BUILD YOUR OWN AI
HOW TO START?

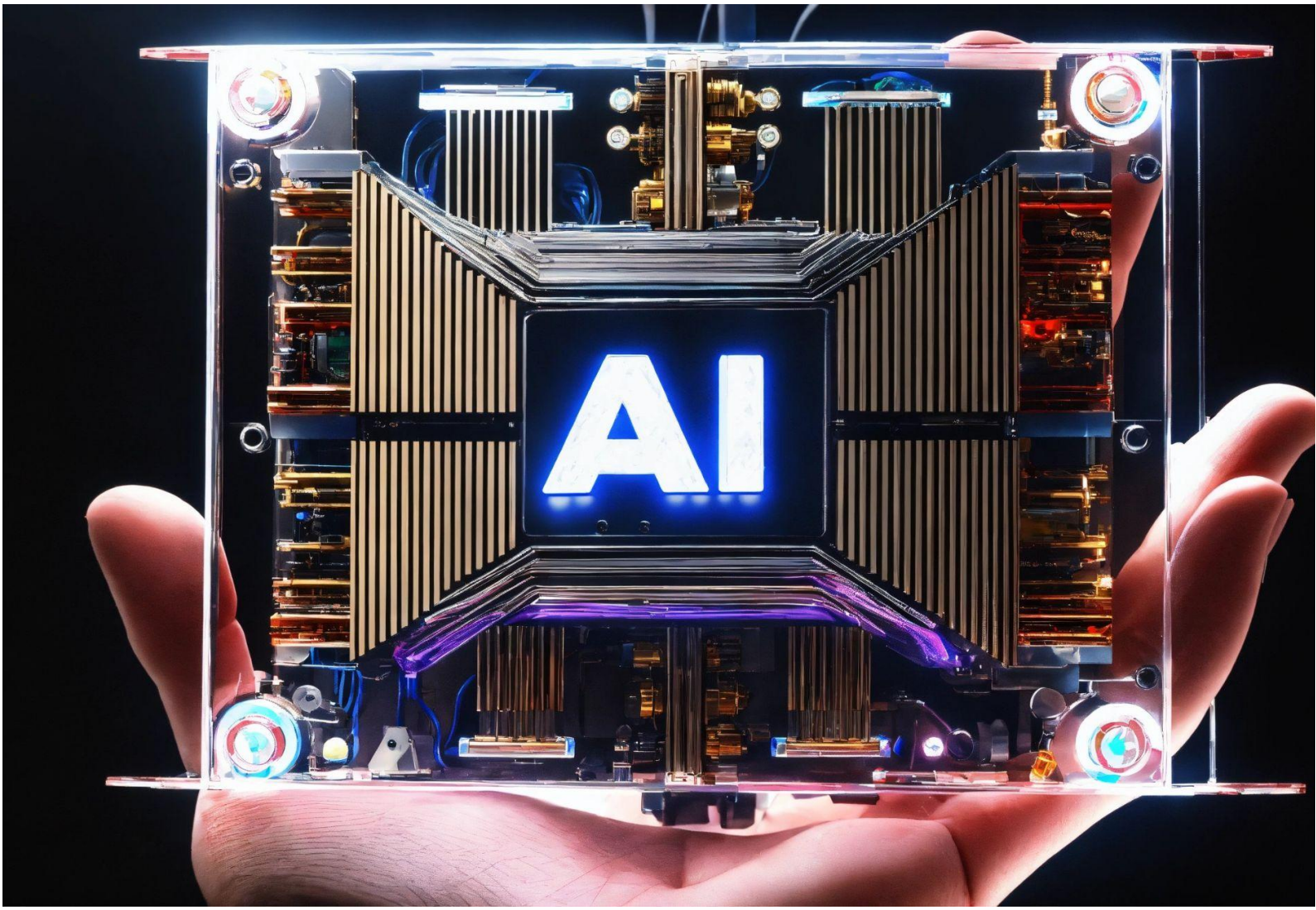
Fabio Matricardi
- AI writer -

Your Guide to Building Your Own AI with Open-Source Tools

Table of Contents

CHAPTER 01

Embarking on the AI Odyssey
Your Guide to Building Your Own AI with Open-Source Tools



CHAPTER 01

Embarking on the AI Odyssey

If you are here is because you may have heard of or even used ChatGPT and you believe that Artificial Intelligence must be free and accessible to everyone. If you are here it means also that you have some expectations: to learn what Artificial Intelligence is and how you can use it for free.

Do you know what is the problem of Artificial intelligence technology? It looks like magic! But it is not a black box: it is a well engineered system that has been designed to give the intended results.
Large Language Models (LLM) looks like a real magic: you write what you want, with your own words (sometimes also with mistakes…) and the Artificial intelligence gives you an amazing reply. It is so amazing that sometimes it goes really over the expectations.

But if I tell you that you can learn how to build your own Artificial Intelligence for FREE and that is not HARD at all? Will you believe me?
Buckle up and get ready to start your journey into Large Language Models. The only things you need are:

- a computer with internet access
- desire to learn
- consistency

Yes that is All!

In the era of technological marvels, artificial intelligence (AI) stands as a transformative force, revolutionizing industries and redefining our world. From self-driving cars to intelligent assistants, AI's impact is undeniable, yet its intricacies often remain shrouded in mystery. This book, "Build Your Own AI - How to Start?", aims to demystify the realm of AI, empowering you to embark on your own AI odyssey.

CHAPTER 02

What is Artificial intelligence?

Artificial intelligence (AI) is the ability of machines to simulate human intelligence. This means that AI can perform tasks that typically require human intelligence, such as learning, reasoning, and problem-solving. AI is a rapidly growing field, and it is already having a major impact on our lives. For example, AI is used in self-driving cars, facial recognition software, and medical diagnosis.

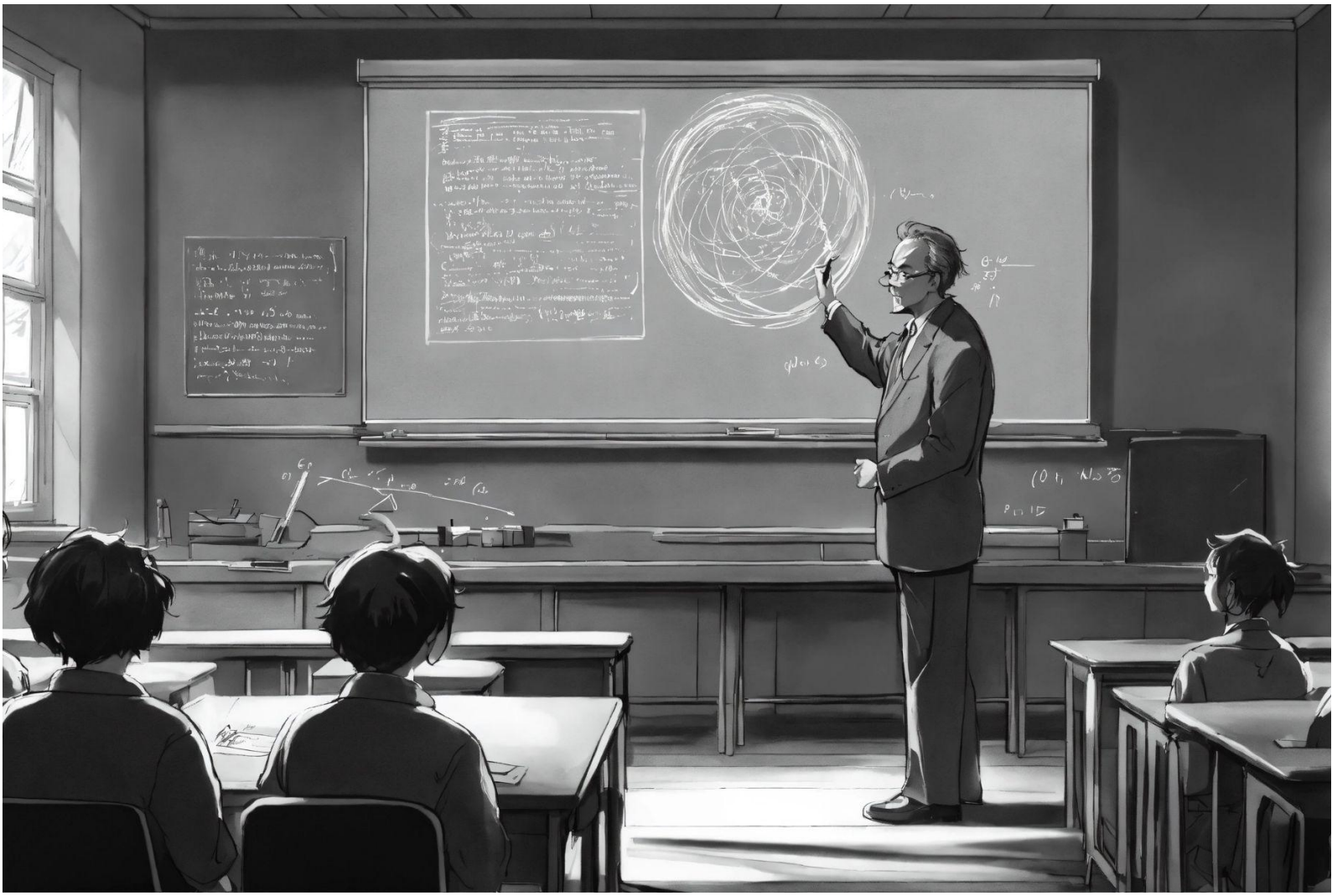
AI, the ability of machines to simulate human intelligence, encompasses a diverse range of techniques and approaches. Machine learning (ML), a subset of AI, enables machines to learn from data, identifying patterns and improving their performance without explicit programming. Natural language processing (NLP), another facet of AI, focuses on the interaction between computers and human language, enabling machines to understand, interpret, and generate human language.

While proprietary AI tools have dominated the landscape, open-source alternatives are rapidly gaining traction. These open-source tools offer accessibility, transparency, and a collaborative spirit, empowering individuals to delve into the world of AI without the constraints of proprietary licenses.

Generative AI, a branch of AI that focuses on creating new content, has emerged as a powerful tool for producing creative text formats, translating languages, and even generating realistic images. ChatGPT, a proprietary generative AI model, has garnered immense popularity for its ability to engage in open-ended, informative conversations.

This book will guide you through the process of harnessing the power of open-source generative AI tools to create your own AI applications. You will learn about the fundamentals of AI, explore the capabilities of open-source generative AI models, and delve into practical tutorials that will equip you with the skills to build your own AI creations.

Whether you're a seasoned programmer, an aspiring AI enthusiast, or simply curious about the potential of AI, this book will serve as your companion on your AI journey. Together, we will explore the vast landscape of AI, unlocking its potential and empowering you to become an AI innovator.



CHAPTER 03

How to get started?

Learning how to use generative AI like chatbots and large language models can seem daunting if you don't have a lot of coding experience. But with the open source Hugging Face Hub and some simple Python libraries, getting started is easier than ever.

Hugging Face Hub

The Hugging Face platform provides free access to many popular AI tools, more than 200.000 models like GPT-3. You can use these models in your own applications without needing special hardware or advanced programming skills. The Hugging Face Hub makes it easy for anyone to start using generative AI by providing a simple interface and documentation on how to get started with different models.

Why we need Open Source Large Language Models?

If you're concerned about your privacy or you run a business that interacts with privacy-concerned users, it may not be a very wise idea to send you or your customer's data to OpenAI or other providers. With Open source LLMs (better if running running on your own computers/server…) you maintain full control over your data.

Be a Rebel… Learn by doing!

If you have watched any tutorial on Youtube or on the web, they all start with taking a good foundational course and then move on…
To begin, you'll want to install the Hugging Face Transformers library in Python. This provides access to many state-of-the-art large language models like GPT-3, PaLM and Claude as well as pretrained model prompting APIs. Once installed, you can start using these powerful AI systems right from your Jupyter Notebook or IDE by just importing the libraries.

Well DON'T do that!

What is the goal?

The goal is to learn how to run an AI and ask your questions, change tasks and have fun with it. The next step is to learn how to improve your troubleshooting skills and choose your own path in Artificial Intelligence.

What are the tools?

1. Python programming basic skills
2. Basic knowledge of Artificial Intelligence
3. A Google account (free)
4. Practice (Know-how)

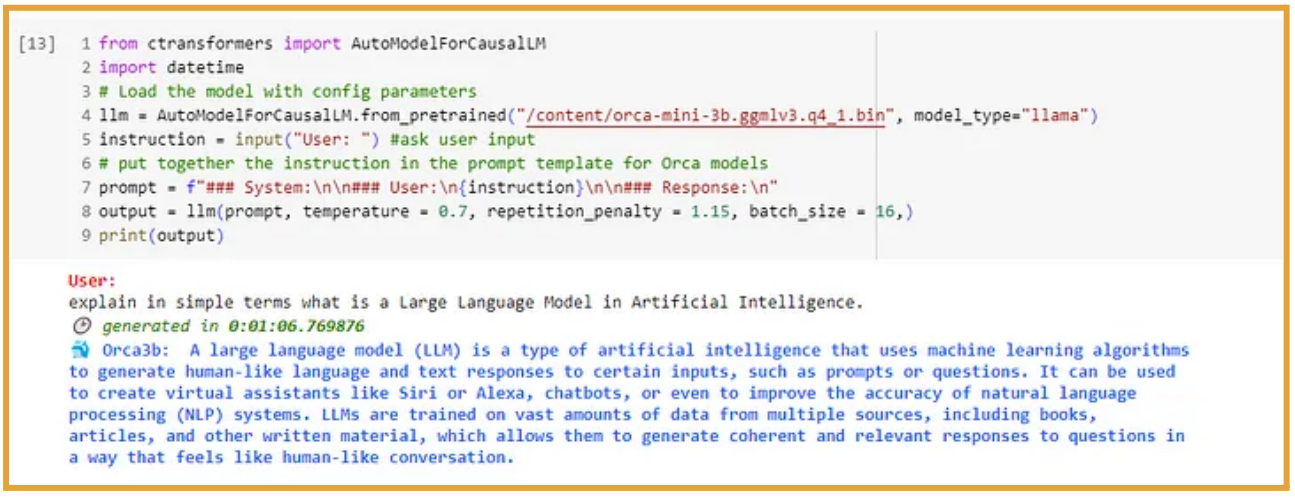
We will start with the practice and introduce during the process the required knowledge to set you free and autonomous in your learning.There are a lot communities and platform that hosts Large Language Models (from now on LLM), many are free, few are not. We will focus on the free ones available.

CHAPTER 04

You can start using an AI in 9 lines of code

Artificial Intelligence (AI) is a branch of computer science that focuses on creating machines and software that can perform tasks that would normally require human intelligence to complete: learning, problem-solving, decision making, perception... Some examples of AI applications include virtual assistants like Siri or Alexa, self-driving cars, facial recognition technology, chatbots, recommendation engines, and more.

We use a programming language, like python, because it allows humans to interact with Language Model. With Python we need only 10 lines of code to create our own basic AI!



```
[13]  1 from ctransformers import AutoModelForCausalLM
      2 import datetime
      3 # Load the model with config parameters
      4 llm = AutoModelForCausalLM.from_pretrained("/content/orca-mini-3b.ggmlv3.q4_1.bin", model_type="llama")
      5 instruction = input("User: ") #ask user input
      6 # put together the instruction in the prompt template for Orca models
      7 prompt = f"### System:\n\n### User:\n{instruction}\n\n### Response:\n"
      8 output = llm(prompt, temperature = 0.7, repetition_penalty = 1.15, batch_size = 16,)
      9 print(output)

User:
explain in simple terms what is a Large Language Model in Artificial Intelligence.
generated in 0:01:06.769876
Orca3b:  A large language model (LLM) is a type of artificial intelligence that uses machine learning algorithms
to generate human-like language and text responses to certain inputs, such as prompts or questions. It can be used
to create virtual assistants like Siri or Alexa, chatbots, or even to improve the accuracy of natural language
processing (NLP) systems. LLMs are trained on vast amounts of data from multiple sources, including books,
articles, and other written material, which allows them to generate coherent and relevant responses to questions in
a way that feels like human-like conversation.
```

an example of what we will do in 10 minutes, 9 lines of code

OK, but what the hell is a Large Language Model?

LLMs are like super smart computer programs that use a special kind of learning to understand, summarize, make new things, and even guess what might happen next. Language is really important because it helps us talk to each other and to machines too. In artificial intelligence, a language model helps computers understand and come up with new ideas.

To do this, the computer program needs a lot of information. It looks at many examples and learns the patterns and connections between words. These connections are like special clues that tell the program which words are most likely to come next and make sense in a sentence.

Python programming Basic Skills

Python is the most used and simple programming language. It has a syntax similar to the English language and has a huge online community with free tutorial, courses and video.

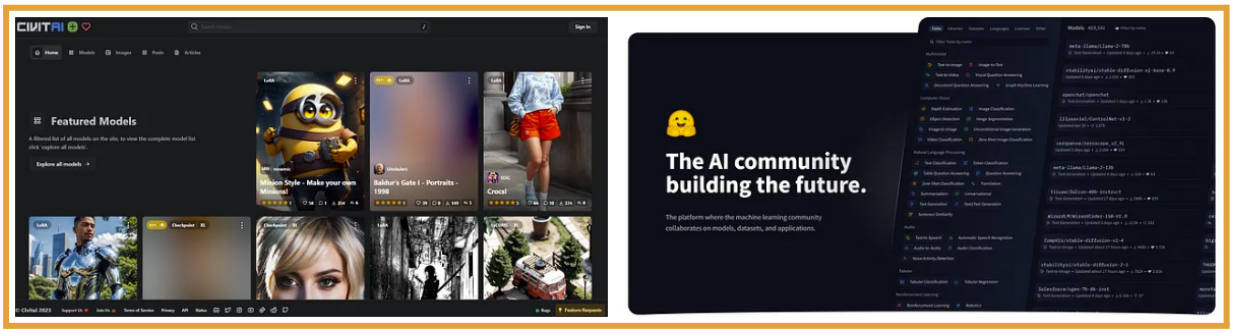
Python has lots of special features and tools that make it easier for us to build and work with large language models and with Artificial Intelligence in general. It is the most used language in artificial intelligence.

We will start with the things we need for our goal: start experimenting yourself changing little things here and there. Feed your curiosity searching on-line 😉.

A Google account (free)

We will learn how to use Large Language Models (from now on LLM) from Hugging Face in a Free Google Colaboratory notebooks. This will be the foundation step to run any free LLM from Hugging Face.

Open Source LLM are mainly hosted on Hugging Face and on CivitAI: the first one is covering many language related tasks (like text generation), the second one is focused on Image generation. Hugging Face is a platform that connects data scientists, researchers, and ML engineers to support open-source projects. It provides tools to build, train and deploy ML models from Open Source code and technologies.



CivitAI and Hugging Face

If you already know everything about Google Colab you can skip the next section and go directly to the libraries.

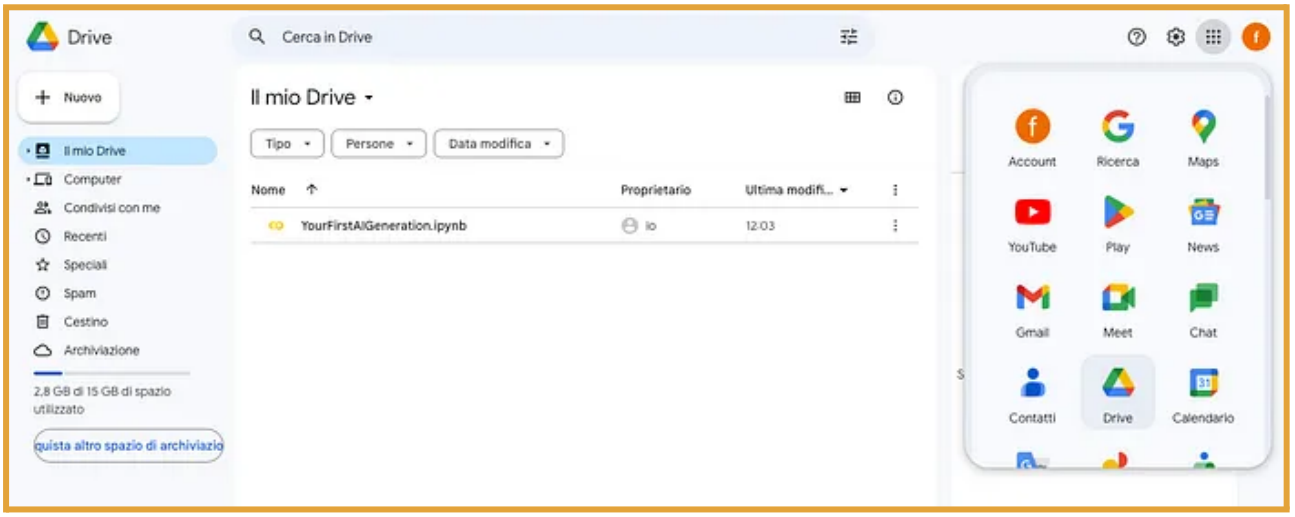
You need a Playground (Google Colab setup)

Learning is hard… but it is also fun. So why don't we create a learning playground by using Google Colaboratory Notebook? It is online and it is 100% free.

Google Colaboratory Notebook; if you have ever heard about Jupyter Notebook they are the same but hosted on Google Servers. If you don't know what they are, well they are an interactive python (and other 40 languages…) environment where you can run the code step by step: this means also that if you make mistakes you don't have to run again your program 😁.

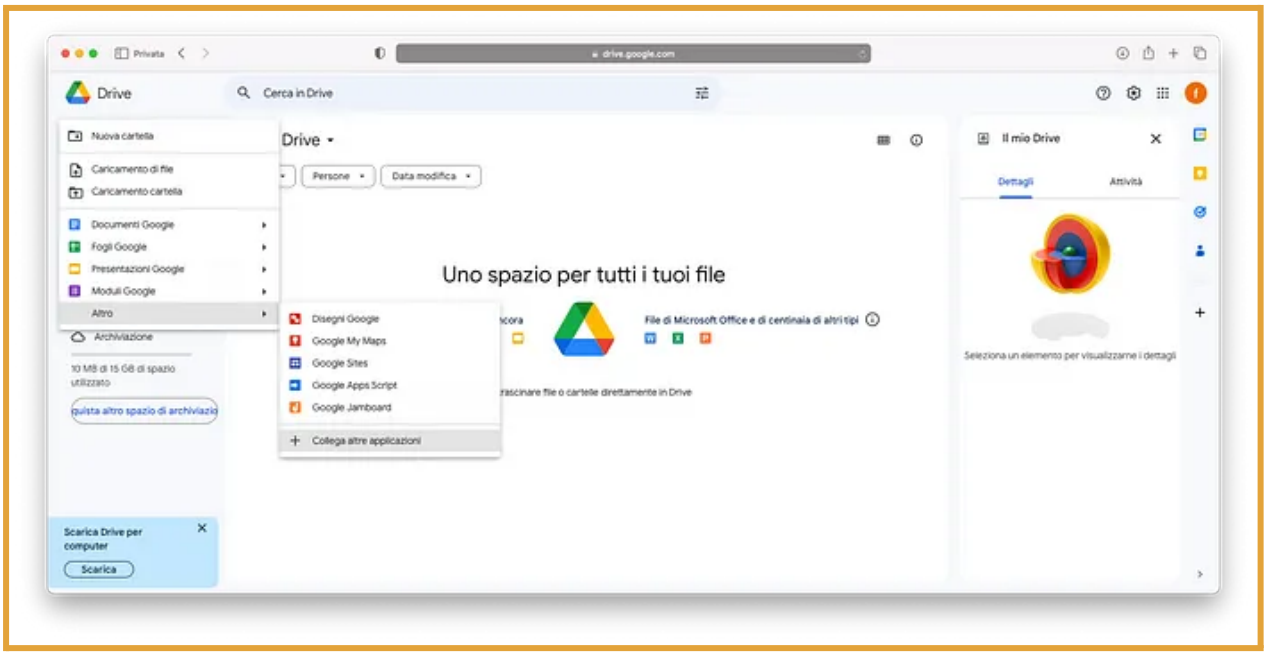
Google Colab notebooks are free! You simply need a Google account, even an existing one. In this case I have an account I never used for Google Colaboratory. To get the feature is really simple.

Go to your google account and open Google drive

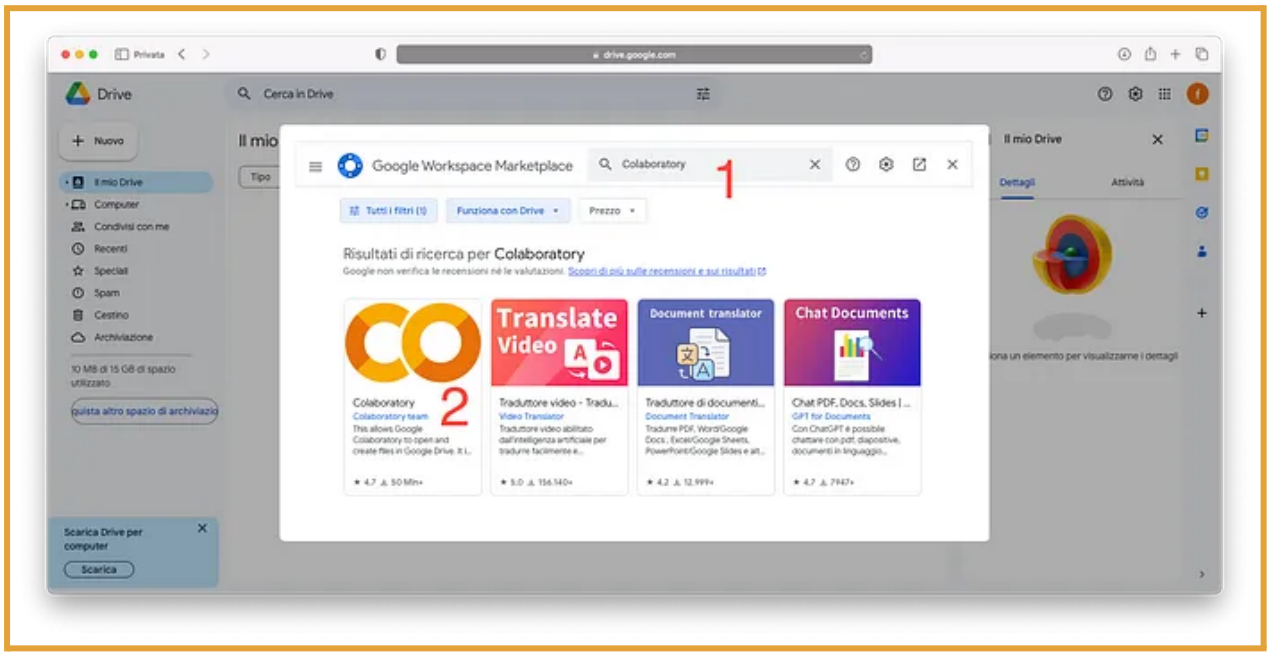


click on the top right to access all the apps from Google connected to your account

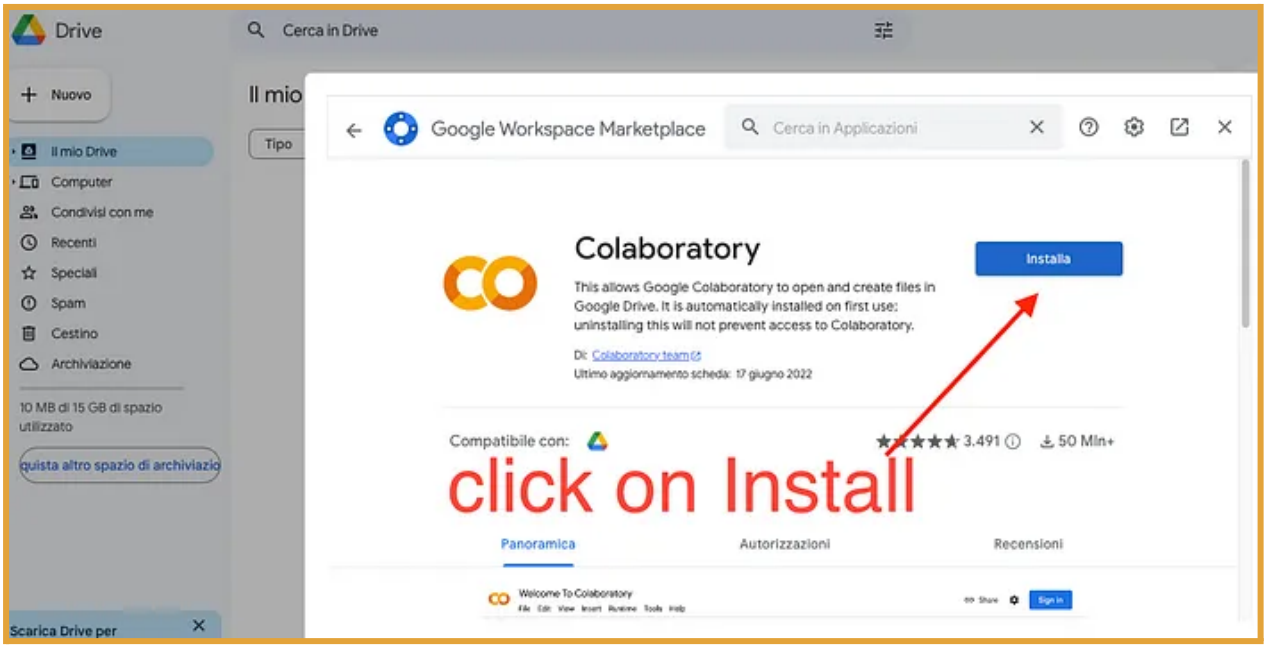
On the top left you find + New button go to other and Connect new apps



On the search bar type “Colaboratory” and click on the card (2)

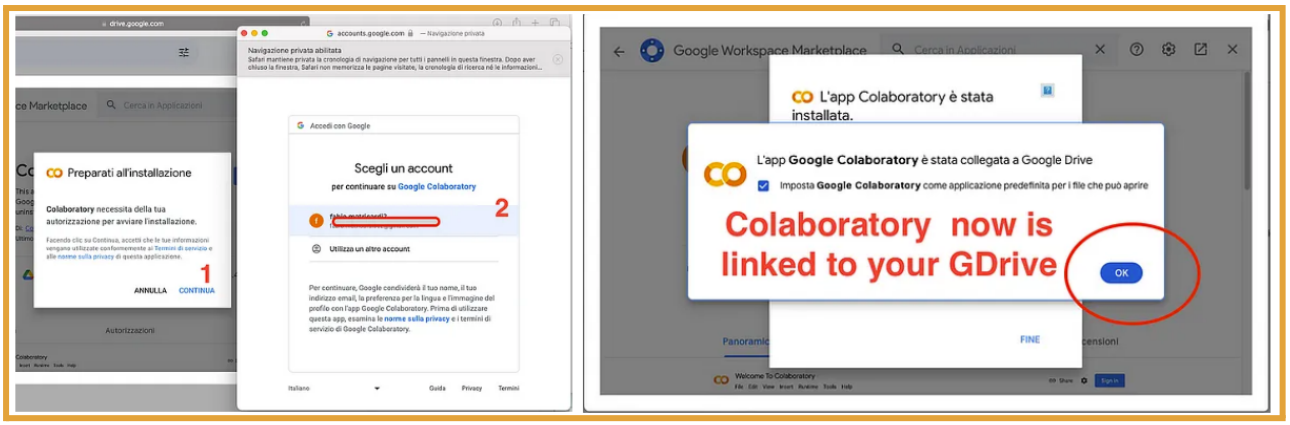


When the complete description page is loaded click on Install



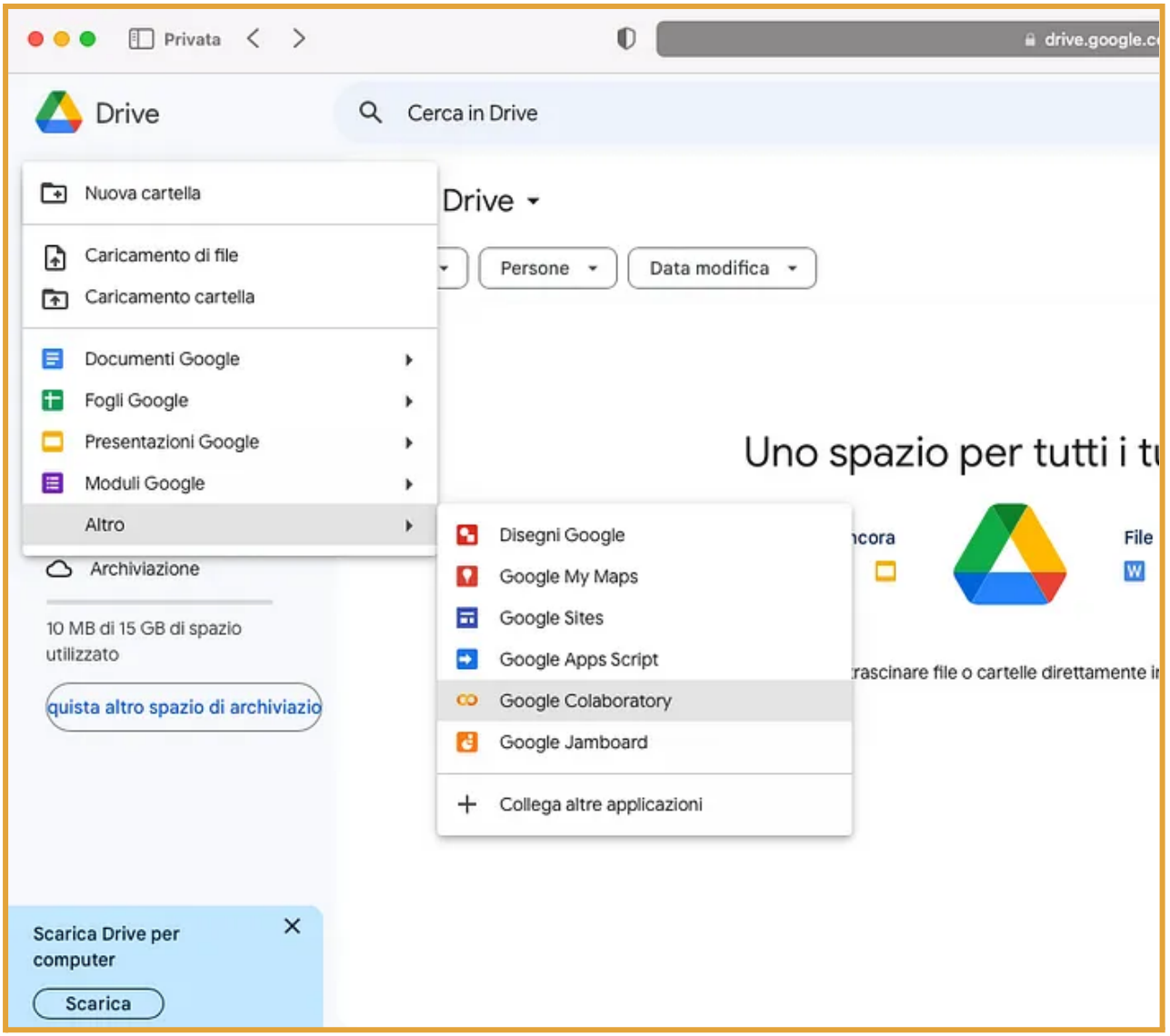
Google Colaboraty extension for Google accounts — free

Google now will ask you to confirm for what account you want to install the feature: click on your email address one and done!



You are ready to go — click on OK

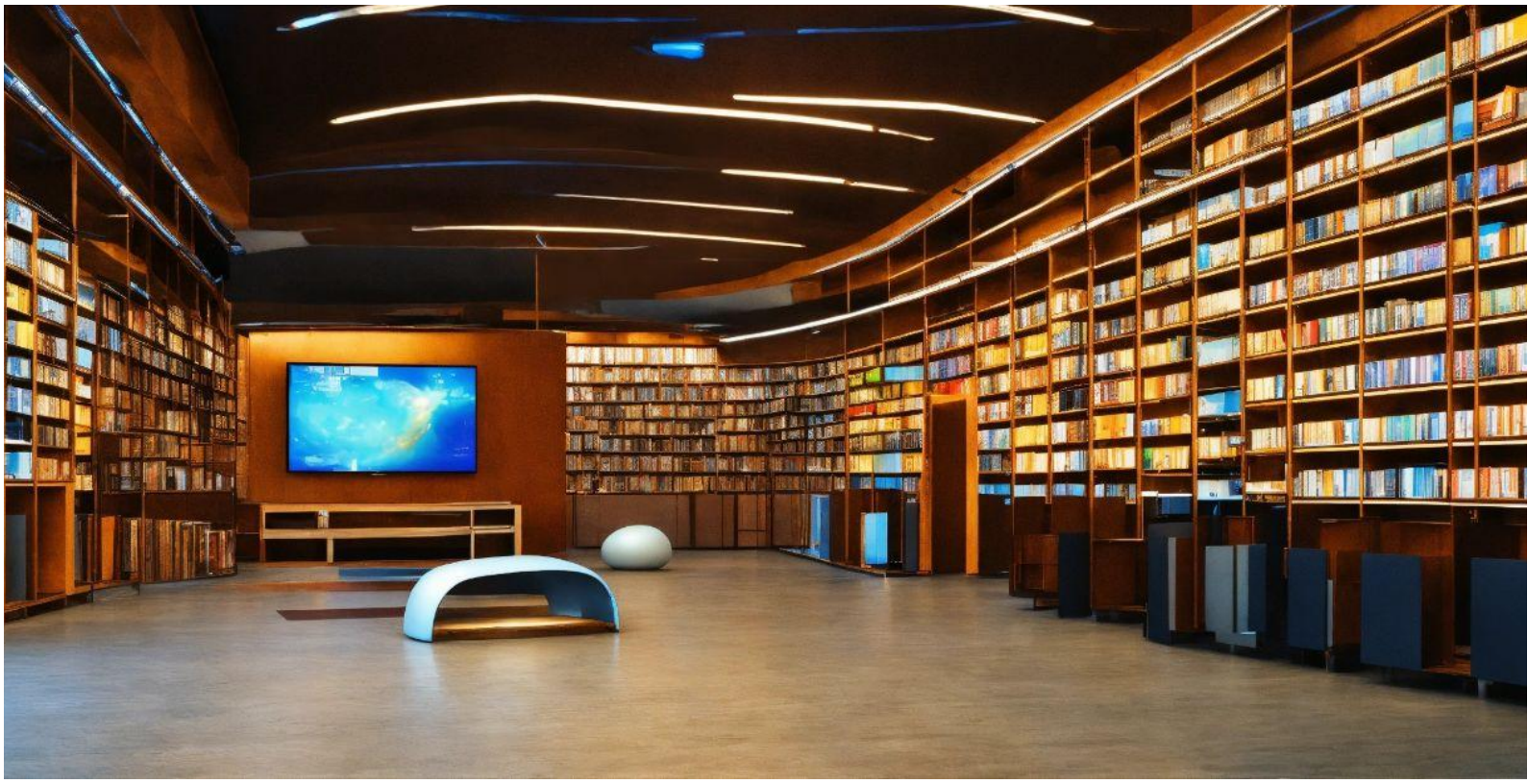
Now you will find the Option to create a New Colaboratory Notebook file from Google Drive



Open a New Google Colab Notebook

CHAPTER 05

You need Libraries

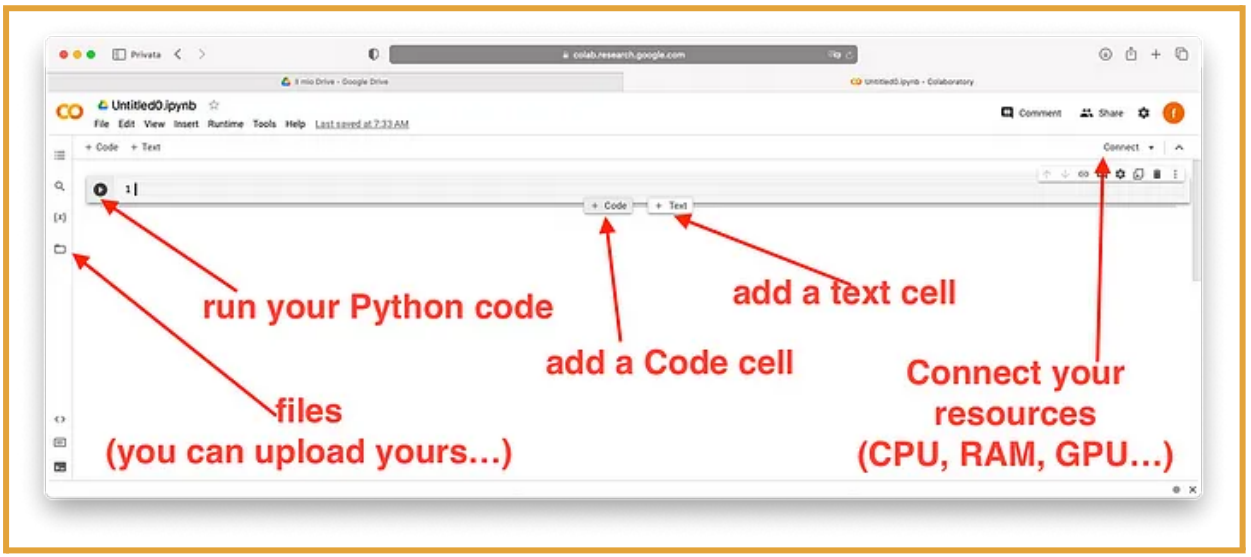


CHAPTER 05

You need libraries (AI transformers)

It is time to code. With Colab is straight forward. The web app is easy to understand.

You can choose to write a Python code or a text (with Markdown features); you can immediately run your code with the play/run button in the cell. You can also upload or download documents



The web application environment

In our scenario we will keep it simple: we need only 2 libraries to be installed. CTransformers library, a very powerful and simple tool to load and run Large Language models in quantized format (small space, big trained parameters…). And then, just for giving a little of good look to the application we will install also the Rich library.

- CTransformers page
- Rich library page

In the first code cell type, we install the CTransformers (RICH is already coming with Google Colab, no need to install) and we download a quantized (reduced) model called Orca-mini in the 3Billion parameters version:

```
%%capture
!pip install ctransformers>=0.2.24
!wget
https://huggingface.co/TheBloke/orca_mini_3B-GGML/resolve/main/orca-mini-3b.ggmlv3.q4_1.bin
```

Note that we put a %%capture to remove the logs during the installations. You can remove it and see the difference. In the second code cell we basically do all the job (ten lines of code, including the comments (in python comments starts with the #):
In order:

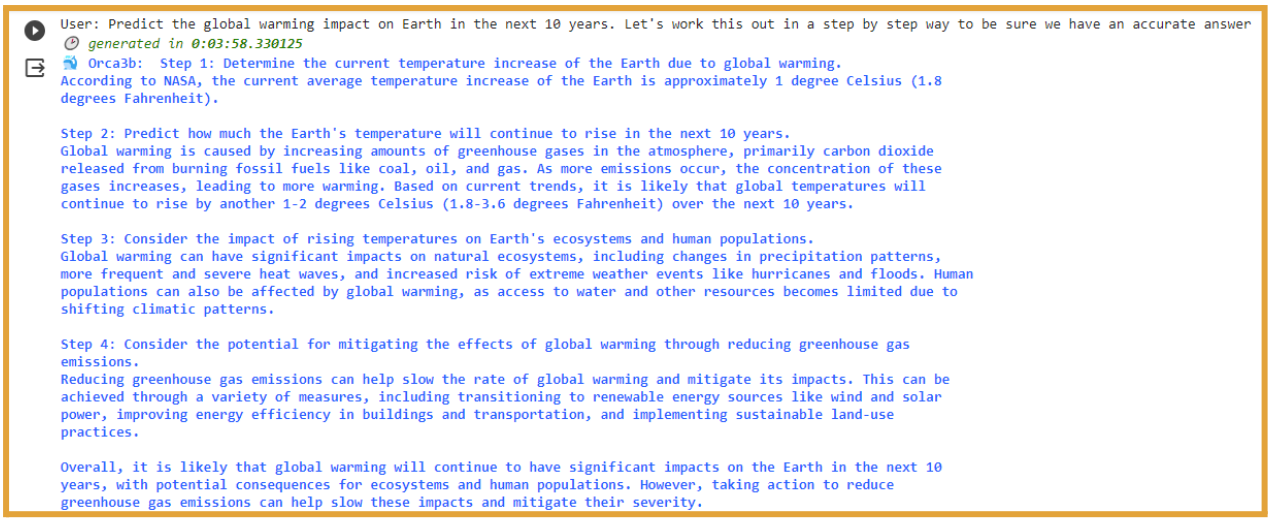
- we load the model in the RAM of Google Colab
- we ask the user to type a question (or an entire prompt, if you want): this is the instruction to the AI
- we merge the instruction with a template (called prompt here): this is crucial because every model has been trained to accept the instructions in a specific way
- finally we put in the variable output the generation and then we print them

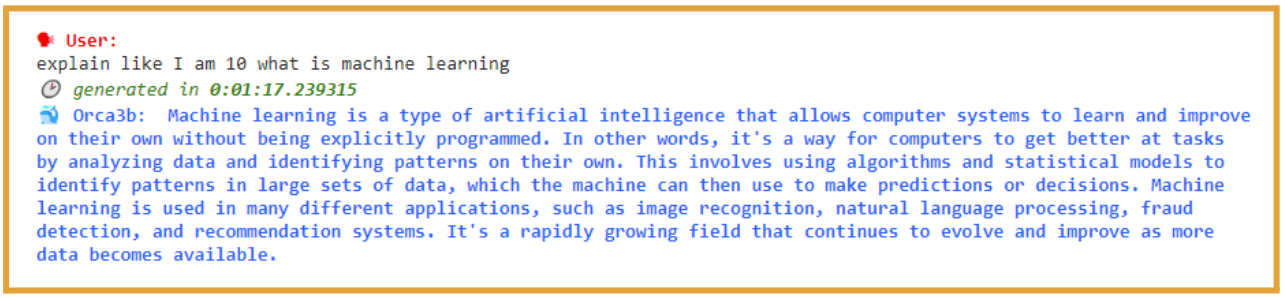
```
from ctransformers import AutoModelForCausalLM
# Load the model in the RAM ready for the generation caall
llm = AutoModelForCausalLM.from_pretrained(
                            "/content/orca-mini-3b.ggmlv3.q4_1.bin",
                            model_type="llama")
instruction = input("User: ")
# Putting together the instruction and the prompt template
prompt = f"### System:\n\n### User:\n{instruction}\n\n### Response:\n"
# Call the text generation
output = llm(prompt,
             temperature = 0.7,
             repetition_penalty = 1.15,
             max_new_tokens = 1024,
             batch_size = 16,)
# printing the output.. well you cannot see it all
print(f"Orca3b: {output}")
```

If you want to continue the test on your new-born AI copy the python code from the line that starts with instruction to the last line into a new cell and run it!😁

```
instruction = input("User: ")
# Putting together the instruction and the prompt template
prompt = f"### System:\n\n### User:\n{instruction}\n\n### Response:\n"
# Call the text generation
output = llm(prompt, temperature = 0.7, repetition_penalty = 1.15,
             max_new_tokens = 1024, batch_size = 16,)
# printing the output.. well you cannot see it all
print(f"Orca3b: {output}")
```

Here an example of the results che can get:





CHAPTER 06

Machine Learning and Generative AI

CHAPTER 06

Machine Learning and Generative AI

Choosing the Right Artificial Intelligence Technology for you to use, means that you know the basics of AI features, tasks and landscape.

Artificial intelligence is transforming industries across the board, from healthcare to finance. But with so many options available, it can be difficult to determine which type of AI will best serve your business needs and goals.

Do we really need a Language Model for every Business application?

The two leading technologies in this space are Generative AI and Machine Learning. Both have their advantages but also some key differences that should factor into any decision about adopting AI for the workplace.
Machine Learning (ML) and Generative AI are two distinct subfields of Artificial Intelligence (AI). Both involve training algorithms using vast amounts of data to make predictions or generate new content, but they differ significantly in their approach and capabilities.

Machine Learning

Machine learning algorithms are trained to recognize patterns within large datasets without generating new content.
Machine Learning involves feeding data into an algorithm to enable it to identify patterns and relationships within the input. These algorithms then are used to extract insights for informed decisions or predictions about future events based on the observed trends.
For instance, ML algorithms can classify images, recognize speech, predict stock prices, and recommend products to users. They typically operate by optimizing a loss function, which measures how well the model performs relative to the desired output. This process allows the algorithm to adjust its parameters until it achieves satisfactory accuracy.

Generative AI

Generative AI systems like ChatGPT, are Generative Adversarial Networks (GANs). We can think of them as “creators” — they take input in the form of text prompts and generate human-like responses based on patterns learned from training data.

Generative AI (like Large Language Models) goes beyond merely identifying patterns; it seeks to generate entirely new content that resembles the original dataset.

In contrast to Machine Learning, which identifies patterns in existing data to make predictions or to guide decisions, Generative AI creates entirely new content that looks similar to the original dataset.

To achieve this, Generative Adversarial Networks (GANs) employ two neural networks: a generator and a discriminator. The generator generates artificial samples, while the discriminator checks whether these samples are genuine or fabricated by comparing them to real examples from the dataset. By engaging in a repeated cycle, the generator enhances its capacity to produce lifelike outputs, while the discriminator develops greater proficiency in discerning between generated and authentic data.

The complex result of this interaction leads to the production of incredibly precise and believable counterfeit data, including images, videos, audio, and even text.

So what is best?

They are both cool! But because of their specifications it is important to pick the right solution that fit your use case.

So what use cases might lend themselves better to generative AI versus machine learning? As a rule of thumb here the 2 main areas:

- Generating unique, human-written text is a strength of systems like ChatGPT — it can be used for things like customer service chatbots or automated marketing campaigns.
- Machine learning excels at analyzing large datasets and detecting patterns in complex information like images, audio files, or video footage. It's ideal for tasks like fraud detection, predictive analytics, and natural language processing.

Let's give them few more thoughts.

Machine Learning use cases

- Predictive maintenance using time series data from sensors in machinery to identify potential issues before they lead to downtime or failures, enabling preventative maintenance.
- Fraud detection by analyzing large datasets of transaction data and identifying anomalies or patterns associated with fraudulent behavior using supervised ML models like random forest classifiers.

- Medical diagnosis: Machine learning can be used to diagnose a variety of medical conditions, such as cancer, heart disease, and diabetes. Machine learning algorithms can be trained on data that includes patient records, medical images, and other relevant information, and then used to identify patterns that are indicative of a particular disease.
- Recommendation systems: Machine learning can be used to develop recommendation systems that suggest products, movies, music, and other items to users based on their past behavior and preferences. Recommendation systems are used by a variety of companies, such as Amazon, Netflix, and Spotify.

Large Language Models use cases

- Chatbot and assistants in your job, in customer service, in consulting.
- Content creation, including text summarization, writing etc. LLMs like GPT-3 are trained to generate human-like text from prompts, making them useful for content generation tasks where creativity is important.
- Text classification to categorize documents, emails or social media posts into predefined categories based on their content by fine-tuning an LLM on the desired category labels during training. This allows for fast and accurate text classification tasks.

- Art and music generation: Generative models can be used to create new works of art and music. Generative models can be trained on data that includes examples of art and music, and then used to generate new works in a similar style.
- Data augmentation: Generative models can be used to augment existing datasets by generating new data points. Data augmentation can be used to improve the performance of machine learning models, as it provides them with more data to train on.

A comparison

LLMs and ML are both powerful AI technologies with the potential to solve a wide range of problems. However, they are not equally suited for all applications.

LLMs are best suited for tasks that require natural language understanding, such as chatbots and machine translation.
ML is best suited for tasks that require pattern recognition, such as image classification and fraud detection.

When choosing between LLMs and ML, it is important to consider the following factors:

- The nature of the task
- The available data
- The desired performance

If you are not sure which technology to use, have a look to the next Chapter.

In addition to the factors discussed above, there are a few other things to keep in mind when choosing between LLMs and ML.

- LLMs are still under development, and their capabilities are constantly improving. ML is a more mature technology, but it is also more complex.
- LLMs are more expensive to train and operate than ML models.
- ML models are more transparent than LLMs, which means that it is easier to understand how they work.

CHAPTER 07

Where to go from here on...

CHAPTER 07

Where to go next?

The main purpose of this book is to point out that AI and GPT like models are not that hard and are not a black box at all.

Maybe you saw in the generation call some things you really do not understand, like

```
# Call the text generation
output = llm(prompt, temperature = 0.7,
             repetition_penalty = 1.15,
             max_new_tokens = 1024,
             batch_size = 16,)
```

Do not worry!
The important discovery here is that now you know that to start is not that hard. Sure, there are a lot of things to be discovered and learnt yet…
Also you may have not yet clear what technology is better for you to start: Machine Learning or Generative AI? Both of them are the present and future of our daily lives, and are changing the world we live in.

If you want to learn more, you can follow me on Medium.

https://medium.com/@fabio.matricardi

I got you covered there! I have plenty of additional guides and articles to help you continue the journey in learning Artificial Intelligence.

See you there soon!