

chat-proxy

Progetto di web chat verso motore LLM (proxy) per task chat e supporto coding privato/riservato.

Obiettivi generali

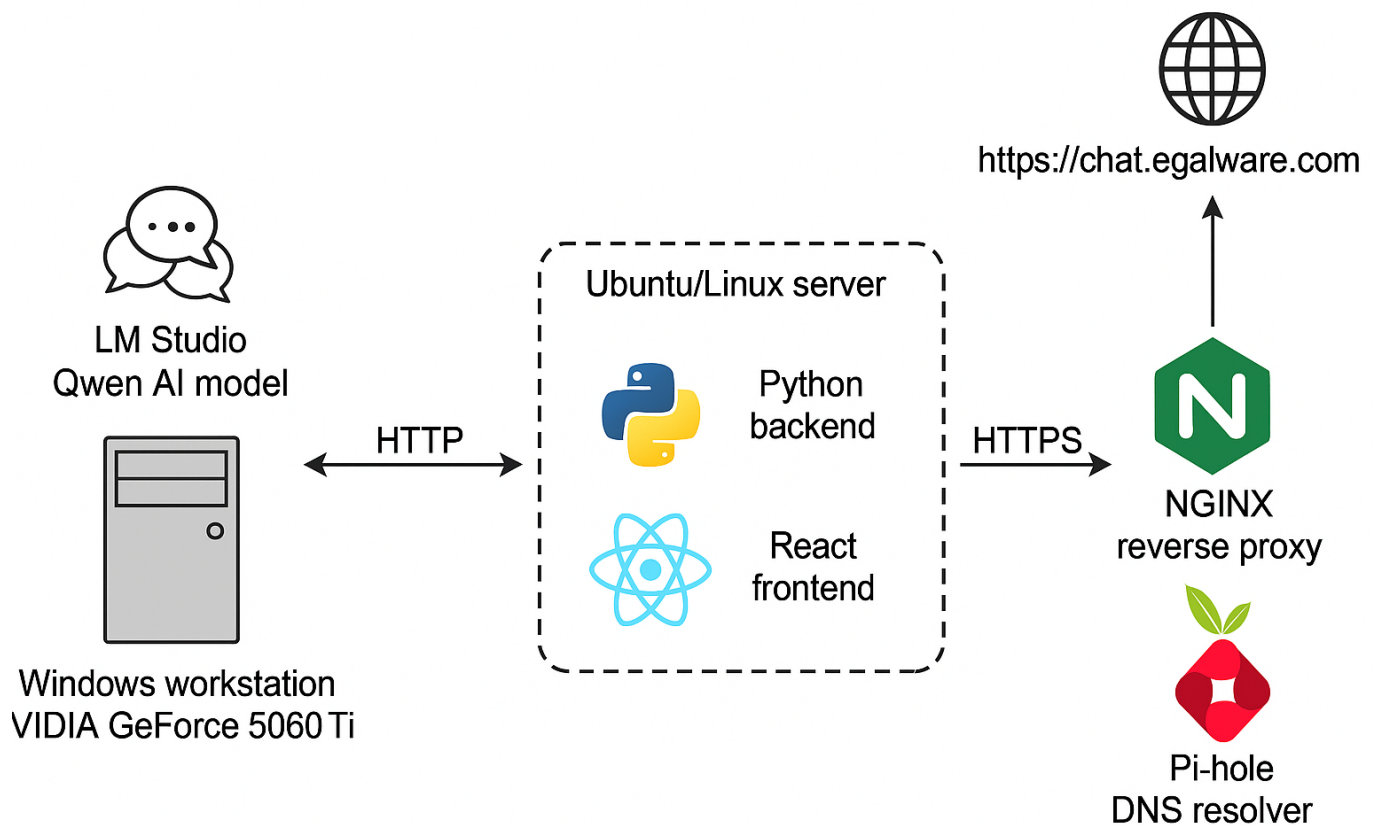
L'obiettivo iniziale è avere un agente AI basato su soluzioni opensource

- da eseguire localmente su HW presente in ufficio
- dove poter effettuare chiamate anche con codice sorgente proprietario senza temere di inviare info riservate esternamente
- per poter avere un agente sempre disponibile con le risorse allocate
- da poter successivamente addestrare con risorse interne tra cui
 - wiki
 - sorgenti di codice aziendale

Setup soluzione

La soluzione è basata sul seguente stack

- LM Studio per esecuzione modello LLM locale (al momento su workstation Sam + scheda video AMD e poi NVidia)
- Abilitazione LM Studio x chiamate locali su porta 1234
- virtual machine linux con soluzione backend/frontend di proxy/caching verso il modello AI di LM Studio



Startup

Al momento per l' esecuzione della soluzione, sulla virtual machine di proxy, vanno avviati backend (python) e frontend (node) manualmente.

Avvio soluzione:

backend

```
uvicorn main:app --host 0.0.0.0 --port 8000 --reload
```

frontend

```
npm run dev
```

ToDo's: trasformare in servizi da abilitare all'avvio macchina

Usage

Per utilizzare la soluzione basta andare (in ufficio o via vpn) all'indirizzo

<https://chat.egalware.com>

e da li fare domande all'AI.

Roadmap

Mancano molti punti di ottimizzazione:

- ☐ gestione utenti locali (oauth? openID? user/pwd? username? IP?)
- ☐ gestione sessioni indipendenti (setup REDIS da verificare) per gli utenti con history
- ☐ miglioramento grafica
- ☐ output performances
- ☐ test modelli LLM più consistenti con scheda video + capace
- ☐ completamento logiche RAG
- ☐ fine tuning (o qualunque altra tecnica di post-addestramento) per aggiungere sorgenti private tra cui
 - ☐ wiki aziendali
 - ☐ documentazione
 - ☐ codice sorgente (eventualmente da repo GIT con + versioni)

Versioni

Versione	Note	Data
0.1.2508.2019	Versione test solo locale con LM Studio	2025.08.20

Versione	Note	Data
0.1.2508.2119	Versione con esecuzione locale completa	2025.08.21
0.1.2508.2219	Versione completa e rivisitata graficamente x chat (con memoria sessioni)	2025.08.22
0.2.2509.0317	Miglioramento gestione memoria sessioni	2025.09.03
0.3.2509.0515	Gestione visualizzazione elenco modelli + selezione x sessioni (e x nuove sessioni) del modello richiesto	2025.09.05