

# 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



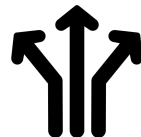
# Manifesto



Knowledge is only valuable if it is useful.



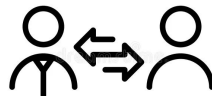
The best way to learn is by doing.



Learning is a non-linear process.



Learning is not a journey, it is guided exploration.



Teaching and learning are complementary.

# Fundamentals

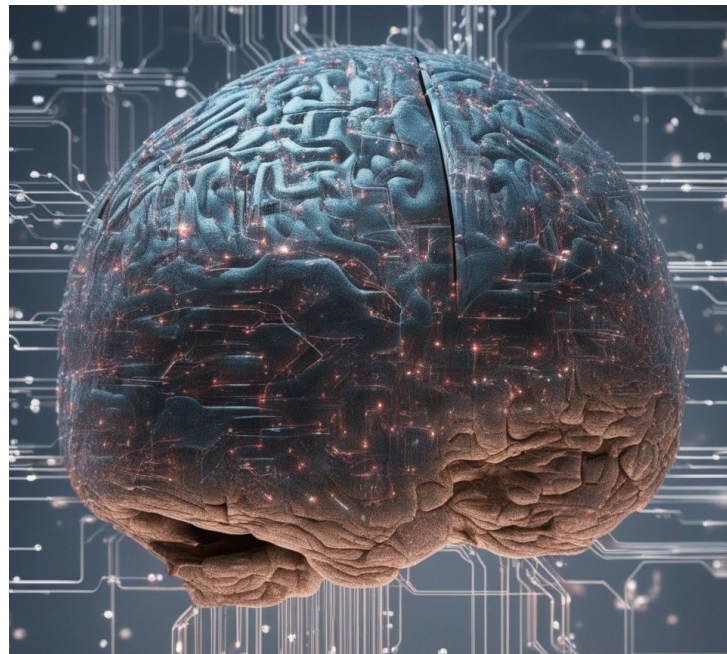


# What is a Large Language Model?

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.

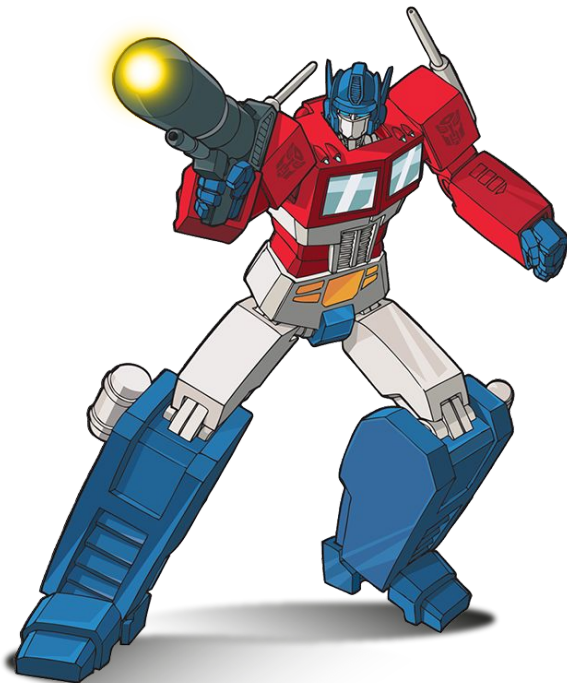
While we're here, ChatGPT is not sentient, nor is it an example of an Artificial General Intelligence (AGI).

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





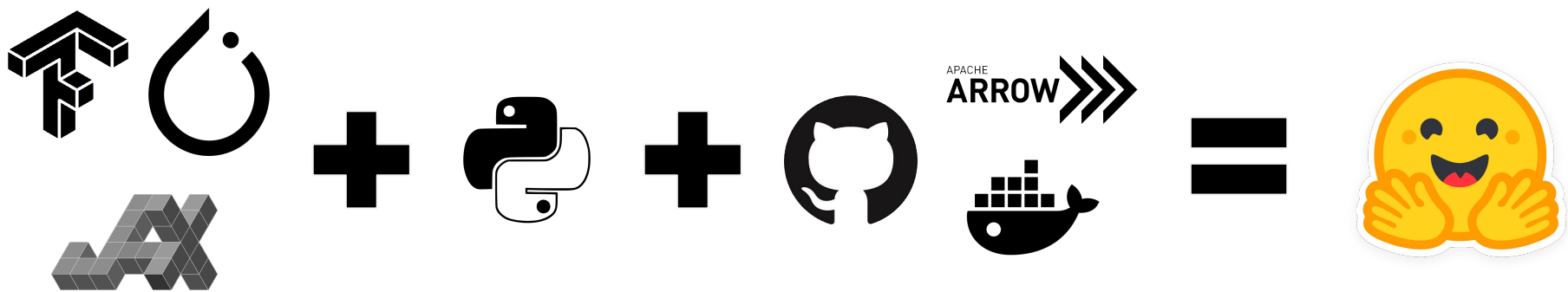
# Hugging Face


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.

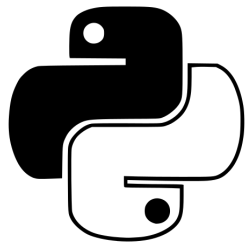
The datasets library allows working with data hosted on the platform, and the transformers library for working with models of this type. There are also other libraries for working with specialized types of models (e.g. diffusers for diffusion models) and data processing and model optimization.





*NLP from scratch* 

# Tools of the Trade



Python 3



Google Colab  
/ Jupyter



**Hugging Face**

Deep Learning and LLMs

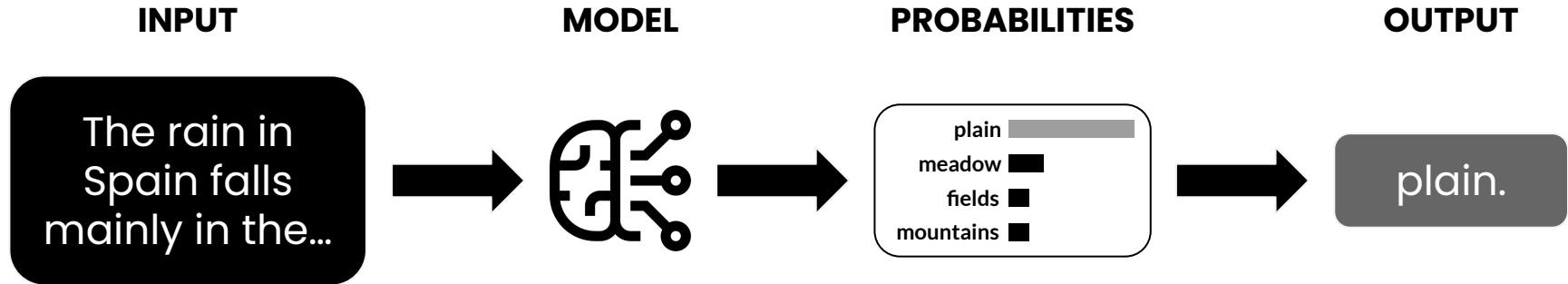
# Generative Text Models

The background of the slide is a dark, intricate network of white lines connecting numerous small, glowing nodes, resembling a complex web or a neural network. Overlaid on this network are various text elements and symbols. In the upper left, a large white rounded rectangle contains the title 'Generative Text Models' in bold black font. Scattered across the network are various alphanumeric strings and words, some of which are highlighted or enclosed in boxes. For example, 'K/51' is prominently displayed in the center-left, and '173356 B01' is visible at the bottom. Other smaller text includes 'REPORTS', 'C65SH', 'AUGUST 10', and '173356 B01'. The overall aesthetic is high-tech and data-driven.

# 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.



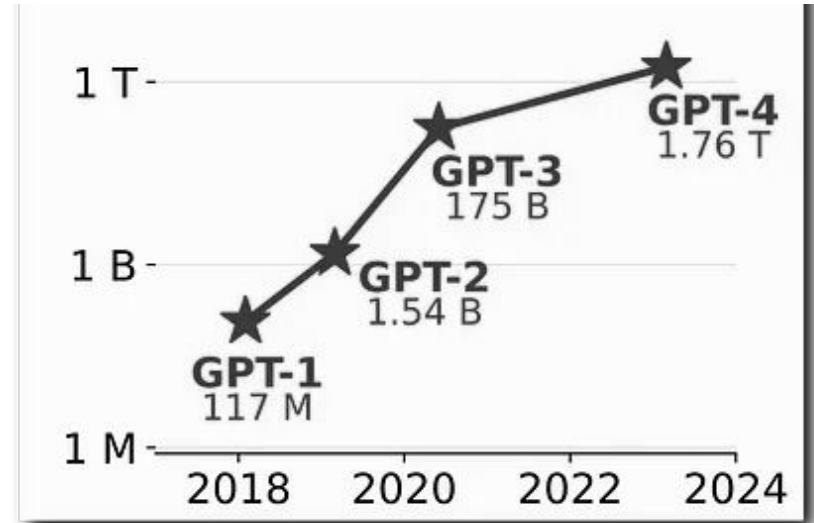
# GPT – The Most Famous LLM

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

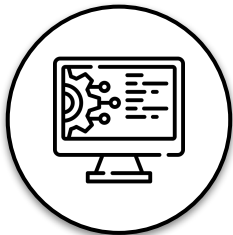
```
conversation = [  
    {"role": "user", "content": "Hello, how  
are you?"},  
    {"role": "assistant", "content": "I'm  
doing great. How can I help you today?"},  
]  
  
Tokenizer = AutoTokenizer.from_pretrained(  
    "microsoft/Phi-3-mini-4k-instruct")  
  
tokenizer.apply_chat_template(conversation,  
    tokenize=False))
```

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

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

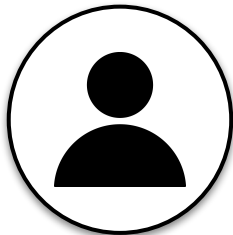
<|endoftext|>

# Message Roles



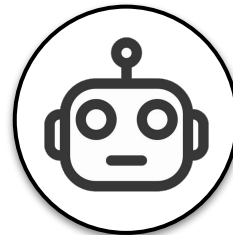
**SYSTEM**

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

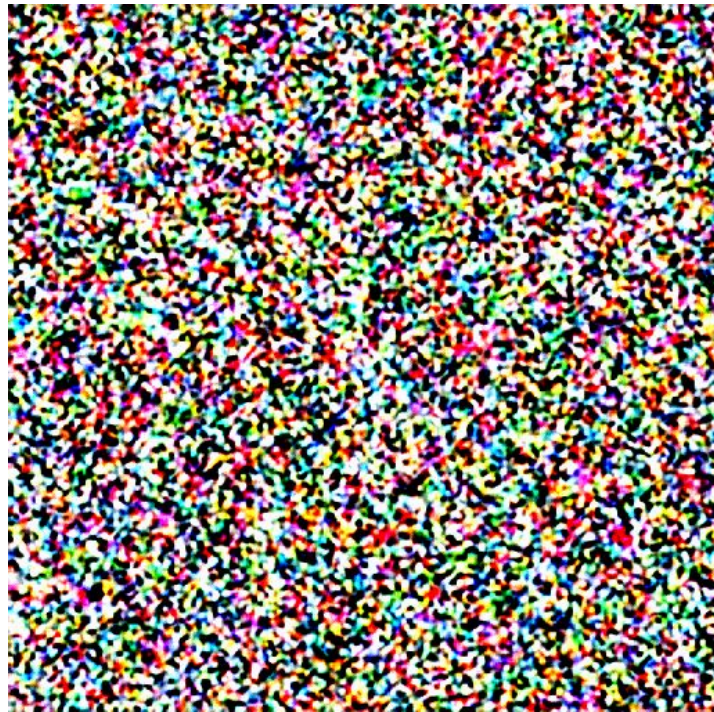




# **Image Generation Models**

# Stable Diffusion

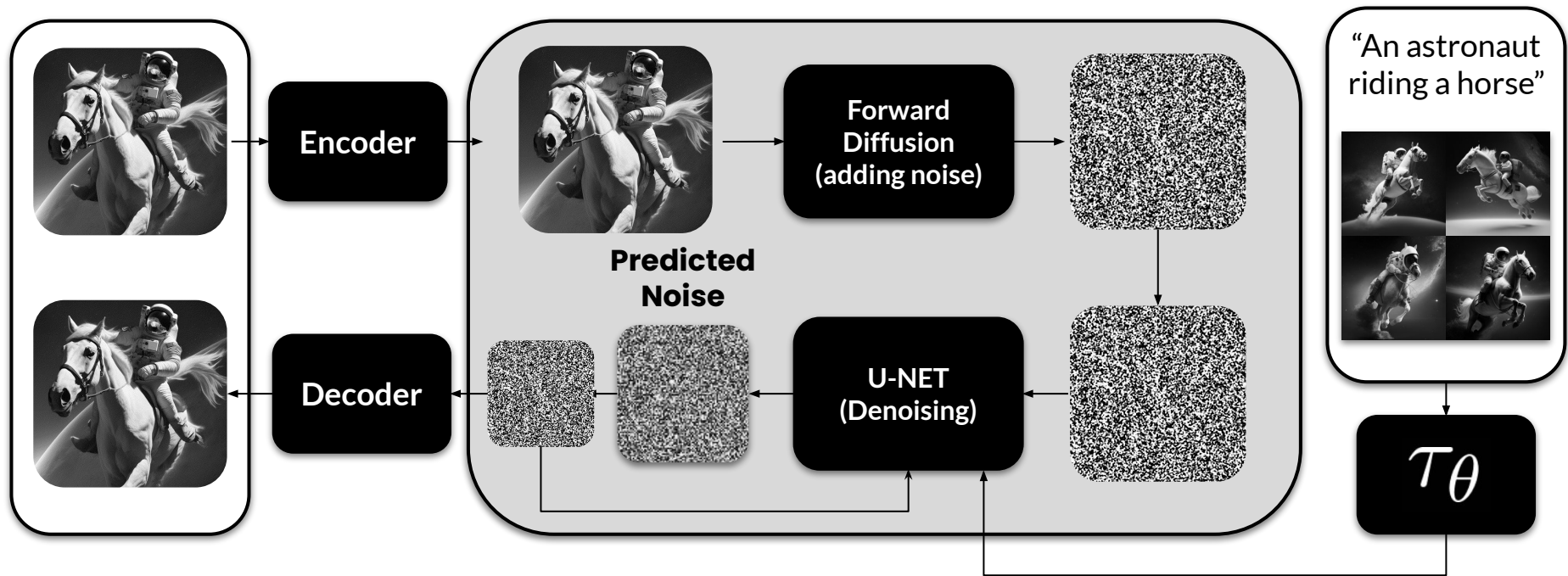
- Latent diffusion models (Rombach et. al, 2021) are a type of generative AI model that can create images by iteratively refining random noise, guided by a learned representation of patterns in data (a “latent space”)
- These models start with random noise and use a neural network to "denoise" step by 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



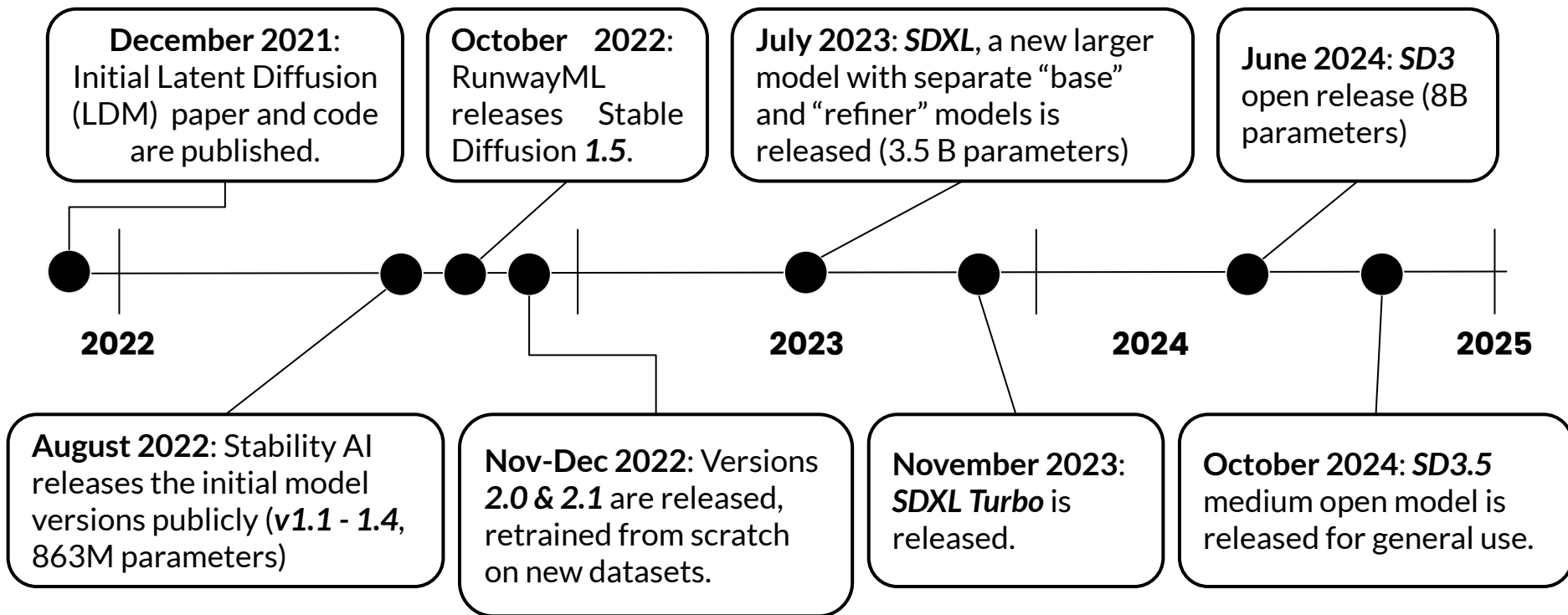
*NLP from scratch*





NLP from scratch 

# Stable Diffusion Timeline



# 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")
```





# 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	
<u><a href="#">Quicktour of the library</a></u>	A presentation of the various APIs in Transformers	 <a href="#">Open in Colab</a>  <a href="#">Open</a>  <a href="#">Studio Lab</a>
<u><a href="#">Summary of the tasks</a></u>	How to run the models of the Transformers library task by task	 <a href="#">Open in Colab</a>  <a href="#">Open</a>  <a href="#">Studio Lab</a>

# Master NLP and LLM Resources List

[https://github.com/  
nlpfromscratch/  
nlp-llms-resources](https://github.com/nlpfromscratch/nlp-llms-resources)

README

# NLP from scratch

## 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:



NLP from scratch

 Thanks!

 [myles@nlpfromscratch.com](mailto:myles@nlpfromscratch.com)

 [nlpfromscratch.com](http://nlpfromscratch.com)

 [linkedin.com/in/mylesharrison](https://www.linkedin.com/in/mylesharrison)