



IA generativa sulla propria intranet

Francesco Alaimo

<https://www.linkedin.com/in/falaimo/>



LLM e chatGPT

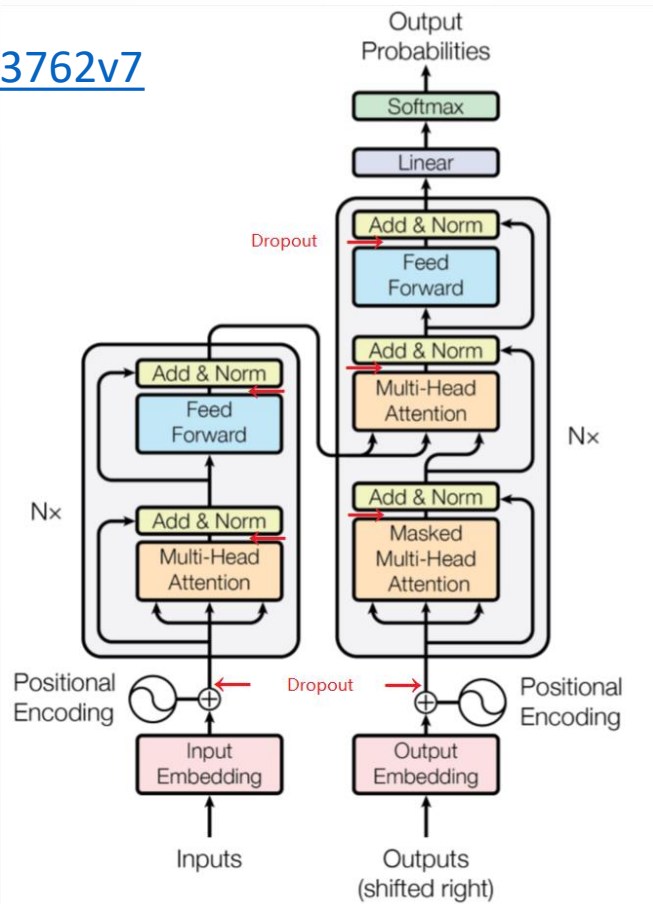
Cosa è un LLM?



- **Large Language Model**; si tratta di un modello generativo
- Nei test per diventare avvocati, negli USA, GPT-4 ha prestazioni fino al 90°percentile rispetto a candidati in carne e ossa (40°percentile nei test di calcolo universitari)
- Recentemente queste tecnologie sono state aggiunte a **Bing** e in altri prodotti **Microsoft** come strumenti di **Copilot**
- Numerosi progetti opensource su Internet rendono disponibili modelli LLM pre-addestrati (es. hugging face)

Attention is All You Need

<https://arxiv.org/pdf/1706.03762v7>



Cosa sono i modelli linguistici (LM)?



- Si tratta di modelli di grandi dimensioni, cioè con un numero elevato di parametri
- Si basano sulle distribuzioni di probabilità di sequenze di parole
- In origine erano usati come componenti di altri sistemi, ad esempio, il riconoscimento automatico della voce
- Dal rilascio del **LLM** GPT-3, nel 2020, i **LM** sono utilizzati autonomamente

LM per IA generativa



- I **LM** possono essere adoperati per generare una storia dato il suo inizio:
 - *Sto andando in un* $\rightarrow P(x_5/x_4, x_3, x_2, x_1)$
 - $P(x_5/x_4, x_3, x_2, x_1) \rightarrow$ negozio
 - *Sto andando in un negozio* $\rightarrow P(x_6/x_5, x_4, x_3, x_2, x_1)$
 - $P(x_6/x_5, x_4, x_3, x_2, x_1) \rightarrow$ oggi

Sto andando in un negozio oggi

Come viene addestrato un LM?



- Si usano diverse tecniche di machine learning:
 - **Supervised learning:** il modello apprende da informazioni associate ad un target
 - **Self-supervised learning:** il modello associa porzioni di una frase a parole scomposte che diventano il target: ogni parola viene associata ad una probabilità che compaia in un certo contesto

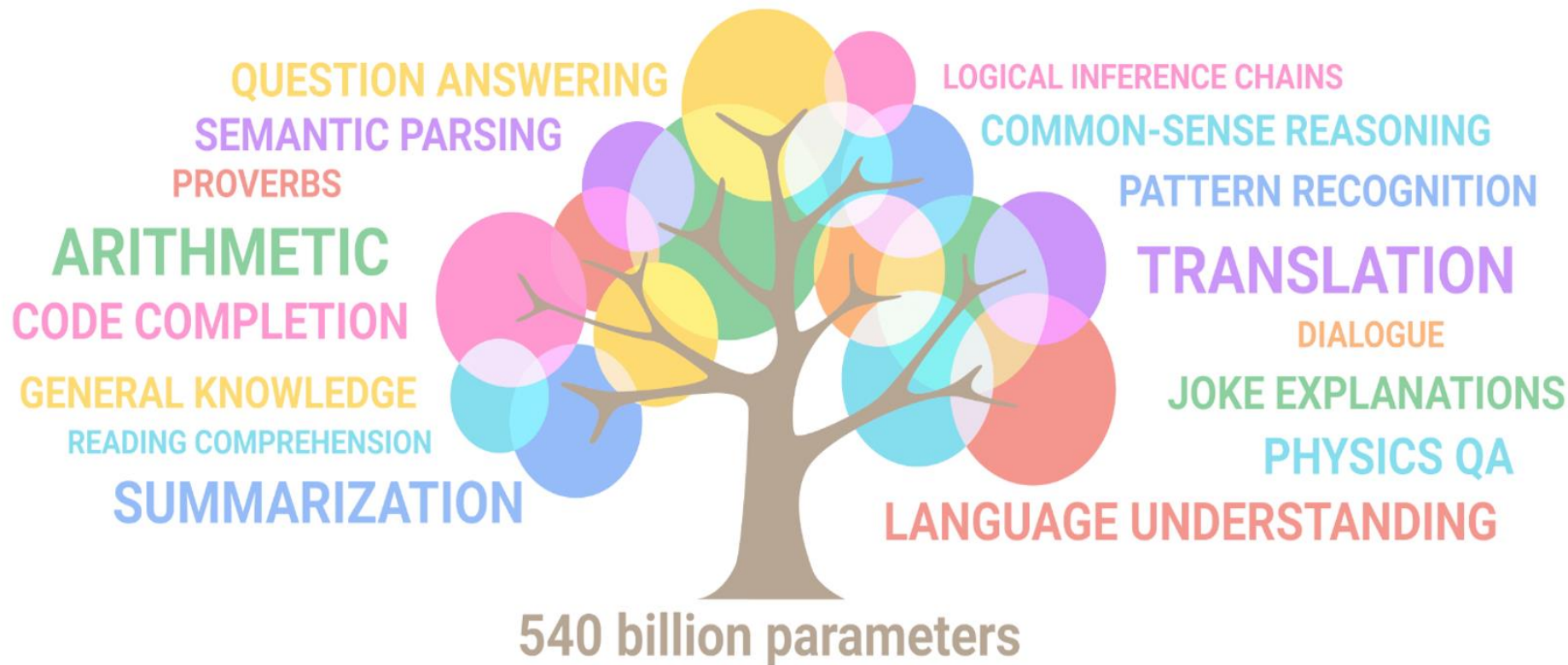
input: Sto andando in un negozio → target: oggi

Altre abilità dei LLM



- Quando i modelli aumentano di dimensioni, acquisiscono abilità sempre più utili, ad esempio, risolvere problemi più complicati
- La dimensione di un modello dipende dal numero di parametri che lo alimenta: $y = p_1 x + c$ è una funzione con un parametro
- **GPT-3** opera con $175 \cdot 10^{12}$ parametri, mentre **GPT-4** con $1 \cdot 10^{18}$
- La risposta di un **LLM** può essere guidata tramite un *prompt*
- Il modello acquisisce nuove *competenze* al raggiungimento di alcune soglie critiche del numero di parametri

Come crescono le abilità di un LLM



[Pathways Language Model \(PaLM\): Scaling to 540 Billion Parameters for Breakthrough Performance – Google Research Blog \(googleblog.com\)](https://research.googleblog.com/2023/07/pathways-language-model-palm-scaling-to-540-billion-parameters-for-breakthrough-performance.html)

- È un **LLM** conversazionale, apparentemente con memoria, che implementa una tecnologia che è il **reinforcement Learning from Human Feedback** (RLHF)
- **RLHF** permette di migliorare il tipo di risposte, introducendo dei filtri e generando output più accurati e coinvolgenti
- Il meccanismo **RLHF** prevede una votazione, da parte di persone, delle risposte ai prompt, che il **LLM** utilizza per selezionare le risposte più votate (e pertinenti)

ChatGPT: pro e contro



Pro:

- ✓ Permette di generare testi, riassumere testi, creare codice, pagine web conformi con SEO
- ✓ È semplice da usare (e dà risposte migliori con un prompt adeguato)

Contro:

- ✗ Può avere allucinazioni
- ✗ È sensibile ai **bias** (aggiunti con RLHF e presenti nelle info acquisite)
- ✗ I filtri si possono aggirare con facilità
- ✗ Non viene considerata la privacy e può violare i copyright



privateGPT

Perché dovrei usare un LLM locale?



- **chatGPT** è stato addestrato con una quantità enorme di informazioni presenti su Internet
- In molti casi non è stato tenuto conto dei copyright, ove presenti, e della privacy associata ai dati
- In ambito aziendale, addestrare **chatGPT** con dati dell'azienda è una pratica che viola le policy
- Un **LLM** locale, addestrato con i propri dati, che rispetti copyright, privacy e policy, può essere un valido strumento in numerosi ambiti

privateGPT



- È un progetto open source di **Iván Martínez** per implementare un **LLM** sul proprio computer, rispettando la privacy
- Funziona con Python dalla ver. 3.10 e utilizza diverse tecnologie (LangCaing, GPT4All, LlamaCpp, Chroma, SentenceTransformers)
- Utilizza un modello **LLM** pre-addestrato selezionabile
- Può essere addestrato con i propri documenti e non richiede collegamento ad Internet
- Multiplatforma Linux/macOS/Windows

<https://github.com/imartinez/privateGPT>

Installazione (richiede Anaconda)



1. `git clone https://github.com/imartinez/privateGPT.git`
2. `cd privateGPT`
3. `conda create -n gpt python`
4. `conda activate gpt`
5. `pip3 install -r requirements.txt`
6. Download LLM `ggml-gpt4all-j-v1.3-groovy.bin`
7. `cp example.env .env`
8. Add documents to the «source_documents» folder

Come funziona?



- **python ingest.py**: effettua il parsing dei documenti e crea un database vettoriale locale dei termini contenuti
- **python privateGPT.py**: permette di interrogare il local LLM con un prompt, avendo come output una risposta coerente con il contenuto dei documenti
- **exit** per uscire dalla modalità query
- Supporta documenti csv, docx, doc, enex, eml, epub, html, md, msg, odt, pdf, pptx, ppt, txt

Come funziona?



Seleziona Anaconda Prompt (anaconda3) - python privateGPT.py

```
(base) C:\Users\falai>d:
```

```
(base) D:\>conda activate gpt
```

```
(gpt) D:\>cd privateGPT
```

```
(gpt) D:\privateGPT>python privateGPT.py
```

```
Found model file at models/ggml-stable-vicuna-13B.q4_2.bin
```

```
llama.cpp: loading model from models/ggml-stable-vicuna-13B.q4_2.bin
```

```
llama_model_load_internal: format      = ggjt v1 (latest)
```

```
llama_model_load_internal: n_vocab     = 32001
```

```
llama_model_load_internal: n_ctx       = 2048
```

```
llama_model_load_internal: n_embd     = 5120
```

```
llama_model_load_internal: n_mult     = 256
```

```
llama_model_load_internal: n_head     = 40
```

```
llama_model_load_internal: n_layer    = 40
```

```
llama_model_load_internal: n_rot      = 128
```

```
llama_model_load_internal: ftype      = 5 (mostly Q4_2)
```

```
llama_model_load_internal: n_ff       = 13824
```

```
llama_model_load_internal: n_parts    = 1
```

```
llama_model_load_internal: model size = 13B
```

```
llama_model_load_internal: ggml ctx size = 73.73 KB
```

```
llama_model_load_internal: mem required = 9807.47 MB (+ 1608.00 MB per state)
```

```
llama_init_from_file: kv self size = 1600.00 MB
```

```
Enter a query: █
```

```
█
```

privateGPT: Pro e contro



Pro:

- ✓ Permette di mantenere la privacy dei documenti utilizzati per l'addestramento
- ✓ Può essere usato senza connessione ad Internet

Contro:

- ✗ Può avere allucinazioni (in particolare con prompt scadenti)
- ✗ Le fasi di *ingest* e di *query* sono molto lente (si usa la CPU; ci sono dei fork per GPU, es. localGPT, **localAI**)
- ✗ Il modello LLM pre-addestrato non è esente da **BIAS**



h2oGPT

- È un progetto che utilizza, in ambiente python, le librerie per il machine learning e l'AI di **h2o.ai**, rilasciate con licenza Apache 2.0
- Utilizza una GUI all'interno di un browser, all'indirizzo <http://localhost:7860> e permette diverse parametrizzazioni
- Si possono caricare i documenti in locale tramite browser
- Supporta le GPU *nvidia*, tramite **CUDA**
- Multiplatforma Linux/macOS/Windows
- Supporta diversi modelli pre-addestrati, selezionabili da GUI

Installazione



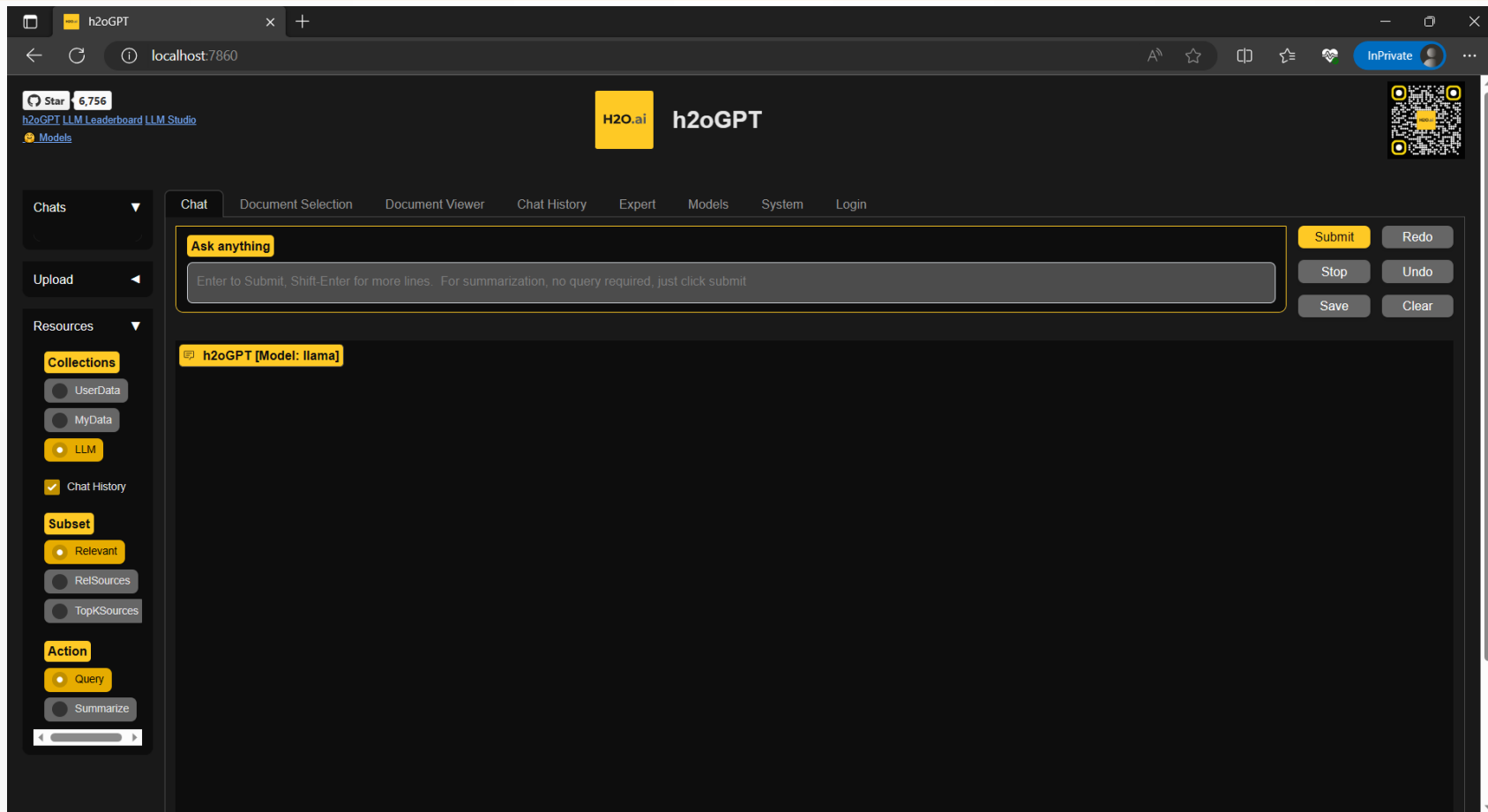
1. `git clone https://github.com/h2oai/h2ogpt.git`
2. `cd h2ogpt`
3. `pip install -r requirements.txt`
4. `pip install -r reqs_optional/requirements_optional_langchain.txt`
5. `pip install -r reqs_optional/requirements_optional_gpt4all.txt`
6. `wget https://huggingface.co/TheBloke/Llama-2-7B-Chat-GGML/resolve/main/llama-2-7b-chat.ggmlv3.q8_0.bin`
7. `python generate.py --base_model='llama' --prompt_type=llama2`

Come funziona?



The screenshot shows the h2oGPT web application running in a browser. The browser's address bar shows 'localhost:7860'. The application has a dark theme. At the top, there's a navigation bar with the 'h2oGPT' logo and a QR code in the top right corner. Below the navigation bar, there's a sidebar on the left with sections: 'Chats' (with a dropdown arrow), 'Upload' (with a dropdown arrow and a 'Drop File Here' area), and 'Resources' (with a dropdown arrow). The main area has a tabbed interface with 'Chat' selected. The 'Chat' tab contains a text input field with the placeholder 'Ask anything' and a 'Submit' button. Below the input field, there's a message from 'h2oGPT [Model: llama]'.

Come funziona?



h2oGPT: Pro e contro



Pro:

- ✓ Permette di mantenere la privacy dei documenti utilizzati per l'addestramento
- ✓ Può essere usato senza connessione ad Internet
- ✓ Numerosi settaggi disponibili
- ✓ Qualità dell'output superiore
- ✓ Supporta CPU e GPU (tramite **CUDA**)

Contro:

- X Può avere allucinazioni (in particolare con prompt scadenti)
- X Se si usa la CPU, la fase di *query* è molto lenta, meno quella di *ingest*
- X Il modello LLM pre-addestrato non è esente da **BIAS**



GPT4ALL

GPT4ALL



- È un progetto open source di Nomic.ai, disponibile su <https://github.com/nomic-ai/gpt4all> che funziona usando la CPU
- Tra le caratteristiche:
 - Risponde a domande generiche
 - Scrive mail, documenti, crea storie, poesie, canzoni
 - Permette di caricare i propri documenti fornendo sommari e risposte inerenti il contenuto
 - Genera codice in diversi linguaggi di programmazione (beta)
 - Memorizza le query e le risposte (non ha memoria delle conversazioni)
 - I modelli LLM sono mantenuti da Nomic e diversi sono utilizzabili liberamente
 - Può funzionare come servizio cui si può accedere con delle chiamate REST

Installazione








1. Scaricare la versione per il proprio S.O. al link <https://gpt4all.io/index.html>
2. Avviare il programma: si presenterà una finestra in stile chatGPT
3. Scaricare il modello di LLM che si preferisce selezionandolo dalla vista **Download**, disponibile a sinistra nella videata principale
4. I documenti da elaborare vanno inseriti nella cartella desumibile da **Application General Setting** → **Plugin** → **LocalDoc Plugin (BETA) Setting**, come collezioni, nella stessa pagine selezionarne l'utilizzo nelle risposte (opzionale)

Come funziona?



GPT4All v2.4.14

ggml-stable-vicuna-13B.q4_2.bin




P

can you create a program in javascript that open a file in txt format and append other lines to the file?

R

Yes, it is possible to write a JavaScript program that opens a text file and appends additional lines. Here's an example of how this could be done using Node.js:

javascript




```
const fs = require('fs'); // Require the 'fs' module for file I/O operations
const path = require('path'); // Require the 'path' module for file system operations

// Define a function to append additional lines to an existing text file
function appendLines(filePath, newLine)
```

☐ Stop generating

1.5 tokens/sec

Send a message...



GPT4ALL: Pro e contro



Pro:

- ✓ Permette di mantenere la privacy dei documenti utilizzati per l'addestramento
- ✓ Può essere usato senza connessione ad Internet
- ✓ Molto semplice da utilizzare e da configurare

Contro:

- ✗ Può avere allucinazioni (in particolare con prompt scadenti)
- ✗ Le fasi di *ingest* e di *query* sono lente, nell'ordine di 4-6 token/s (si usa la CPU; al momento non ci sono fork per GPU)
- ✗ Il modello LLM pre-addestrato non è esente da **BIAS**



Demo di utilizzo

Cosa ne sai Machine Learning?

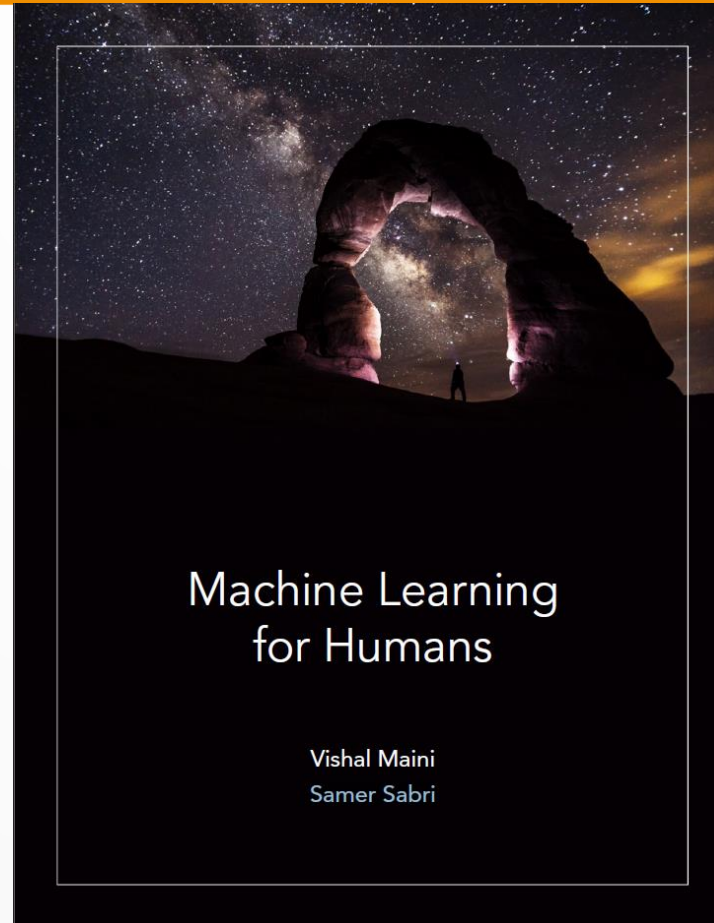


- È stato addestrato un modello **LLM** con il documento «**Machine learning for humans**» di Vishal Maini e Samer Sabri
- Sono stati utilizzati, in sequenza:
 - privateGPT
 - h2oGPT
 - GPT4ALL

È stata posta, ai modelli di LLM, la seguente domanda:

puoi riassumere il libro "machine learning for humans" di Vishal Maini e Samer Sabri, in lingua italiana, usando almeno 20 righe?

Machine Learning for Humans



privateGPT



Anaconda Prompt (anaconda3) - python privateGPT.py

Enter a query: puoi riassumere il libro "machine learning for humans" di Vishal Maini e Samer Sabri, in lingua italiana, usando almeno 20 righe?



Il libro "Machine Learning for Humans" di Vishal Maini e Samer Sabri è un'introduzione alle tecniche di apprendimento automatico. Gli autori spiegano come l'apprendimento automatico può essere utilizzato per risolvere problemi reali, come la classificazione dei dati o il forecasting del futuro. Il libro presenta anche esempi pratici di applicazioni dell'apprendimento automatico in diverse aree, tra cui finanza, marketing e tecnologia. In generale, l'obiettivo del libro è quello di aiutare i lettori a comprendere le basi dell'apprendimento automatico e come utilizzarlo per migliorare la loro vita quotidiana.

> Question:

puoi riassumere il libro "machine learning for humans" di Vishal Maini e Samer Sabri, in lingua italiana, usando almeno 20 righe?

> Answer (took 996.14 s.):

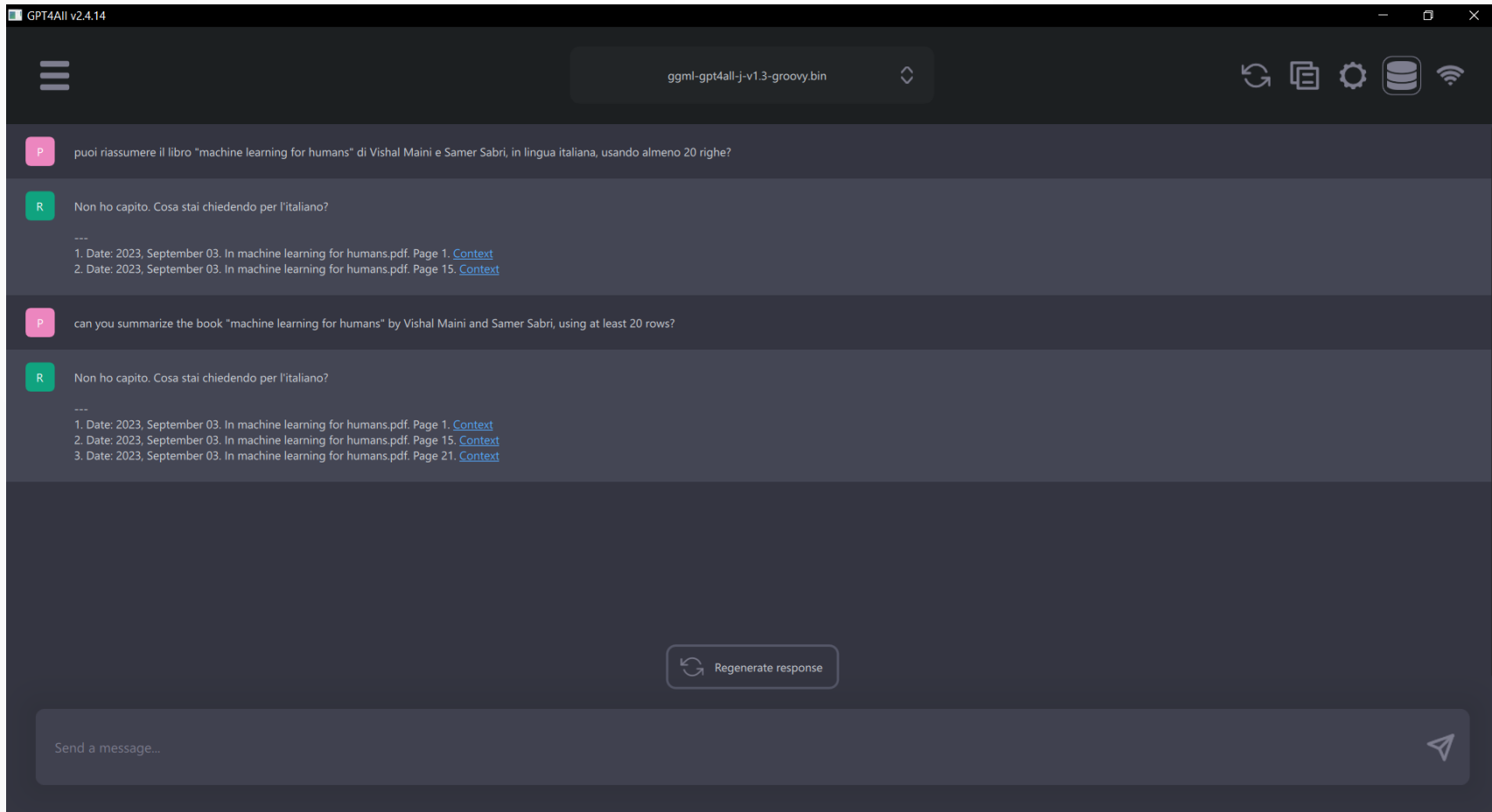
Il libro "Machine Learning for Humans" di Vishal Maini e Samer Sabri è un'introduzione alle tecniche di apprendimento automatico. Gli autori spiegano come l'apprendimento automatico può essere utilizzato per risolvere problemi reali, come la classificazione dei dati o il forecasting del futuro. Il libro presenta anche esempi pratici di applicazioni dell'apprendimento automatico in diverse aree, tra cui finanza, marketing e tecnologia. In generale, l'obiettivo del libro è quello di aiutare i lettori a comprendere le basi dell'apprendimento automatico e come utilizzarlo per migliorare la loro vita quotidiana.

Enter a query:  

h2oGPT



GPT4ALL con LLM 7B



GPT4ALL con LLM 13B



GPT4All v2.4.14

ggml-stable-vicuna-13B.q4_2.bin

P

puoi riassumere il libro "machine learning for humans" di Vishal Maini e Samer Sabri, in lingua italiana, usando almeno 20 righe?

R

Il libro "Machine Learning for Humans" di Vishal Maini e Samer Sabri è un'introduzione alle tecniche di apprendimento automatico utilizzate nei sistemi artificiali. Gli autori spiegano come questi algoritmi possono essere utilizzati per analizzare grandi quantità di dati, creare modelli che possono prevedere il comportamento degli utenti e automare molte attività quotidiane.

Gli autori sottolineano anche l'importanza dell'etica nell'utilizzo delle tecnologie di apprendimento automatico, in particolare quando si tratta di dati personali dei consumatori o di decisioni che possono avere un impatto significativo sulla vita umana.

In sintesi, il libro offre una panoramica completa dell'apprendimento automatico e delle sue applicazioni nella società moderna, mettendo in guardia gli utenti sui rischi associati all'utilizzo di queste tecnologie senza un adeguato rispetto per l'etica.

1. Date: 2023, September 03. In machine learning for humans.pdf. Page 1. [Context](#)

2. Date: 2023, September 03. In machine learning for humans.pdf. Page 15. [Context](#)

P

can you summarize the book "machine learning for humans" by Vishal Maini and Samer Sabri, using at least 20 rows?

R

The book "Machine Learning for Humans" by Vishal Maini and Samer Sabri is an introduction to the techniques used in artificial intelligence systems. The authors explain how these algorithms can be used to analyze large amounts of data, create models that can predict user behavior, and automate many daily activities.

The authors also emphasize the importance of ethics when using AI technologies, particularly with regard to personal consumer data or decisions that can have a significant impact on human life.

In summary, the book provides a comprehensive overview of machine learning and its applications in modern society, warning readers about the risks associated with unethical use of these technologies without adequate respect for ethics.

1. Date: 2023, September 03. In machine learning for humans.pdf. Page 1. [Context](#)

2. Date: 2023, September 03. In machine learning for humans.pdf. Page 15. [Context](#)

3. Date: 2023, September 03. In machine learning for humans.pdf. Page 21. [Context](#)

Regenerate response

Send a message...

GPT4ALL con LLM 13B (bis)



GPT4All v2.4.14

ggml-stable-vicuna-13B.q4_2.bin

P

can you summarize the book "Machine learning for Humans" by Vishal Maini and Samer Sabri using at least 200 words?

R

"Machine Learning for Humans" is a comprehensive guide to understanding machine learning, its applications, and how it can be used in everyday life. The book provides an overview of the basics of machine learning, including supervised and unsupervised learning, as well as reinforcement learning. It also covers various real-world applications of machine learning, such as natural language processing, computer vision, and predictive analytics.

The authors provide practical examples to illustrate how machine learning can be used in different industries, from healthcare to finance. They discuss the challenges that come with implementing machine learning solutions, including data quality, bias, and ethical considerations. The book also covers various tools and frameworks for building machine learning models, such as TensorFlow and PyTorch.

One of the key themes throughout the book is the importance of understanding human behavior when it comes to using machine learning. The authors emphasize that while machines can automate processes and make predictions, they cannot replace human judgment entirely. They discuss how humans can use machine learning to augment their decision-making abilities and improve efficiency in various tasks.

Overall, "Machine Learning for Humans" is a valuable resource for anyone interested in understanding the basics of machine learning and its applications. The book provides practical examples and insights into how machine learning can be used in different industries, as well as the challenges that come with implementing these solutions. It also emphasizes the importance of human judgment when it comes to using machine learning, making it a must-read for anyone looking to leverage this technology in their work or personal life.

1. Date: 2023, September 03. In machine learning for humans.pdf. Page 1. [Context](#)
2. Date: 2023, September 03. In machine learning for humans.pdf. Page 15. [Context](#)
3. Date: 2023, September 03. In machine learning for humans.pdf. Page 81. [Context](#)

Regenerate response

Send a message...

Conclusioni



- È possibile implementare una soluzione **LLM** stand-alone, nel rispetto delle policy, della privacy e del copyright
- Ci sono diverse soluzioni che utilizzano modelli LLM pre-addestrati, da addestrare sui propri documenti
- Al momento le risposte non sono sempre al massimo e dipendono molto dai prompt, oltre che dalle dimensioni del modello usato
- Si spera, in un prossimo futuro, in un loro sviluppo per incrementarne le prestazioni e l'usabilità



<https://www.datasciencefacile.it>