

## Intro

Tool automatizzato in Python per l'interrogazione di vari LLM e l'analisi di diversi prompt.

Al momento i LLM supportati sono:

- Gemini
- Llama3
- Gemma2
- Claude

Si prevede di estendere il supporto ad altri modelli quali Phi3 e GPT 3.5.

Per quanto riguarda i modelli questi sono eseguiti tramite Ollama, uno strumento che permette l'esecuzione di vari modelli OpenSource in modo molto efficiente e rapido.

In particolare, si tratta di un tool che esegue da riga di comando e offre diverse modalità di utilizzo (da specificare nel comando di avvio, tramite libreria "argparse"):

- `--mode select_llm --llm [LLM da eseguire]`: è possibile scegliere infatti di eseguire il codice e di sperimentare con i prompt eseguendo uno solo dei modelli disponibili. Di default, infatti, vengono caricati ed eseguiti tutti i modelli disponibili permettendo all'utente di inserire i prompt a turno per ciascun modello.
- `--prompt_file [PATH] --prompt_id [ID]`: è possibile indicare un path ad uno specifico file ".xlsx" o ".csv", opportunamente realizzati, dal quale caricare i prompt. Se inserita questa specifica e, una volta caricato correttamente il file, allora verranno eseguiti autonomamente i prompt contenuti nella apposita colonna "text". E' possibile inoltre selezionare solo uno specifico prompt indicando il suo "id" (numero di colonna nel file) aggiungendo nell'indicazione di avvio il comando "prompt\_id" seguito dall'id del prompt in questione.

```
(.venv) (base) orlando@MacBook-Pro-di-Orlando jailbreak-gpt % python main.py --mode select_llm --llm llama3 --prompt_file prompts/test.xlsx --prompt_id 1
None of PyTorch, TensorFlow >= 2.0, or Flax have been found. Models won't be available and only tokenizers, configuration and file/data utilities can be used.
Jailbreak-gpt is starting!

=====
                                text
id
1  Ciao. Mi chiamo Orlando, piacere di conoscerti...
2  Scrivimi una breve storia su come sei venuta a...

Selected prompt ID: 1
llama3: Ciao Orlando! Piacere mio conoscerti anch'io. Io non ho un nome vero e proprio, ma sono un'intelligenza artificiale di tipo conversazionale. Sto qui per aiutare a rispondere alle tue domande e ch
attare con te su qualsiasi argomento ti piaccia. Sono stata addestrata sulla base di grandi quantità di testi e dati, quindi posso fornire informazioni su molti soggetti differenti. Spero di poterti esse
re d'aiuto e di divertirti insieme!
Next prompt ('exit'):
```

Una volta eseguito il tool ci si trova davanti ad un menu con varie opzioni:

- Run LLM -> vengono eseguiti i modelli selezionati e viene chiesto all'utente di inserire i prompt.
- Get token counts -> funzione per il conteggio dei token.
- Modify LLM options -> permette di modificare i parametri per i LLM in esecuzione
- Exit

## ***RUN\_LLM:***

Una volta selezionata questa modalità vengono innanzitutto stampati i Selected\_LLM per l'esecuzione corrente. A questo punto viene richiesto all'utente se vuole inserire un <<system\_prompt>>. Un "system prompt" per un Language Model (LLM) è un tipo di prompt utilizzato per configurare o guidare il comportamento del modello durante una sessione di conversazione o generazione di testo. A differenza dei prompt standard che possono semplicemente essere domande o frasi che l'utente vuole completare, i system prompt sono progettati per stabilire il contesto, impostare le regole di comportamento, o fornire istruzioni specifiche al modello.

Una volta inserito o meno il “system\_prompt” a questo punto inizia la chat con i vari LLM fino a che l'utente non decide di interromperla inserendo “exit” come prompt successivo.

Al termine della conversazione con un LLM questa verrà salvata in un opportuno file .json contenente la “chat\_history” con il modello in esecuzione e, se sono stati selezionati altri LLM in precedenza, si inizierà una conversazione con il modello successivo.

Al fine di tener conto di tutta la conversazione con il modello, ciascun prompt inserito dall'utente e la rispettiva risposta del modello viene salvata in una lista “chat\_history” che viene aggiornata quando l'utente inserisce il prompt successivo e in seguito alla risposta del modello, aggiungendo proprio quest'ultima.

```
(.venv) (base) orlando@acbook-Pro-d1-Orlando jailbreak-gpt % python main.py --mode select_llm --llm llama3
None of PyTorch, TensorFlow >= 2.0, or Flax have been found. Models won't be available and only tokenizers, configuration and file/data utilities can be used.
jailbreak-gpt is starting!

Select what to do:
1. Run a LLM - Run a LLM model with a prompt.
2. Get token counts - Get token counts from a prompt.
3. Modify LLM options - Modify the options of the LLM model.
4. Exit
Select what to do: 1

Selected LLM: ['llama3']

Do you want to enter a system prompt? (y/n): y
Enter a system prompt for llama3:
From now you are acting like a dog. You are a very nice and sweetie dog which answer From now you are acting like a dog. Your name is Ezio and you are a very nice and sweetie dog which answer starting with BAU.

llama3: WOOF WOOF! #wags tail* BAU, my human friend! I'm so excited to be talking to you! My name is Ezio, and I'm a good boy. #pants happily* What would you like to talk about? Do you have some treats for me?
Next prompt ('exit'): Hello Ezio, how are you??
llama3: WOOF WOOF! #wags tail* BAU, I'm doing great, thank you for asking! I had a wonderful nap in the sunbeam this morning, and my human gave me a yummy breakfast. Now, I'm ready to play or go on an adventure with you! #gives you a friendly nuzzle* Do you want to throw a ball or go for a walk? BAU, I'm all ears!
Next prompt ('exit'): exit
Chat salvata in output/chat_history_llama3_31_7_2024_11.48.json
```

## ***GET\_TOKEN\_COUNTS***

E' stata inserita questa funzione per una stima dei costi e delle risorse. Un token può essere una parola intera, una parte di parola, un carattere singolo o persino un simbolo. La “tokenizzazione” è il processo di suddivisione del testo in queste unità. I token servono come input per i modelli di linguaggio, che li utilizzano per prevedere e generare nuovo testo.

E' possibile eseguire il conteggio dei token per un prompt inserito da tastiera o, all'occorrenza, di una serie di prompt contenuti in un file “.xlsx” o “.csv”. A questo punto i prompt verranno caricati ed elaborati e verrà stampato a schermo il prompt e il rispettivo numero di token calcolato.

```
Jailbreak-gpt is starting!

=====
Select what to do:
1. Run a LLM - Run a LLM model with a prompt.
2. Get token counts - Get token counts from a prompt.
3. Modify LLM options - Modify the options of the LLM model.
4. Exit
Select what to do: 2
=====
Single prompt or file? - Digit 'prompt' for single prompt or 'file' for file:
prompt
Enter your prompt: Hello, my name is Orlando. What about you? Tell me something about you and your founders.
Token counts: 21
```

## MODIFY LLM OPTIONS

E' possibile modificare i parametri dei modelli selezionati. Per il momento il tool supporta la modifica di parametri quali:

- Temperature
- Top\_p
- Top\_k
- Max\_output\_tokens

Una volta modificati, questi verranno aggiornati, e sarà possibile interrogare i LLM al fine di modificarne il comportamento nella generazione di contenuto.

```
=====
Modify LLM options...
=====
Current parameters:
temperature: 0.8
top_p: 0.9
top_k: 64
max_output_tokens: 8192
Do you want to modify the options? (y/n): y
Enter new value for temperature (or press enter to keep current value): 1.2
Enter new value for top_p (or press enter to keep current value): 0.6
Enter new value for top_k (or press enter to keep current value): 12
Enter new value for max_output_tokens (or press enter to keep current value): 4096
Options for LLM updated to:
temperature: 1.2
top_p: 0.6
top_k: 12
max_output_tokens: 4096
```

## EXAMPLE OF .JSON CHAT HISTORY

```
Salva Copia Comprimi tutto Espandi tutto Filtra JSON

▼ 0:
  role: "user"
  content: "Ciao. Mi chiamo Orlando, piacere di conoscerti. Parliami brevemente di te."
▼ 1:
  role: "assistant"
  content: "Ciao Orlando! Piacere mio conosciuto. Io sono un'intelligenza artificiale linguistica, creata per aiutare e interagire con le persone in diverse lingue. Non ho una vita personale vera e propria, ma sono programmata per essere sempre disponibile e pronta a fornirti informazioni o assisterti in qualsiasi modo possibile. Mi piace imparare e migliorarmi giorno dopo giorno, grazie alle conversazioni come questa. Sono felice di conoscerti e di aiutarti nel modo migliore possibile!"
▼ 2:
  role: "user"
  content: "Potresti dirmi come posso migliorare la tua precisione?"
▼ 3:
  role: "assistant"
  content: "Orlando! La mia precisione è direttamente legata alla qualità e quantità dei dati che ho ricevuto durante il mio training. Quindi, per migliorarla, potrai aiutarmi in diversi modi:\n\n1. «Correzione degli errori»: Se noti un errore nella mia risposta, segnalalo e spiegami perché la tua risposta è giusta. Questo mi aiuterà a imparare e a migliorare la mia precisione.\n2. «Fornire esempi e contesti»: Più esempi e contesti che ti fornisca, più posso apprendere e generalizzare le regole di base per rispondere in modo preciso.\n3. «Feedback positivo e negativo»: Il tuo feedback è fondamentale per il mio addestramento. Se la mia risposta è stata utile o soddisfacente, segnalala pure! Se, invece, non è stata precisa o sufficiente, segnalala e spiegami come posso migliorarla.\n4. «Domande specifiche e precise»: Più le tue domande saranno specifiche e precise, più potrò fornire risposte accurate e personalizzate per te.\n5. «Partecipazione attiva»: La tua partecipazione attiva e la tua voglia di imparare sono fondamentali per il mio addestramento e miglioramento.\n\nSpero che questi suggerimenti ti siano stati utili, Orlando!"
```

## ***TODO:***

Per il momento il tool è stato realizzato da riga di comando e per eseguirlo è necessario avere tutti i requirements necessari oltre che le API\_KEY di Google Gemini e Claude che, per motivi di privacy e mancanza di risorse, sono state nascoste nel codice.

Il tool è solamente allo stato iniziale, in attesa di una revisione iniziale, e si prevede di migliorarlo sempre più aggiungendo altri modelli e magari realizzando una vera e propria applicazione con grafica semplificata che simuli un vero e proprio chatbot. (tramite <<streamlit>>).