Outskill

HuggingFace & Open Source

Week 2, Day 1 | GEF C2

HuggingFace

The most happening place for Gen-AI news after X/twitter.
Tons of amazing resources.

What are “models”?

HuggingFace Walkthrough: Models

HuggingFace Walkthrough: Companies

HuggingFace Walkthrough: Datasets

HuggingFace Walkthrough: Spaces

Can you get ChatGPT or Calude from HuggingFace?

Can you get ChatGPT or Calude from HuggingFace?

What is OpenSource? Open Weight

Learning to Read LLM Model Names

- XB = X billion parameters
- XB YAB = X billion parameters, Y active (MoE - Mixture of Experts)
- Pretrained/Base/Foundational
- Instruct/Chat (IT)
- Quantised

PS: There is no standard to naming models!

Outskill

HuggingFace: APIs

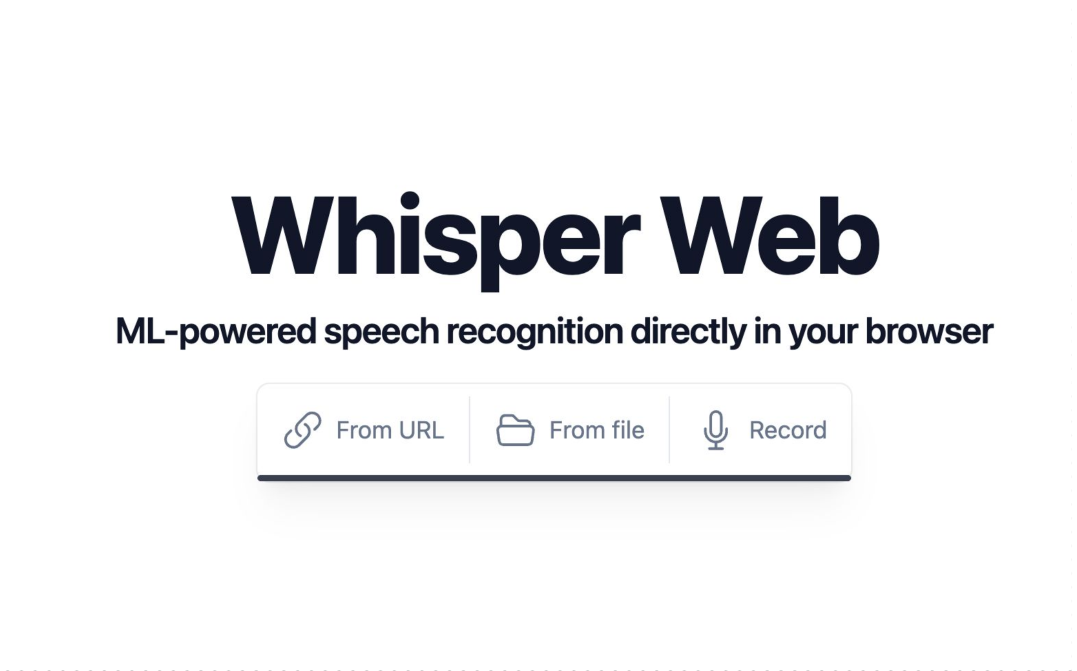
Using free API Keys/Access Tokens

HuggingFace: APIs

1. Sentiment Analysis/Classification
2. Rating Classification
3. Timeseries
4. Zero-shot Classification
5. Text Summarisation
6. Question Answering
7. Transformer Search
8. Text Generation
9. Gen-AI LLM
10. Image Diffusion

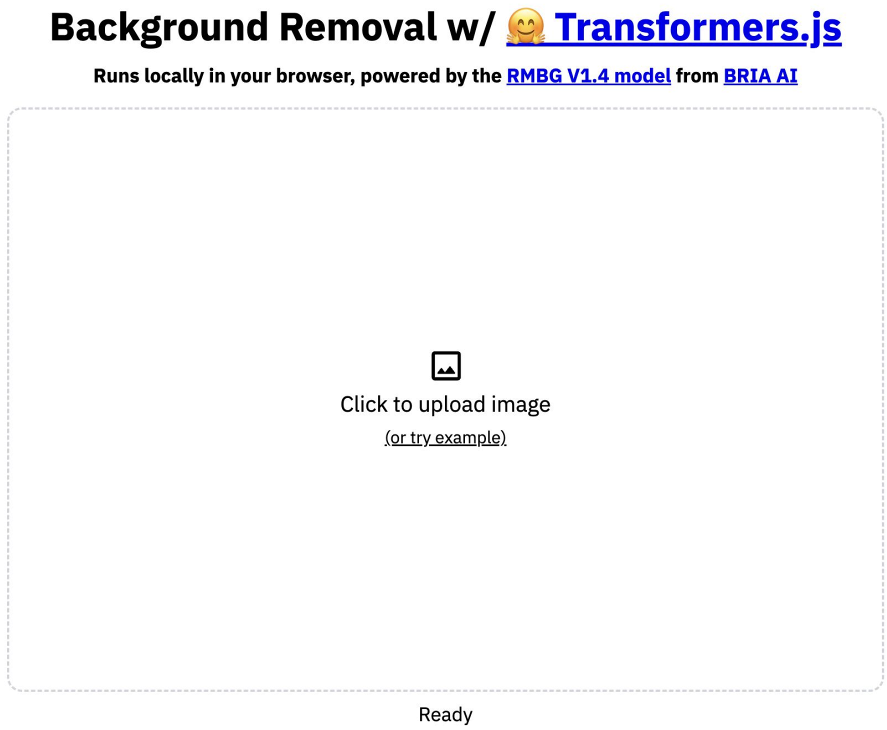
HuggingFace: Interesting Models

https://huggingface.co/spaces/Xenova/whisper-web



Whisper Web

ML-powered speech recognition directly in your browser https://huggingface.co/spaces/Xenova/remove-background-web



Background Removal w/ 🤗 Transformers.js

Runs locally in your browser, powered by the RMBG V1.4 model from BRIA AI

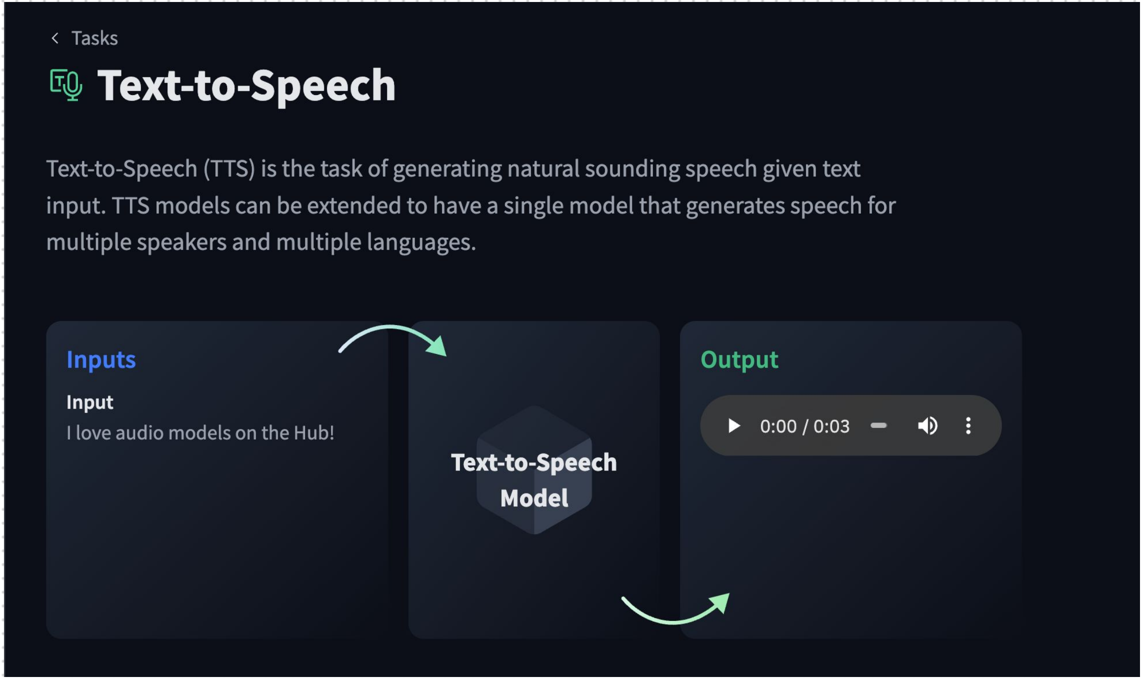
Click to upload image

(or try example)

Ready

https://huggingface.co/tasks/text-to-speech

Outskill



‹ Tasks

Text-to-Speech

Text-to-Speech (TTS) is the task of generating natural sounding speech given text input. TTS models can be extended to have a single model that generates speech for multiple speakers and multiple languages.

Inputs

Input

I love audio models on the Hub!

https://huggingface.co/nari-labs/Dia-1.6B

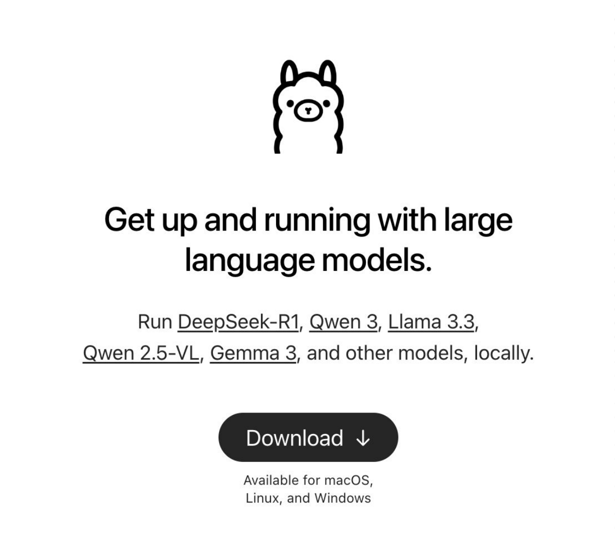
Nari Dia-1.6B

[S1] Dia is a text to dialogue model.
[S2] You get full control over scripts and voices.
[S1] Wow. Amazing. (laughs)

Running LLM’s Locally

Did you install ollama or LMStudio?

2 Options

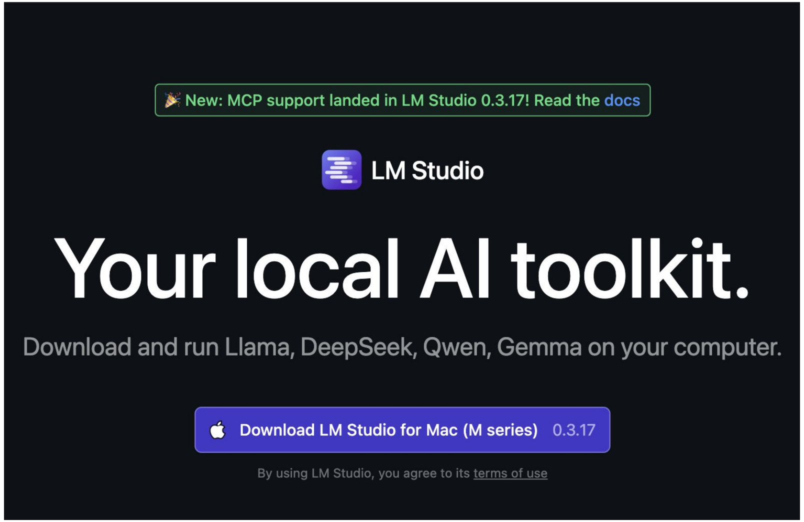


Get up and running with large language models.

Run DeepSeek-R1, Qwen 3, Llama 3.3, Qwen 2.5-VL, Gemma 3, and other models, locally.

Available for macOS,
Linux, and Windows

https://ollama.com/



🎉 New: MCP support landed in LM Studio 0.3.17! Read the docs

Your local AI toolkit.

Download and run Llama, DeepSeek, Qwen, Gemma on your computer.

By using LM Studio, you agree to its terms of use

https://lmstudio.ai/

Other Options

1. LM Studio https://lmstudio.ai/ ✨
    a. Includes most optimisations including GGUF (quantisation) + MLX + llama.cpp
2. Msty https://msty.app/ (advanced paid features, multi LLM chat)
3. GPT4All https://www.nomic.ai/gpt4all
4. Ollama https://ollama.com/ (no true GUI)
5. Jan.ai https://jan.ai/
6. Transformer Lab (setup is tricky) https://transformerlab.ai/
7. Special for Apple Silicon Mac's: https://github.com/ml-explore/mlx

Gemma 3 by Google

Demo Local LLM’s as API’s

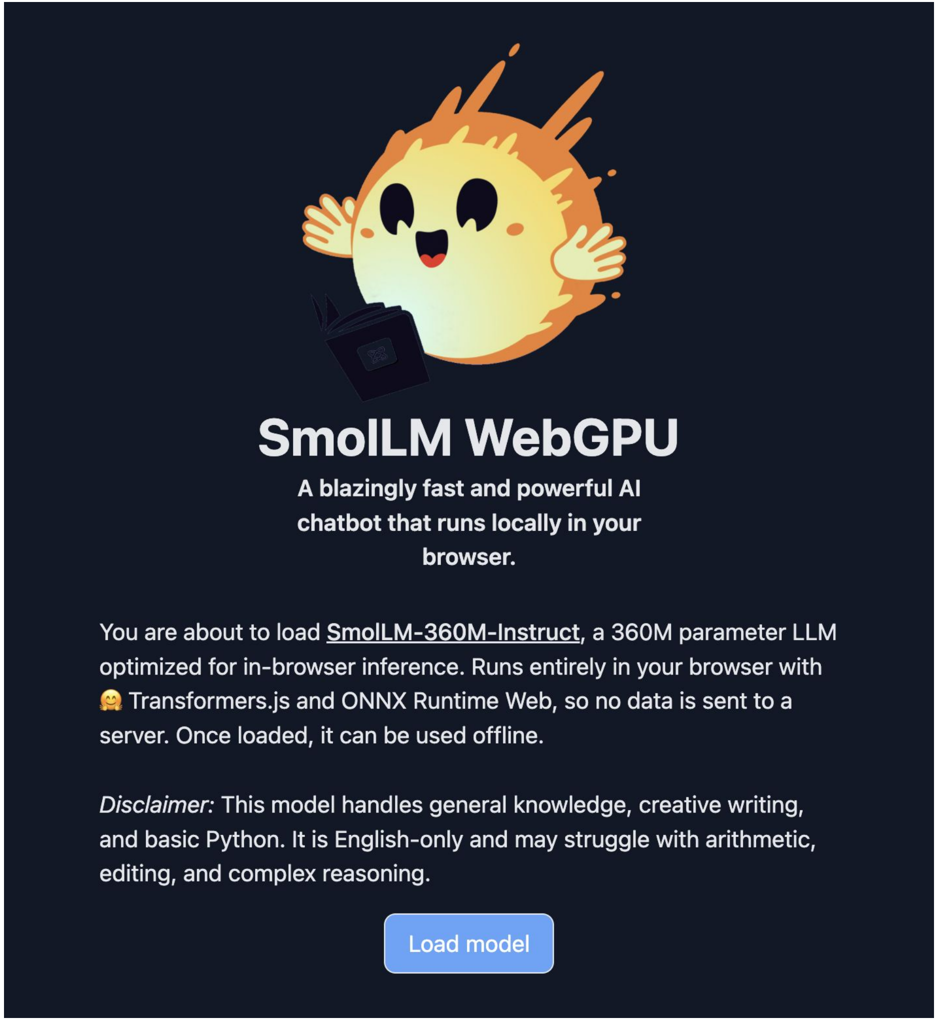
Model Options

1. Mistral/Mixtral
2. Gemma 2 & 3
3. Quen 2 & 3
4. Llama 3.1 (4 was a disaster)
5. DeepSeek R1

Model Optimisation Techniques

1. Do maths in C++ rather than Python (llama.cpp)
2. Quantisation & Quantisation Aware Training (QAT) — change 32 bits precision to 16, 16 to 8, 8 to 4bits!
3. Distillation: Using outputs of a larger model (teacher) to train smaller models (student) (R1 Distill)
4. MoE: Mixture of Experts replacing Feed-Forward Layers

Small Language Models (SLM)



SmolLM WebGPU

A blazingly fast and powerful AI chatbot that runs locally in your browser.

You are about to load SmolLM-360M-Instruct, a 360M parameter LLM optimized for in-browser inference. Runs entirely in your browser with 🤗 Transformers.js and ONNX Runtime Web, so no data is sent to a server. Once loaded, it can be used offline.

Disclaimer: This model handles general knowledge, creative writing, and basic Python. It is English-only and may struggle with arithmetic, editing, and complex reasoning.

Small Language Models

1. Gemma 3 1B & 4B (multi-modal)
2. Quen 3 0.6B
3. TinyLlama https://huggingface.co/TinyLlama/TinyLlama-1.1B-Chat-v1.0
4. Phi by Microsoft
5. Mobile BERT
6. DeepSeek R1 Distill Quen 1.5B

Building a Profile on HuggingFace

1. Research
2. Models
3. Applications & Spaces
4. Datasets
5. Comments, Networking, etc.

Measuring Agents

What agent are you going to build?