Hands-on Hugging Face

Myles Harrison

December 14th, 2024

@nlpfromscratch

Agenda

01 Introduction

02 Fundamentals

03 Generative Text Models

04 Image Generation Models

05 Conclusion & Resources

Disclaimer

The materials presented herein are the author's own and do not necessarily reflect the views of Hugging Face, nor any other organization or individual.

The author is solely responsible for the accuracy and completeness of the information presented, and any errors or omissions are unintentional.

This presentation is not endorsed, affiliated with, nor sponsored by Hugging Face, and the author has no affiliation with this company.

Who am I?

- Data Scientist
- Entrepreneur
- Consultant
- Educator
- Community Builder

Jupyter AI
Jupyter AI integrates generative AI models into Jupyter, allowing you leverage LLMs in notebooks and improve your productivity
An %%ai magic that turns the Jupyter notebook into a reproducible generative AI playground
Jupyternaut

Manifesto

Knowledge is only valuable if it is useful.

The best way to learn is by doing.

Learning is a non-linear process.

Learning is not a journey, it is guided exploration.

Teaching and learning are complementary.

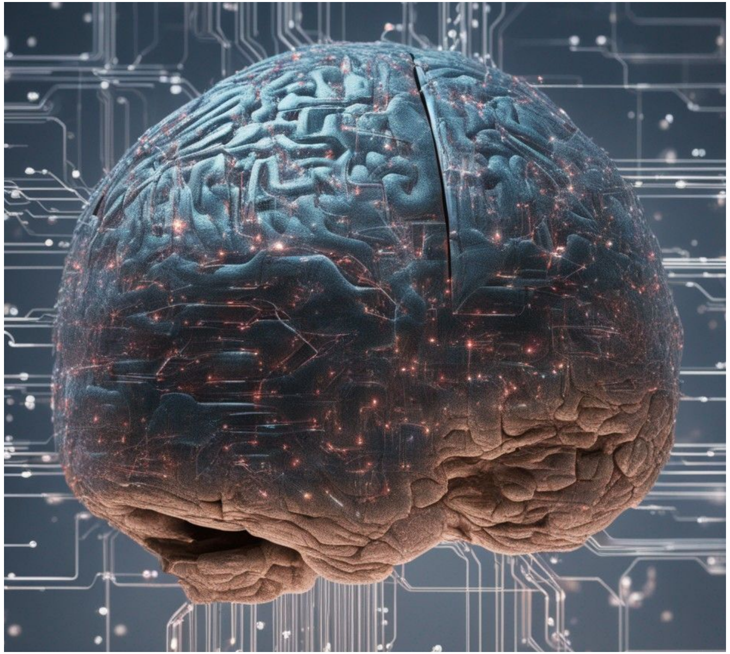
Fundamentals

What is a Large Language Model?

ChatGPT is an example of a large language model (LLM), a type of *deep learning model* trained with hundreds of millions or billions of parameters on very large bodies of text. Large language models currently represent the state of the art in natural language processing (NLP) applications.

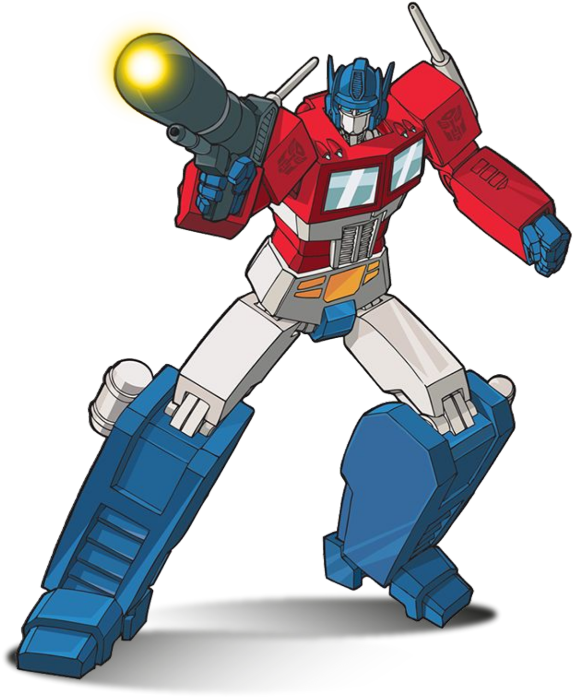
While we're here, ChatGPT is not sentient, nor is it an example of an Artificial General Intelligence (AGI).

Let's take a step back...



The Transformer Architecture

- Groundbreaking paper "Attention is All You Need" from Google researchers (Vaswani et al, 2017) introduced Transformer architecture
- Original application in machine translation but now general purpose and applied to a myriad of other tasks
- Represents the state of the art for LLMs and also applied in domains outside of language (image generation) - virtually all new models based on this architecture
- Popularized by OpenAI and the Generative Pretrained Transformer (GPT) series of models

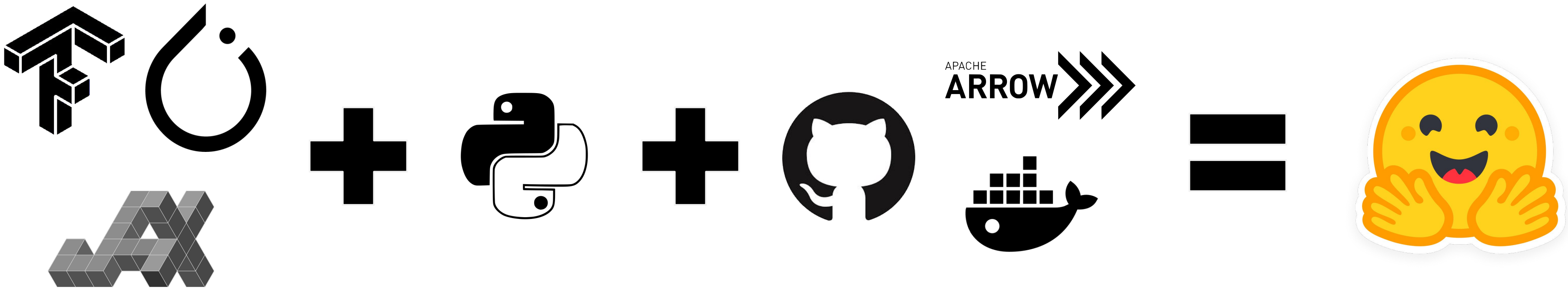


Hugging Face

Hugging Face is a software company founded in 2013 and based in New York city. As of August 2023, the company is in Series 'D' funding with a valuation of $4.5B and backing from companies such as Salesforce, Google, Amazon, IBM, Nvidia, AMD, and Intel.

While this name refers to the company, it also refers to the software and platform they develop for working with large language models and data in the natural language processing and other domains.

The `datasets` library allows working with data hosted on the platform, and the `transformers` library for working with models of this type. There are also other libraries for working with specialized types of models (*e.g.* `diffusers` for diffusion models) and data processing and model optimization.


APACHE
ARROW

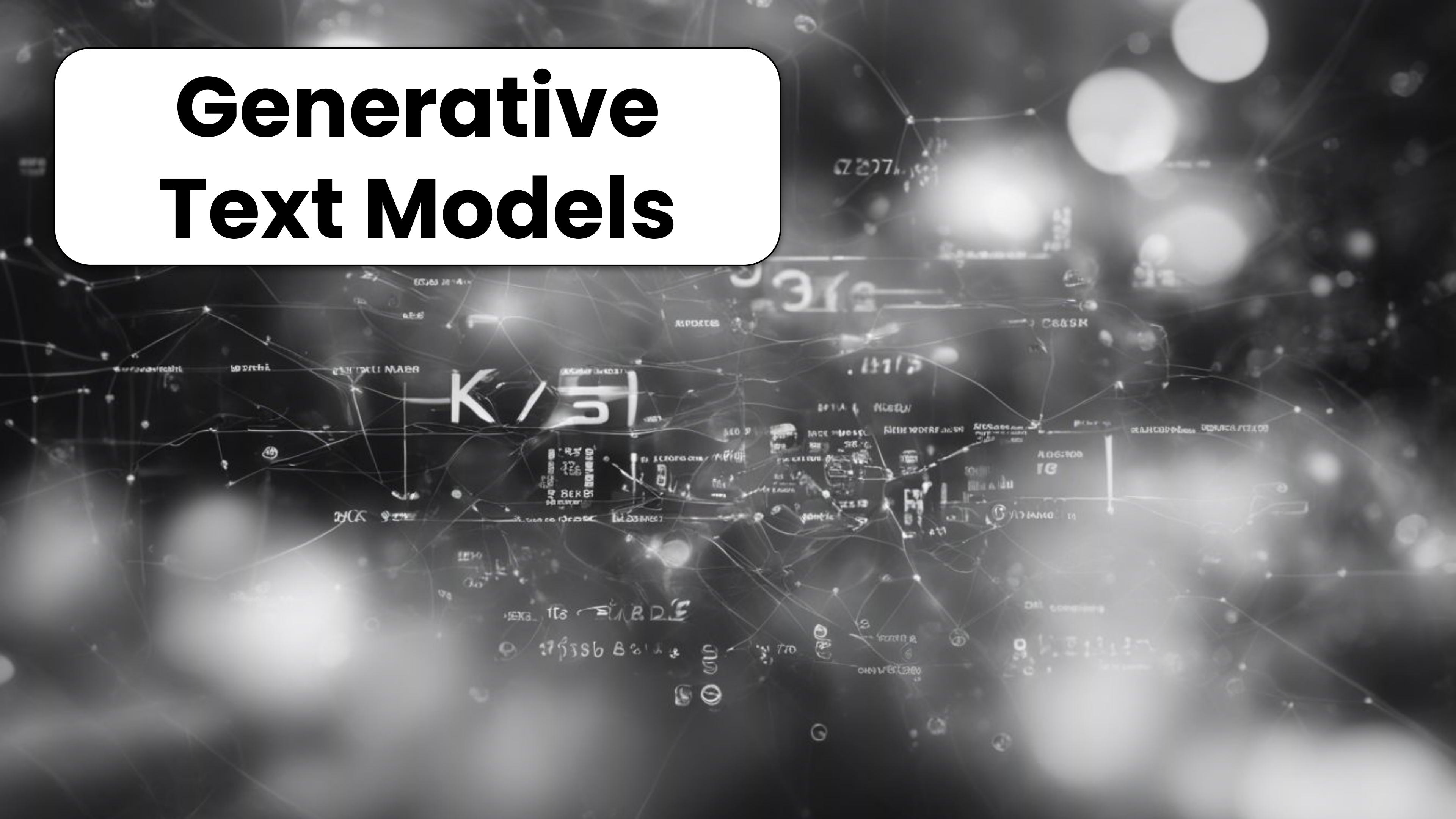
Tools of the Trade

Python 3

Google Colab
/ Jupyter

Deep Learning and LLMs

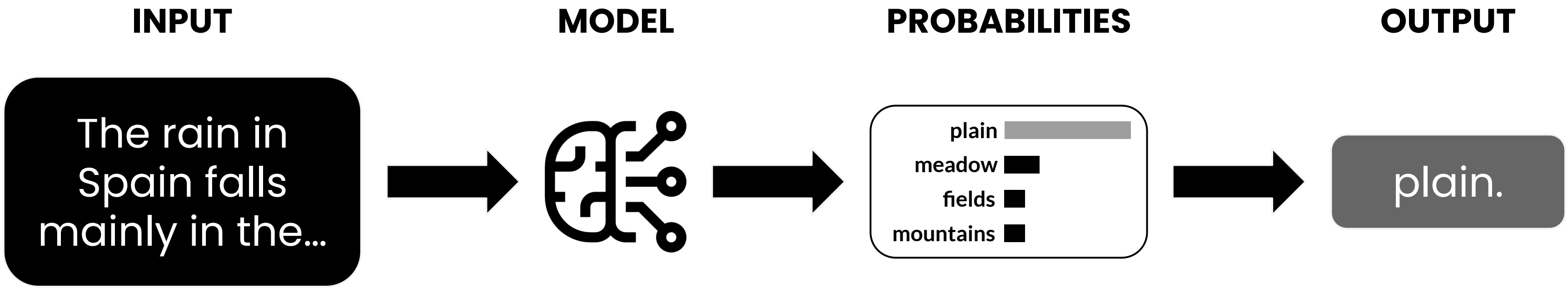
Generative Text Models



How do LLMs generate text?

When generating text, the model assigns probabilities to all possible tokens based on its understanding of the entire context. It then selects the next token in the output based on these probabilities.

There are different parameters we can specify when generating text from a model to vary the outputs thereof.



GPT - The Most Famous LLM

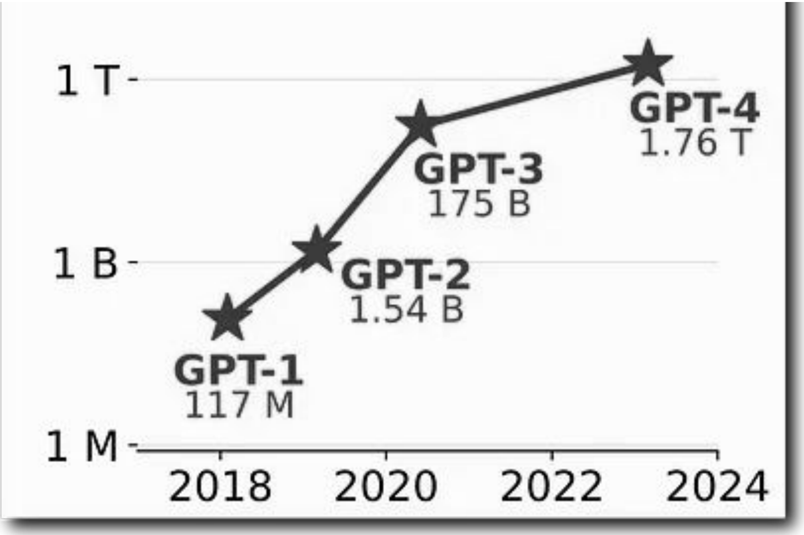
OpenAI

Undoubtedly, the most popularly known generative text model is that of the Generative Pretrained Transformer (GPT) by OpenAI.

The GPT series of models are of ever increasing size and trained on increasingly large and more comprehensive datasets (right)

While GPT-3 remains proprietary and only available to use through the OpenAI API, the weights of GPT-2 are publicly available and can also be accessed through Hugging Face.

Let's take a look at generating text with GPT-2.



GPT Series Parameter Counts by model
Image credit: Francesco Casalegno

Instruction-Tuned (“Chat”) Models

```
conversation = [
  {"role": "user", "content": "Hello, how
are you?"},
  {"role": "assistant", "content": "I'm
doing great. How can I help you today?"},
]

Tokenizer = AutoTokenizer.from_pretrained(
"microsoft/Phi-3-mini-4k-instruct")

tokenizer.apply_chat_template(conversation,
tokenize=False))
```

<|user|>Hello, how are you?<|end|>

<|assistant|>
I'm doing great. How can I help you today?<|end|>

<|endoftext|>

Message Roles

SYSTEM

Sets the behavior of the assistant - how it should behave at the conversation level (optional)

USER

Provide requests or input to which the assistant will respond (*i.e.* the prompts)

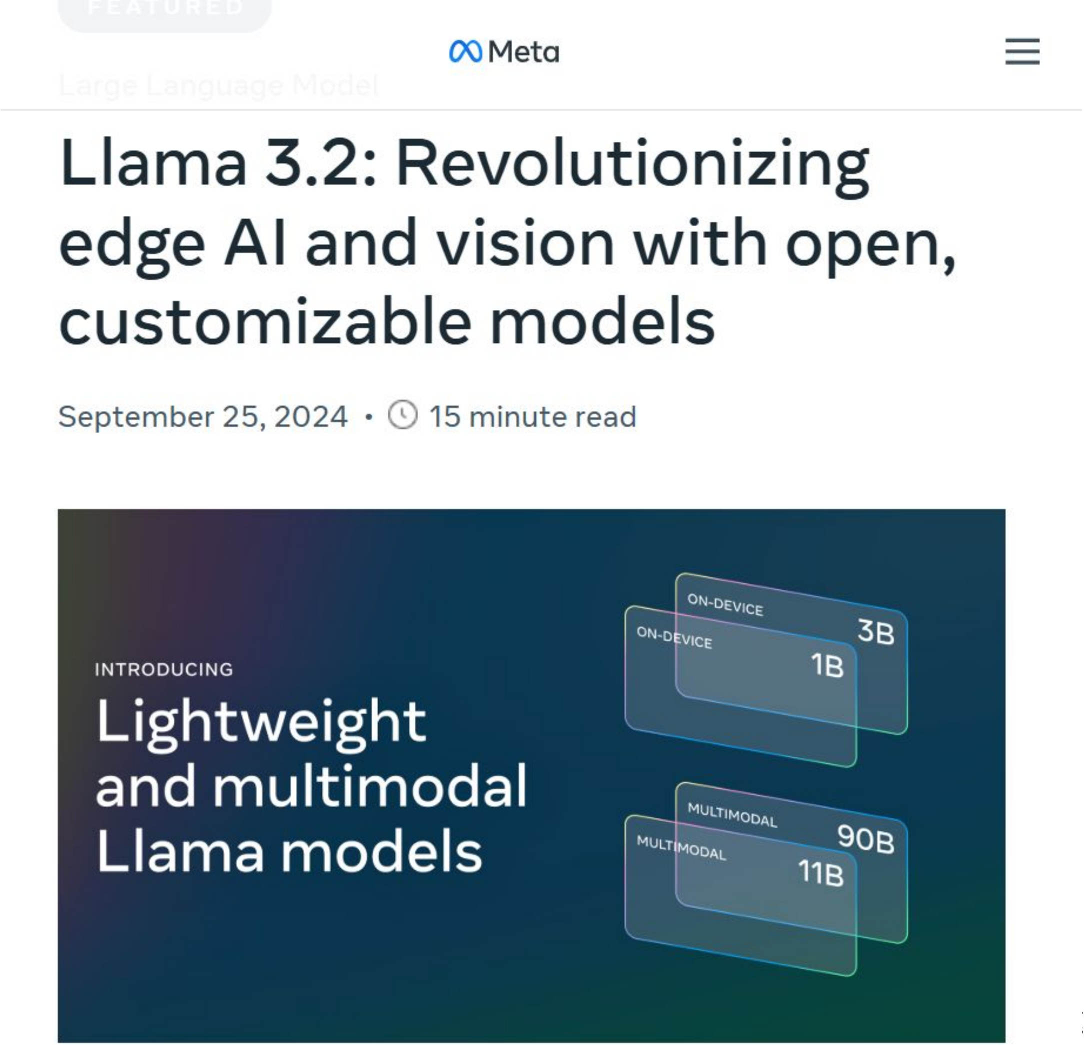
ASSISTANT

Responses from the model. Can be used to include conversation history when it is important (optional)

Meta LLaMA 3.2

Released September 25th, 2024

Family of LLMs with various sizes and text-only (1B and 3B) and multimodal versions (11B and 90B)



FEATURED

Meta

Large Language Model

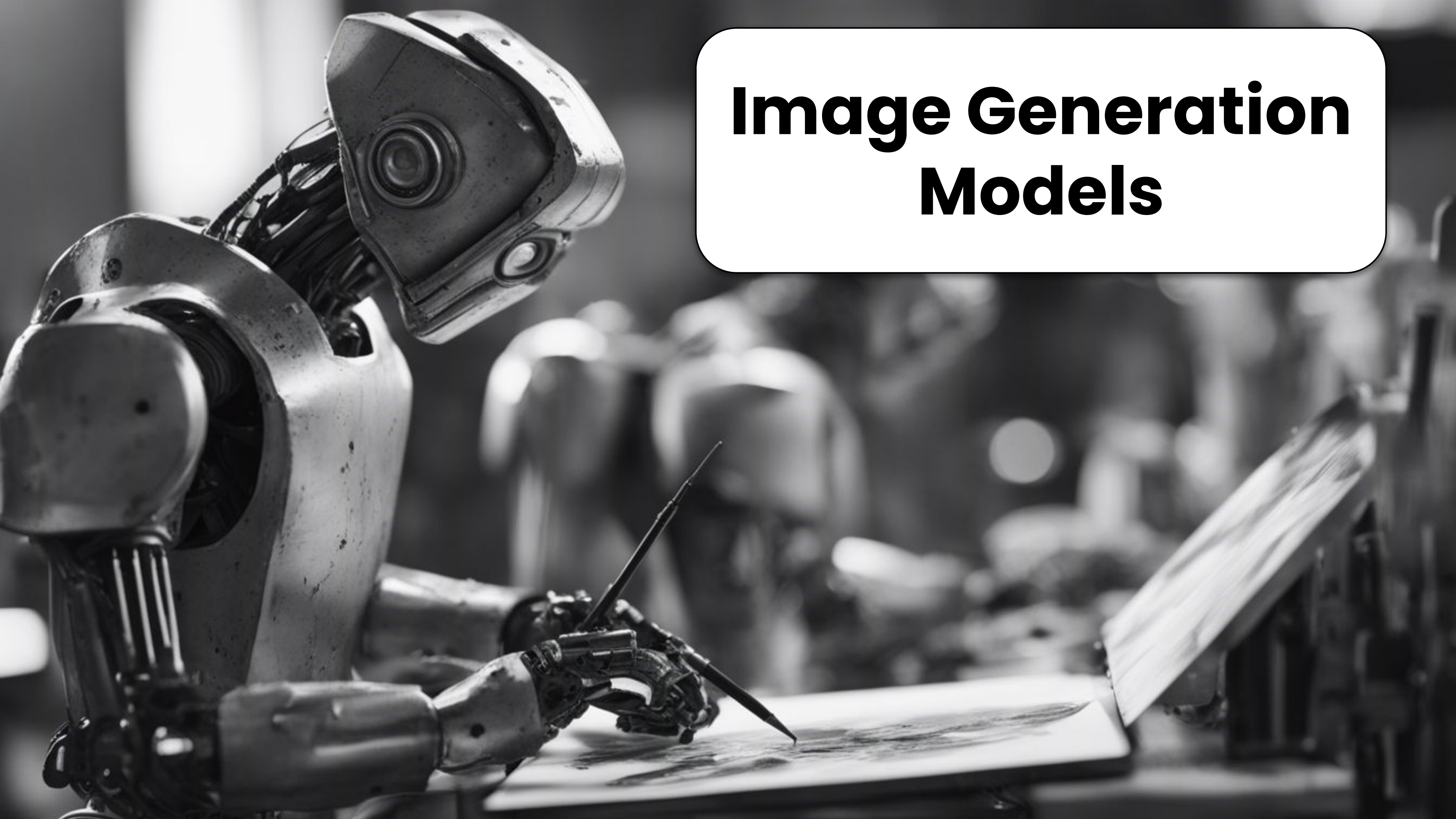
Llama 3.2: Revolutionizing edge AI and vision with open, customizable models

September 25, 2024 • 15 minute read

INTRODUCING

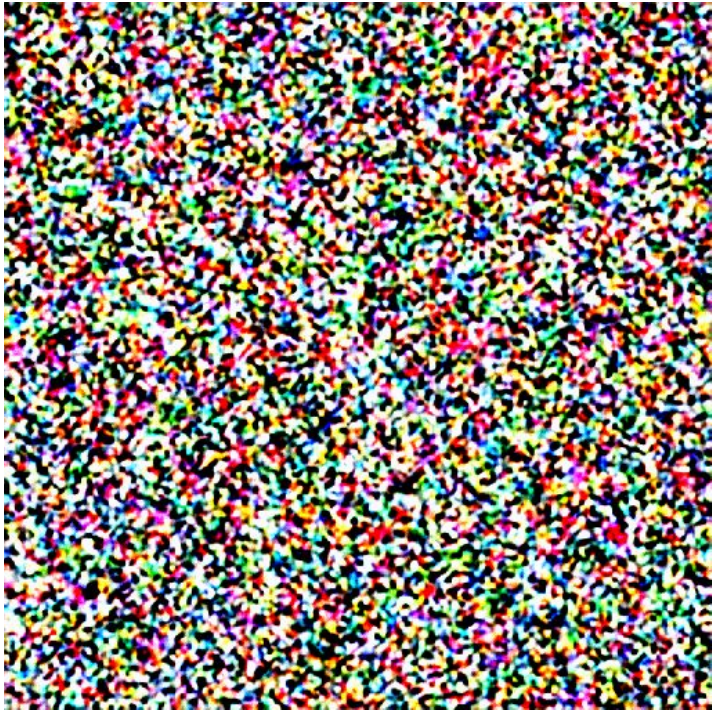
Lightweight and multimodal Llama models

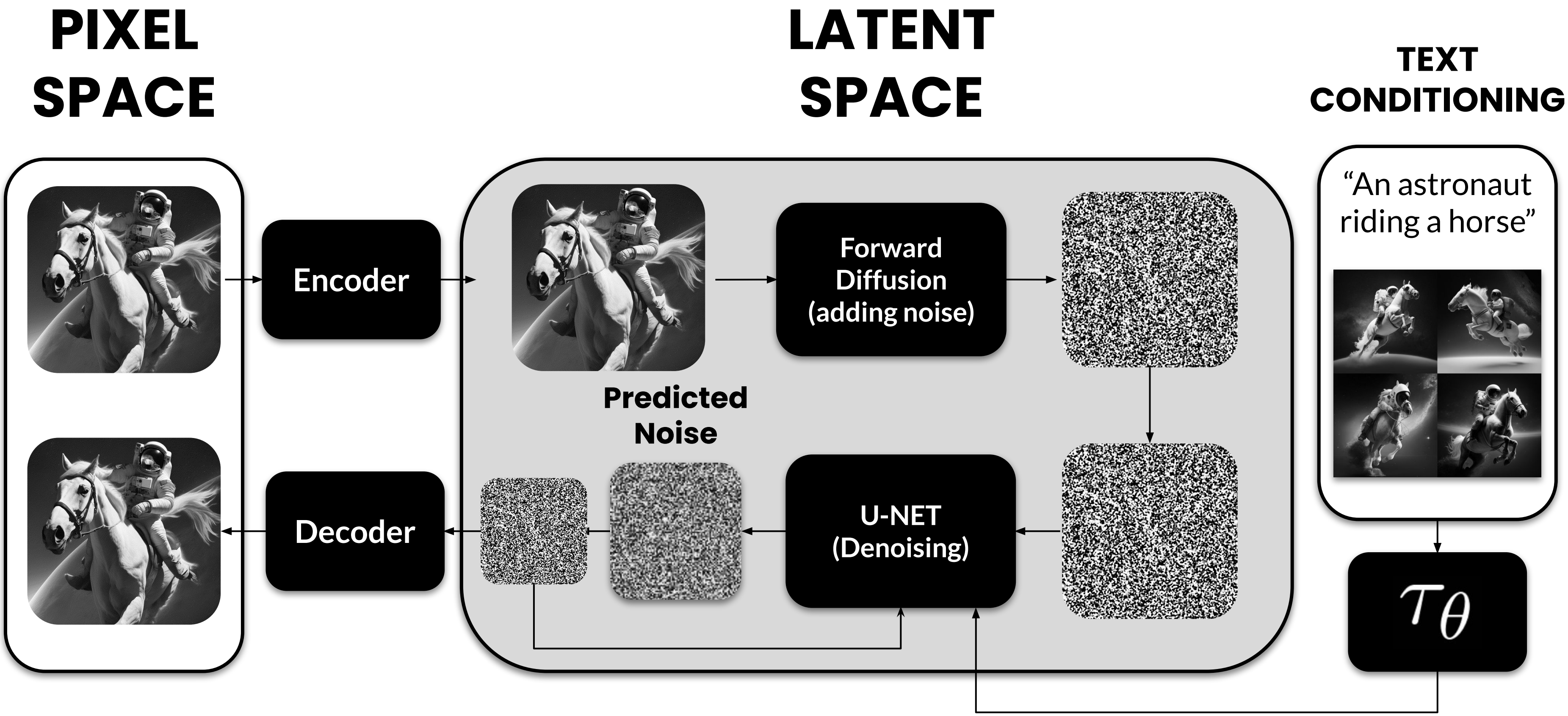
Image Generation Models



Stable Diffusion

- Latent diffusion models (Rombach et. al, 2021) are a type of generative AI model that can create images by iteratively refining random noise, guided by a learned representation of patterns in data (a "latent space")
- These models start with random noise and use a neural network to "denoise" step by step, transforming it into a detailed image by following patterns learned from a large dataset of images and captions.
- Model learns to predict the added noise during training, then denoises during prediction
- Can be conditioned with text via the important CLIP model learning representations between text and images (OpenAI, 2021)





PIXEL SPACE

LATENT SPACE

TEXT CONDITIONING

Encoder

Forward Diffusion (adding noise)

Predicted Noise

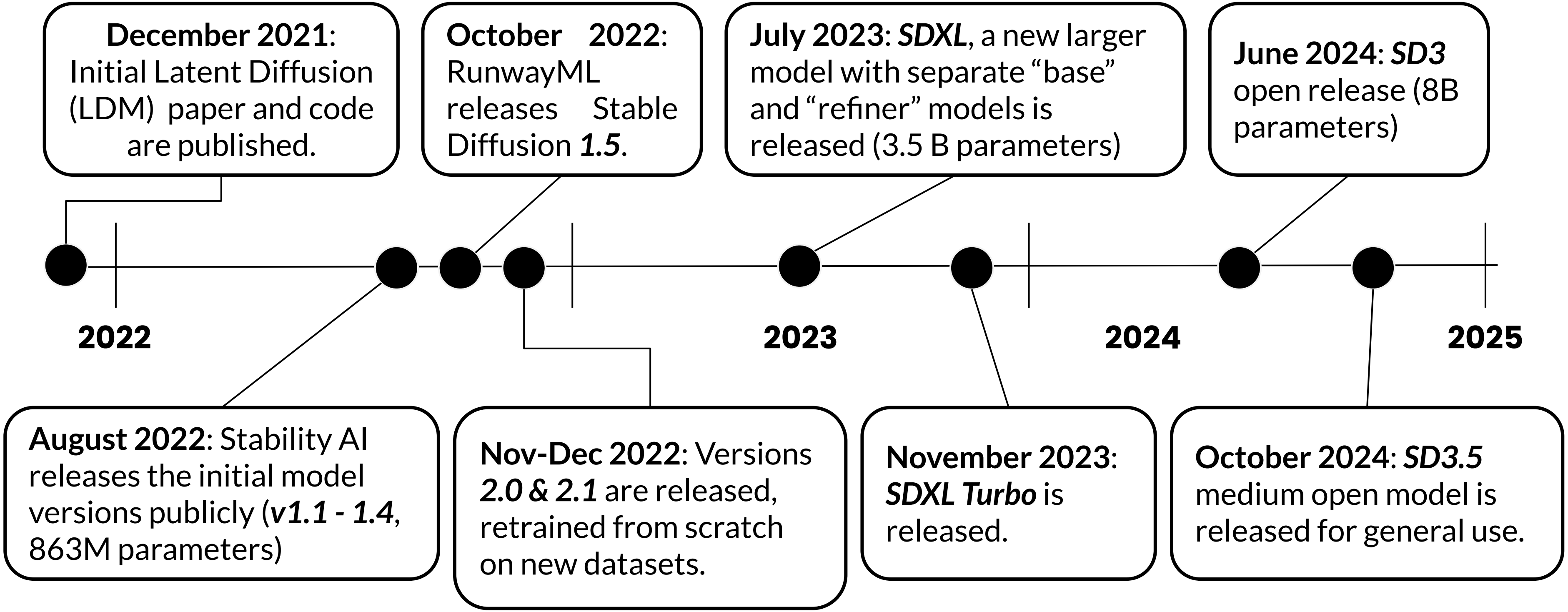
U-NET (Denoising)

Decoder

“An astronaut riding a horse”

τ_θ

Stable Diffusion Timeline



December 2021: Initial Latent Diffusion (LDM) paper and code are published.

August 2022: Stability AI releases the initial model versions publicly (***v1.1 - 1.4***, 863M parameters)

October 2022: RunwayML releases Stable Diffusion ***1.5***.

Nov-Dec 2022: Versions ***2.0 & 2.1*** are released, retrained from scratch on new datasets.

July 2023: ***SDXL***, a new larger model with separate “base” and “refiner” models is released (3.5 B parameters)

November 2023: ***SDXL Turbo*** is released.

June 2024: ***SD3*** open release (8B parameters)

October 2024: ***SD3.5*** medium open model is released for general use.

Stable Diffusion XL (SDXL)

- Released July 2023 by researchers at Stability AI, the successor to Stable Diffusion 2.1
- 3x in size to (core of) original model
- Additional refiner model (image-to-image) for denoising used in a supplementary fashion after base model for high fidelity outputs
- Available through Clipdrop (paid) and on Hugging Face spaces (free, various)
- Now near real-time image generation “as you type” with SDXL Turbo

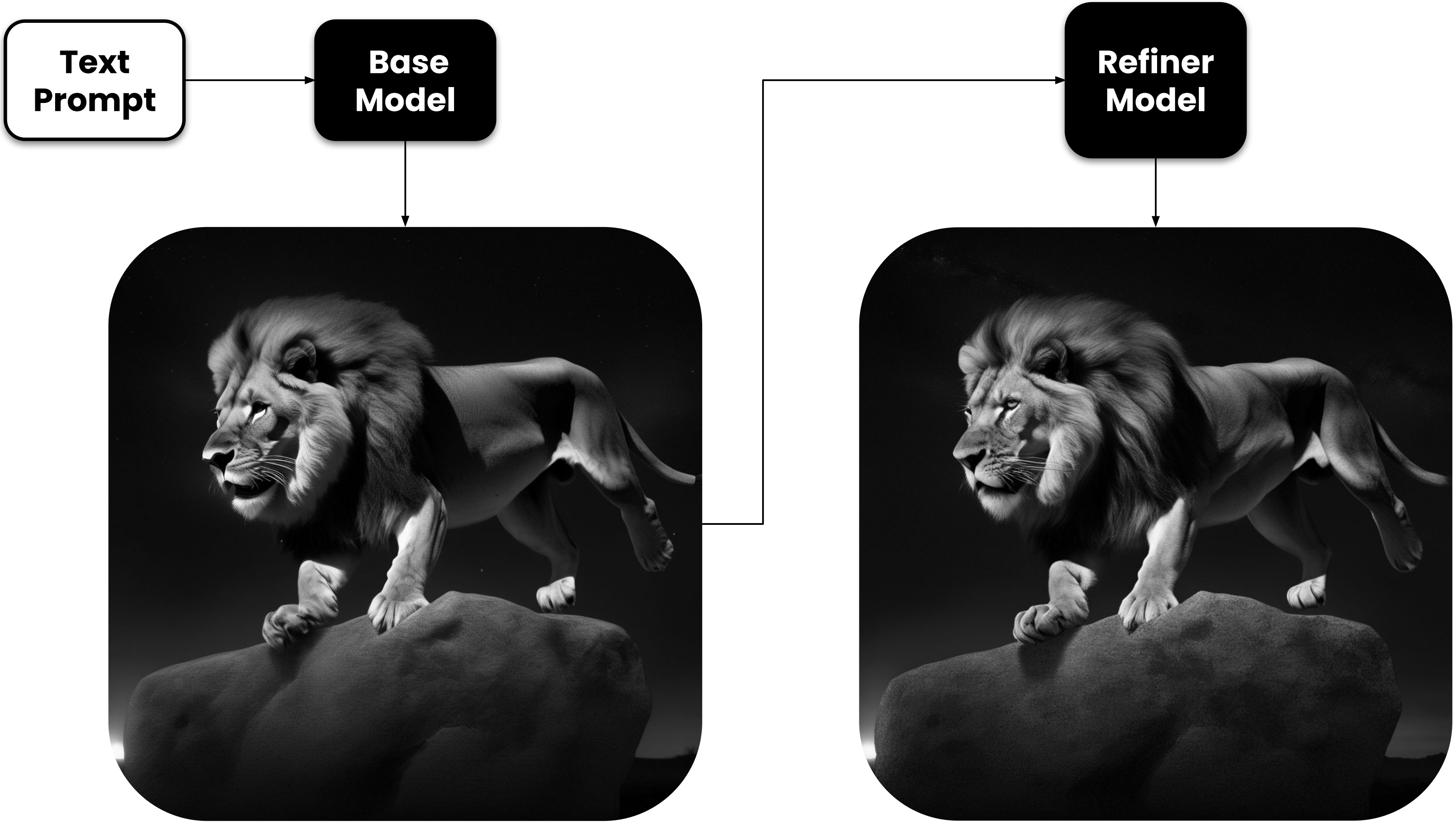
Hugging Face 🤗: SDXL in 5 lines of code

```
from diffusers import AutoPipelineForText2Image
import torch

pipeline =
AutoPipelineForText2Image.from_pretrained(
    "stabilityai/stable-diffusion-xl-base-1.0",
     torch_dtype=torch.float16, variant="fp16",
     use_safetensors=True
).to("cuda")

image = pipeline(prompt="A cute dog in a fuzzy
sweater").images[0]

image.save("dog.png")
```



Text Prompt

Base Model

Refiner Model

Images from huggingface.co/docs/diffusers/en/using-diffusers/sdxl

Flux

- Announced 2024/08/01

- Team of original creators of Stable Diffusion created startup Black Forest Labs

- $231M in seed from a16z

- 12B transformer/diffusion flow-based model in 3 versions: Pro, Dev, and Schnell (Apache 2.0 licensed)

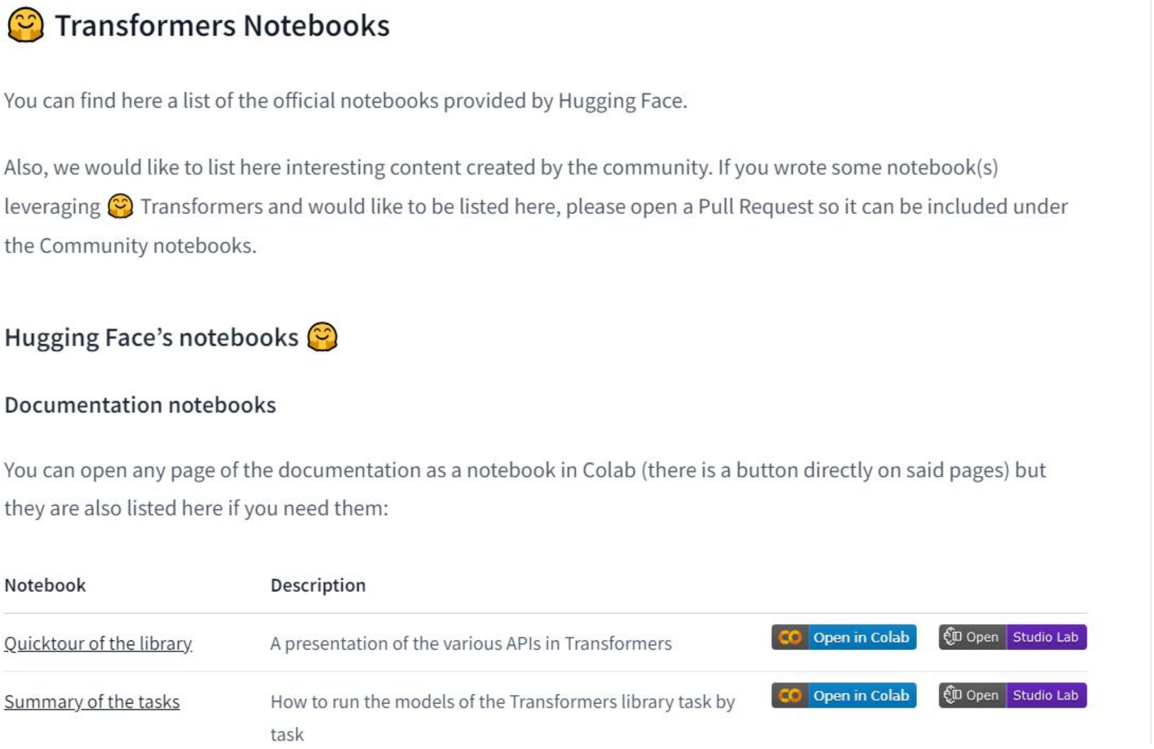
blackforestlabs.ai/announcing-black-forest-labs/

In pixels' dance, AI's craft will rise,
Transforming visions through machine eyes!
From dreams to screens, new worlds unfurled.
AI's brush reshapes our visual world!

Resources

Hugging Face 🤗 Transformers Notebooks

https://huggingface.co/docs/transformers/en/notebooks



🤗 Transformers Notebooks

You can find here a list of the official notebooks provided by Hugging Face.

Also, we would like to list here interesting content created by the community. If you wrote some notebook(s) leveraging 🤗 Transformers and would like to be listed here, please open a Pull Request so it can be included under the Community notebooks.

Hugging Face's notebooks 🤗

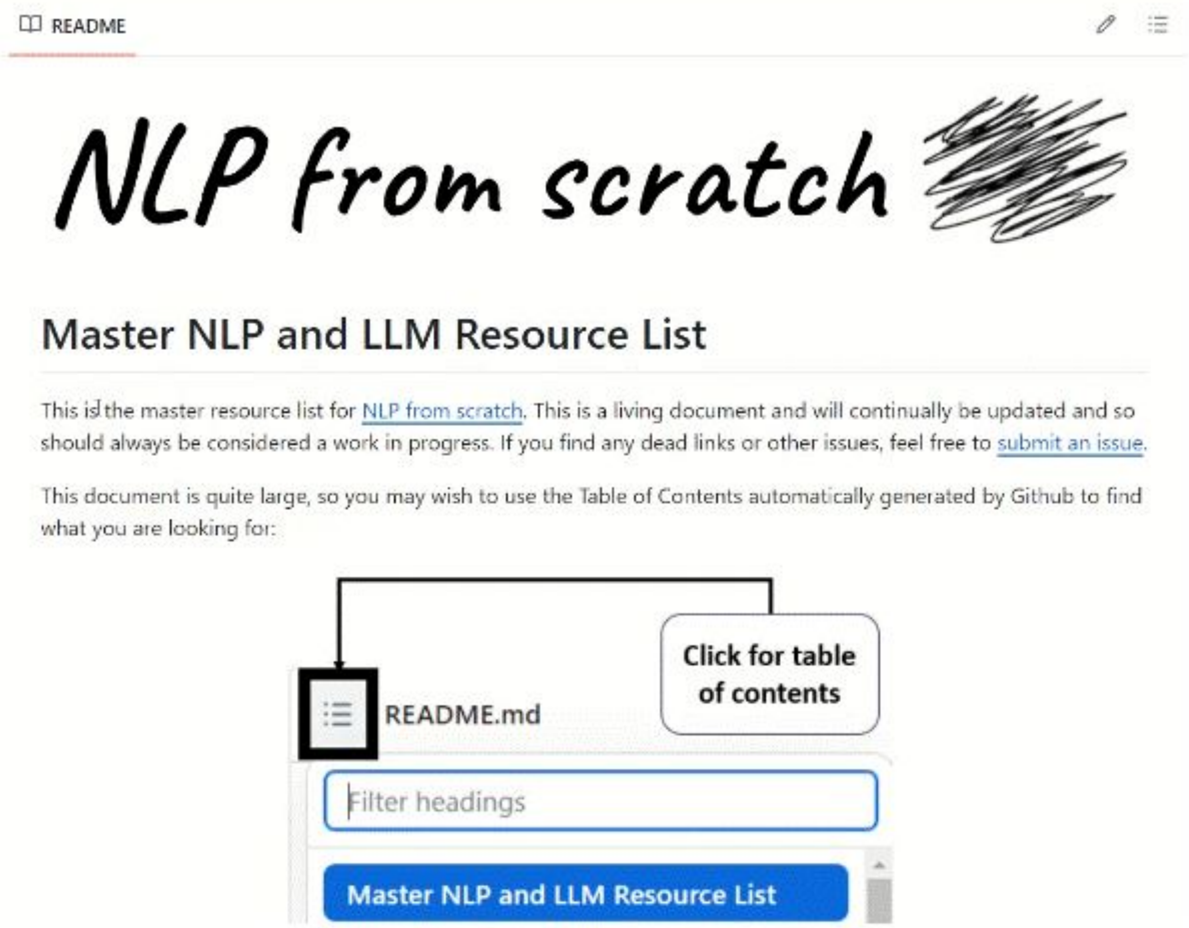
Documentation notebooks

You can open any page of the documentation as a notebook in Colab (there is a button directly on said pages) but they are also listed here if you need them:

Notebook	Description		
Quicktour of the library	A presentation of the various APIs in Transformers	CO Open in Colab	Open Studio Lab
Summary of the tasks	How to run the models of the Transformers library task by task	CO Open in Colab	Open Studio Lab

Master NLP and LLM Resources List

https://github.com/nlpfromscratch/nlp-llms-resources



Master NLP and LLM Resource List

This is the master resource list for NLP from scratch. This is a living document and will continually be updated and so should always be considered a work in progress. If you find any dead links or other issues, feel free to submit an issue.

This document is quite large, so you may wish to use the Table of Contents automatically generated by Github to find what you are looking for:

Thanks!

myles@nlpfromscratch.com

nlpfromscratch.com

linkedin.com/in/mylesharrison