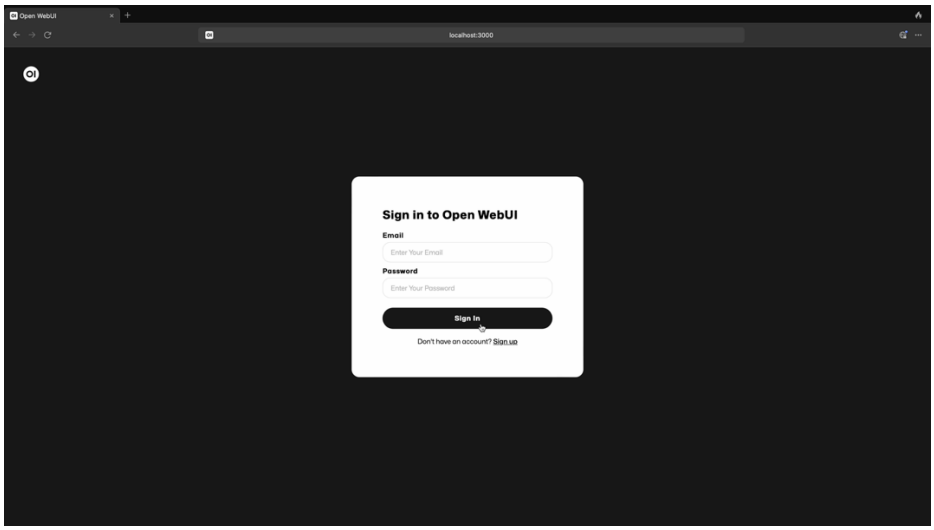
aruba.it

FRAMEWORK OLLAMA + WEBUI

Come Eseguire Modelli di AI Generativa

Aruba Server Dedicati

Framework Ollama + WebUI



Il framework utilizzato si basa su:

- generazione testi - Text Generation (Ollama + WebUI)
- generazione immagini - Image Generation (Automatic + Comfy)
- RAG - Retrieval-Augmented Generation (RAG).

Modelli

Un modello AI (*Artificial Intelligence*) è un sistema computazionale progettato per eseguire azioni che normalmente richiedono intelligenza umana. Queste includono la capacità di comprendere e generare linguaggio naturale (NLP - *Natural Language Processing*), la generazione e la traduzione di testo, il riconoscimento e la generazione di immagini, l'analisi predittiva e molto altro.

I vari modelli AI si differenziano tra loro principalmente per le seguenti caratteristiche:

- **Architettura**
  - La struttura interna del modello che determina come i dati vengono elaborati;
- **Dimensioni**
  - Il numero di parametri nel modello, che influisce sulla sua capacità di apprendimento e accuratezza;
- **Dati e tecniche di addestramento**
  - Il tipo di dati utilizzati per addestrare il modello e i metodi impiegati per migliorare la sua performance;
- **Efficienza e scalabilità**
  - La capacità del modello di utilizzare le risorse computazionali in modo efficiente e di adattarsi a dataset di grandi dimensioni.

I modelli preconfigurati del Framework Ollama + WebUI sono i seguenti:

	TIPOLOGIA [1]	UTILIZZO	PARAMETRI [2]	SITE
GEMMA	LLM	chat	9 Billions	ollama.com/library/gemma
LLAMA 3	LLM	chat	8 Billions	ollama.com/library/llama3
LLAMA 2	LLM	chat	7 Billions	ollama.com/library/llama2
MISTRAL	LLM	chat	7 Billions	ollama.com/library/mistral
PHI3	LLM	chat	3 Billions	ollama.com/library/phi3

1. *LLM (Large Language Model).*
2. *Le dimensioni possono variare nel tempo.*

Settings

Per poter accedere ai settaggi generali del framework procedi come indicato:

- clicca sull'icona del tuo profilo presente in basso a sinistra;
- seleziona la voce **Settings** tra quelle proposte;
- accederai alla sezione **Settings** dove potrai intervenire sulle seguenti voci:

General
Da questa sezione puoi gestire il tema da utilizzare e la lingua con cui utilizzare il framework.

Connections
Da questa sezione puoi gestire la API di OpenAI e Ollama.

Models
Puoi gestire i modelli Ollama.

Interface
Ti permette di personalizzare l'interfaccia WebUI impostando l'attivazione degli add-on.

Personalization
Ti permette di personalizzare le iterazioni con i modelli LLM.

Audio
Puoi configurare i motori per STT (Speech-to-Text) e TTS (Text-to-Speech) e scegliere quale voce utilizzare.

Images
Puoi scegliere quale modello impostare come motore di default per la generazione di immagini.

Chats
Da questa sezione puoi attivare/disattivare lo storico delle chat, e fare import, export, archiviazione o cancellazione delle chat.

Account
Da questa sezione puoi gestire l'immagine del profilo, cambiare il nome e la password.

About
Sono indicate la versione correnti con la possibilità di verificare la presenza di eventuali aggiornamenti disponibili.

Admin Panel - Add User

Per creare un nuovo utente procedi come indicato:
- clicca sul tuo profilo in basso a sinistra;
- seleziona la voce **Admin Panel** tra quelle proposte;
- nella sezione **All Users** sarà visualizzata la lista degli utenti creati;
- clicca sul pulsante + presente in alto a destra;
- si aprirà il livello **Add User**;
- scegli se creare un singolo utente (tab **Form**) o caricare un file CSV contente più utenti (tab **CSV Import**);
- se hai selezionato l'opzione Form devi indicare i seguenti dati:
    - **Role** - è il ruolo che vuoi assegnare all'utente (Pending, User o Admin);
    - **Name** - è il nome assegnato all'utente;
    - **Email** - è la email dell'utente, sarà lo username che l'utente dovrà utilizzare per accedere al framework;
    - **Password** - la password che l'utente dovrà utilizzare per accedere al framework.

Admin Panel - Admin Settings

Per accedere alle impostazioni dell'amministratore procedi come indicato:
- clicca sul tuo profilo in basso a sinistra;
- seleziona la voce **Admin Panel** tra quelle proposte;
- nella sezione **All Users** sarà visualizzata la lista degli utenti creati;
- clicca sull'icona a forma di rondella presente in alto a destra;
- si aprirà il livello **Admin Settings**;
- saranno presenti le seguenti voci:
    - **General**
    - **Users**
    - **Database**
    - **Banners**
    - **Pipelines**

Prompts

Il prompt è l'input che guida l'interazione tra l'utente e il sistema AI, è il testo o l'istruzione che viene fornita al sistema.

Creare un prompt

Per creare un prompt procedi come indicato:

- clicca sulla voce **Workspace** presente nel menu di sinistra;
- nella sezione **Workspace** seleziona la voce **Prompt**;
- a destra dell'area di ricerca Search Documents clicca sul pulsante **+**;
- inserisci **Title**, **Command** e **Prompt Content**.

Importare ed esportare un prompt

- Per importare ed esportare un prompt procedi come indicato:
- clicca sulla voce **Workspace** presente nel menu di sinistra;
- nella sezione **Workspace** seleziona la voce **Prompt**;
- a destra, sotto l'area di ricerca Search Documents, sono presenti due pulsanti;
    - **Import Prompts**;
    - **Export Prompts**.

Documents

Caricare un documento

Per caricare un documento procedi come indicato:

- clicca sulla voce **Workspace** presente nel menu di sinistra;
- nella sezione **Workspace** seleziona la voce **Documents**;
- a destra dell'area di ricerca Search Documents clicca sul pulsante **+**;
- si aprirà il livello **Add Docs** da cui è possibile selezionare un documento da caricare tramite il pulsante **Click here to select documents**;
- contestualmente al caricamento del documento puoi assegnargli uno o più tag tramite la voce **+ Add Tags**.

Tag

I tag servono per classificare i documenti caricati e poterli ricercare in fase di domanda scritta sulla chat.

Documents Settings

Per poter settare i documenti caricati procedi come indicato:

- clicca sulla voce **Workspace** presente nel menu di sinistra;
- nella sezione **Workspace** seleziona la voce **Documents**;
- clicca sul pulsante **Documents Setting** in alto a destra;

- nella sezione **Documents Setting** potrai intervenire sulle seguenti voci:

General

- **General settings**
  - **Scan for documents from /data/docs**
    Esegue la scansione dei documenti che sono presenti nella directory /data/docs.
  - **Embedding Modal Engine**
    Puoi modificare il motore di embedding, cioè il modello che crea le rappresentazioni vettoriali dei chunk, per poi salvarle in un database vettoriale.
  - **Hybrid Search**
  - **Embedding Modal**
  - **Reset Vector Storage**

Puoi cancellare tutti i documenti presenti nello spazio /data/docs.

Chunk Params

- **Chunk Params**
  - **Chunk Size**
    Ti permette di settare il numero massimo di token che vuoi per ogni singolo chunk.
  - **Chunk Overlap**
    Indica di quanti token i vari chunk si devono sovrapporre.
  - **PDF Extract Images (OCR)**

Query Params

- **Query Params**
  Top k Indica il numero dei documenti più rilevanti tra quelli analizzati.
- **PDF Extract Images (OCR)**
  È il prompt, puoi modificarlo gestendo come comportarsi nelle varie casistiche.

Web Params

- **Web Loader Settings**
- **Bypass SSL verification Websites**
- **Youtube Loader Settings**
- **Language**

Chunk

Un chunk rappresenta un blocco di dati o informazioni trattato come se fosse una singola unità, che può essere gestita separatamente. La divisione dei dati o delle informazioni in chunk serve a facilitarne l'elaborazione, l'archiviazione e la ricerca.
Sei i dati da elaborare fossero un libro, i chunk potrebbero essere considerati come i capitoli, le pagine, i paragrafi, le frasi, ecc..
I chunk, in caso di NLP (Natural Language Processing), si conteggiano in token che, a seconda del modello AI utilizzato, corrisponde a parole o caratteri o sub-parole.

Chunk Size

Rappresenta il numero di token, di un documento testuale (DOC, PDF, TXT, ecc.), conteggiati per creare una rappresentazione vettoriale, che sarà poi salvata nel database tramite il processo di embedding.

In caso di NLP un Chunk Size appropriato potrebbe essere compreso tra 1.000 e 10.000 token.

In altri contesti di ML (Machine Learning) potrebbe variare ampiamente.

Chunk Overlap

Indica di quanti token i vari chunk si devono sovrapporre, per garantire che ci sia continuità e coerenza contestuale tra i dati.

In caso di NLP un Chunk Overlap appropriato potrebbe essere un valore tra il 5% ed il 20% del Chunk Size.

In altri contesti di ML potrebbe variare ampiamente.

> Non esistono impostazioni standard dei valori Chunk Size e Chunk Overlap, questi cambiano a seconda dei dataset, dei modelli AI e delle risorse hardware. I valori qui forniti sono puramente di esempio e potrebbero non essere appropriati a specifiche casistiche.

Glossario

AI (Artificial Intelligence)

L'intelligenza artificiale (AI) è una tecnologia avanzata che consente ai sistemi computazionali di replicare i processi dell'intelligenza umana. Attraverso l'implementazione di algoritmi sofisticati e complessi, l'AI è in grado di eseguire attività che tradizionalmente richiederebbero l'intervento umano, come l'apprendimento, il ragionamento, la risoluzione dei problemi e la comprensione del linguaggio naturale.

Chunk

Un chunk rappresenta un blocco di dati o informazioni trattato come se fosse una singola unità, che può essere gestita separatamente. La divisione dei dati o delle informazioni in chunk serve a facilitarne l'elaborazione, l'archiviazione e la ricerca.

CUDA (Compute Unified Device Architecture)

È un'architettura hardware per l'elaborazione parallela creata da NVIDIA. Consente agli sviluppatori di utilizzare le capacità di calcolo delle GPU (Graphics Processing Units) per eseguire operazioni di calcolo intensivo, accelerando notevolmente le prestazioni rispetto all'uso delle CPU (Central Processing Units) tradizionali.

Deep learning

L'apprendimento profondo è quel campo di ricerca dell'apprendimento automatico e dell'intelligenza artificiale che si basa su diversi livelli di rappresentazione, corrispondenti a gerarchie di caratteristiche di fattori o concetti, dove i concetti di alto livello sono definiti sulla base di quelli di basso.

Embedding

È una tecnica utilizzata nell'ambito dell'apprendimento automatico e dell'elaborazione del linguaggio naturale (NLP) per rappresentare dati complessi, come parole, frasi o oggetti, in uno spazio vettoriale.

Fine-tuning

Il fine-tuning rappresenta una tecnica di machine learning che comporta l'ulteriore addestramento di un modello già pre-addestrato su un nuovo dataset specifico per una particolare attività. Questo metodo risulta particolarmente utile nel contesto di modelli di grandi dimensioni, come ad esempio i Large Language Models (LLM), i quali sono stati precedentemente addestrati su enormi volumi di dati generici.

LLM (Large Language Model)

È un modello AI progettato per comprendere e generare linguaggio naturale. Si distingue per l'elevato numero di parametri che vengono addestrati su vaste raccolte di testi. Grazie alla sua grandezza, può acquisire una conoscenza approfondita delle strutture linguistiche e dei contesti. Rappresentano un significativo avanzamento nell'elaborazione del linguaggio naturale, migliorando l'interazione uomo-computer in modo più naturale ed efficiente. A causa della sua complessità, richiede notevoli risorse computazionali sia per l'addestramento che per l'implementazione.

ML (Machine Learning)

È un sottoinsieme dell'intelligenza artificiale (AI) che si concentra sullo sviluppo di algoritmi e modelli che permettono ai computer di apprendere dai dati e fare previsioni o prendere decisioni. Diversamente dai programmi tradizionali, che richiedono istruzioni esplicite per eseguire compiti specifici, i sistemi di machine learning migliorano automaticamente le loro prestazioni nel tempo attraverso l'analisi dei dati. Il ML consente ai computer di affinare le proprie capacità su compiti specifici grazie all'esperienza acquisita dai dati.

NLP (Natural Language Processing)

Il processamento del linguaggio naturale è un campo dell'intelligenza artificiale (AI) che si concentra sull'interazione tra computer e linguaggio umano. L'obiettivo è permettere ai computer di comprendere, interpretare e generare linguaggio naturale in modo significativo e utile.

RAG (Retrieval Augmented Generation)

È il processo di ottimizzazione dell'output di un modello linguistico di grandi dimensioni, in modo che faccia riferimento a una base di conoscenza autorevole al di fuori delle sue fonti di dati di addestramento prima di generare una risposta.

STT (Speech-to-Text)

È una tecnologia che permette di converte il parlato in testo scritto, consentendo ai computer e agli altri dispositivi di comprendere e processare il linguaggio umano parlato.

TTS (Text-to-Speech)

È una tecnologia che permette di convertire il testo scritto in parlato utilizzando tecniche avanzate di ML.