tune a Stable Diffusion model on images of your pet using Dreambooth.

[View on GitHub](#)



## Fast podcast transcriptions

Build an end-to-end podcast transcription app that leverages dozens of containers for super-fast processing.

[View on GitHub](#)



## Document OCR job queue

Use Modal as an infinitely scalable job queue that can service async tasks from a web app.

[View on GitHub](#)



## Parallel processing of Parquet files on S3

Analyze data from the Taxi and Limousine Commission of NYC in parallel.

[View on GitHub](#)



## Retrieval-Augmented Generation for Q&A

Build a question-answering web endpoint that can cite its sources.

[View on GitHub](#)



## Hacker News Slackbot



## Real-time Object Detection

Use Modal to deploy a cron job that periodically queries Hacker News for new posts, and posts the results to Slack.

[View on GitHub](#)

Create a web endpoint that leverages HuggingFace models to do object detection in real-time.

[View on GitHub](#)



## ControlNet playgrounds

Play with all 10 demo Gradio apps from the ControlNet project.

[View on GitHub](#)



© 2024

[About](#)

[Status](#)

[Changelog](#)

[Documentation](#)

[Slack Community](#)

[Pricing](#)

[Examples](#)