BYD
Byte Your Dreams

byteyourdreams.swe@gmail.com

Verbale Esterno · Data: 20/02/2025

Informazioni documento

Luogo	Teams
Orario	16.00 - 17.00
Redattore	L. Zanesco
Verificatore	L. Albertin
Amministratore	Y. Huang
Destinatari	T. Vardanega R. Cardin
Partecipanti Interni	L. Albertin A. Mio O.F. Stiglet Y. Huang A.M. Margarit

Il responsabile: L. Albertin
L'Azienda: Vimar S.p.A.

Registro delle modifiche

Versione	Data	Autore	Verificatore	Dettaglio
0.1.0	21/02/2025	L. Zanesco	L. Albertin	Prima redazione

Indice

Byte Your Dreams

Febbraio 20, 2025

Contents

1 Revisione del periodo precedente

A causa di impegni accademici, durante il periodo precedente, è stato riscontrato un calo drastico di produttività da parte del gruppo.

In data 20/02/2025 il gruppo ha effettuato il colloquio relativo all'*RTB*$_G$ con il *Prof. Cardin*. Da tale incontro sono emerse alcune incomprensioni e possibili criticità nelle tecnologie scelte che verranno discusse e riviste all'interno del *team*$_G$.

2 Ordine del giorno

2.1 Ruoli dei membri

I ruoli non hanno subito alcuna variazione dai periodi precedenti.

2.2 Agent LLM

Per riuscire a gestire il maggior numero di domande possibili, che gli installatori possono effettuare, il gruppo ha pensato la possibilità di implementare un *LLM agent*$_G$. Quest'implementazione vede alla base una classificazione da parte del *LLM*$_G$ del tipo di domanda effettuata dall'utente, con una successiva instradazione in base al tipo di domanda ottenuta.

Per la reale concretizzazione dell'idea, il gruppo ha bisogno di capire a fondo che tipo di prompt utilizzare, e come gestire in maniera opportuna i tipi di domande per la generazione della risposta.

2.3 Applicativo Web

È stata discussa ampiamente la logica dell'*applicativo web*$_G$ realizzato per il *PoC*$_G$. Il gruppo ha esposto all'Azienda il funzionamento generale dell'*applicativo*$_G$, specificando la presenza di componenti che funzionano come *API*$_G$.
Durante tale discussione il gruppo ha mostrato all'Azienda il codice dell'*applicativo*$_G$ e dopo un'analisi dettagliata, l'Azienda ha consigliato di riordinare al meglio la codebase.

Al fine di una comprensione migliore della logica dell'*applicativo*$_G$, l'Azienda ha proposto al gruppo di realizzare dei diagrammi di attività.

3 Chiarimenti ulteriori

3.1 Problema relativo all'LLM

Il gruppo durante alcuni test, ha riscontrato dei problemi con alcune versioni dell'*LLM*$_G$ scelto. In particolare, con le versioni da 1.5B e da 7B di parametri, la maggior parte delle risposte generate è in inglese. L'Azienda ha ipotizzato la possibilità che i modelli siano troppo piccoli ed ha consigliato quindi di provare ad utilizzare modelli più grandi.

3.2 Scarse risorse computazionali

Il gruppo ha riportato che il problema, riportato nel verbale del 22 gennaio 2025, relativo alle scarse risorse computazionali per la generazione della risposta continua a persistere .
A tal proposito l'Azienda ha proposto di provare ad istanziare 2 servizi Ollama differenti. Il primo a carico della CPU, utilizzato per l'operazione di *embedding*$_G$; il secondo servizio, invece, sarà a carico della GPU ed effettuerà l'operazione di generazione delle risposte alle domande.
Questa soluzione dovrebbe risolvere il continuo *swap*$_G$ dei modelli in memoria, riducendo così i tempi di attesa per la generazione delle risposte.