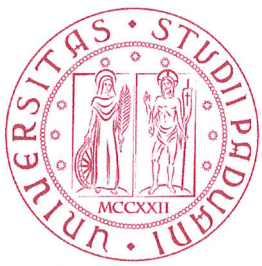
Verbale Riunione 11-03-2024

Gruppo Argo

Informazioni sul documento	
Versione	1.0.0
Approvazione	Sebastiano Lewental
	Zucchetti S.p.A.
Uso	Esterno
Distribuzione	Zucchetti S.p.A.
	Prof. Tullio Vardanega
	Prof. Riccardo Cardin
	Gruppo Argo

Descrizione

Questo documento descrive l'incontro del gruppo Argo con l'azienda Zucchetti S.p.A. per la discussione del capitolato ChatSQL.



UNIVERSITÀ
DEGLI STUDI
DI PADOVA

Registro delle modifiche

Versione	Data	Autore/i	Ruolo	Descrizione
1.0.0	19-03-2024	Sebastiano Lewental	*Responsabile di Progetto*	Approvazione documento
0.1.0	15-03-2024	Mattia Zecchinato	*Verificatore*	Revisione completa
0.0.2	14-03-2024	Riccardo Cavalli	*Redattore*	Stesura del documento
0.0.1	11-03-2024	Raul Pianon, Marco Cristo	*Redattore*	Prima bozza del documento

Indice

1 Informazioni

- **Inizio incontro:** 12:00
- **Fine incontro:** 12:45
- **Tipo incontro:** remoto (Zoom)

1.1 Partecipanti

- **Argo:**
  - Tommaso Stocco
  - Marco Cristo
  - Raul Pianon
  - Sebastiano Lewental
  - Martina Dall'Amico
  - Riccardo Cavalli
- **Zucchetti S.p.A.:**
  - Gregorio Piccoli

2 Riunione

2.0.1 Ritrovo iniziale

Il gruppo si è riunito circa mezz'ora prima dell'appuntamento con l'azienda per discutere la chiarezza delle domande e organizzare una scaletta. Inoltre, sono stati formulati dei quesiti aggiuntivi per garantire un'interazione più fluida con la Proponente. Il gruppo ha affidato il compito di moderare l'incontro al responsabile in carica. Come concordato per via telematica, il referente di Zucchetti si è unito al meeting alle ore 12:00. Dopo una rapida presentazione dell'azienda, il gruppo ha espresso tutti i dubbi e le curiosità in merito al capitolato ChatSQL.

2.1 Argomenti e temi dell'incontro

2.1.1 Materiali di studio

Domanda: Ci sono dei materiali di studio, online o cartacei, che l'azienda consiglia?

Risposta: La reperibilità del materiale di studio, sia online che in versione cartacea, non dovrebbe rappresentare un ostacolo. Ad ogni modo, nel documento di presentazione del capitolato sono riportati alcuni riferimenti, da integrare con ricerche in rete su LLM e SQL.

2.1.2 Origine del capitolato

Domanda: Da dove è nata l'idea di ChatSQL?

Risposta: Zucchetti S.p.A. fornisce supporto ad altre aziende attraverso strumenti di sviluppo (per la produzione di gestionali, ad esempio). Il focus della Proponente è la produttività dei programmatori. Da qualche anno si è diffusa l'idea che ChatGPT possa produrre codice. Ma è davvero così? Spesso si confonde il codice con la struttura di un prodotto (quest'ultima può essere rivista e rielaborata anche a distanza di mesi). Se il cliente tornasse indietro e dicesse "Questo elemento non funziona", quale sarebbe la risposta di ChatGPT? A volte il codice deve essere completamente riscritto, tenendo in considerazione l'analisi dei requisiti e le specifiche tecniche, oltre a questioni di sicurezza. ChatGPT ha problemi di stabilità e di sicurezza, per cui può essere utile nella stesura di codice solamente se dall'altro lato vi è una persona competente. Il 95% dei consigli di ChatGPT non funziona, mentre il 100% non segue gli standard di produzione. Può essere utile per scrivere blocchi di codice? Sì, ma la produzione è tutt'altra storia. Quindi, cosa si può fare realmente con questi sistemi? La Proponente sottolinea come la scrittura di frasi SQL sia una buona pratica da automatizzare. L'ambiente è semplice, così come la scrittura di codice, mentre la sicurezza e gli standard di produzione sono demandati a un altro livello. Si tratta del miglior terreno possibile per ottenere risultati da un LLM. Forse un domani ChatGPT sarà in grado di realizzare prodotti, ma al momento può solamente fornire blocchi di codice più o meno scarni.

2.1.3 Tecnologie

Domanda: Ci sono delle tecnologie o librerie che l'azienda consiglia per svolgere il progetto?

Risposta: L'azienda non impone nessun vincolo sui linguaggi di programmazione e le tecnologie da utilizzare. Tuttavia, il consiglio è di usare Python server side e txtai come libreria open source per la ricerca semantica e l'orchestrazione LLM. Inoltre, la Proponente suggerisce di realizzare una web app con HTML, CSS e JavaScript.

2.1.4 Server

Domanda: Parlando di server side, l'azienda mette a disposizione un server, o è preferibile che il gruppo lavori con i propri strumenti?

Risposta: La Proponente consiglia di lavorare con i propri strumenti o, in alternativa, di appoggiarsi a servizi esterni come Amazon Web Services. Tuttavia, i requisiti del capitolato non dovrebbero appesantire o rallentare le macchine locali.

2.1.5 Esempi concreti

Domanda: Potremmo avere qualche esempio concreto, in modo da chiarire gli obiettivi minimi del capitolato?

Risposta: [chat.lmsys.com] Gli esempi sono realizzati con il supporto di LMSYS Org (Large Model Systems Organization), un'organizzazione di ricerca il cui obiettivo è rendere i modelli di grandi dimensioni accessibili a chiunque. La Proponente sottopone a LLaMA la seguente richiesta in linguaggio naturale:

- "Scrivi una frase SQL che restituisce la lista dei clienti che abitano a Padova".

Vengono confrontati due modelli differenti, uno dei quali dà una soluzione diretta, mentre l'altro fornisce una spiegazione più generale. In entrambi i casi, però, il risultato è una frase SQL standard che può interrogare una tabella generica. Viene mostrata inoltre la classifica LMSYS Chatbot Arena Leaderboard, che vede al primo posto GPT-4. Tornando al prompt precedente, è chiaro che mancano dei dettagli. L'interrogazione deve quindi essere arricchita:

- "Scrivi una frase SQL per la seguente richiesta: dammi i clienti che abitano a Padova".

Un ulteriore miglioramento potrebbe essere l'aggiunta di una descrizione dello schema della tabella.

- "Nel database ho una tabella TAB_CLI che contiene i dati dei clienti, i campi di TAB_CLI sono CODCLI per il codice del cliente, INDCLI per l'indirizzo del cliente e NOMCLI per il nome del cliente. Scrivi una frase SQL per la seguente richiesta: dammi i clienti che abitano a Padova".

Chiaramente il modello sta facendo delle assunzioni, ma possiamo notare che la risposta è più specifica. Qual è l'ostacolo del capitolato? Per ottenere un risultato

preciso, abbiamo dovuto fornire in input una porzione di dizionario dati. Dapprima bisogna capire come scrivere il dizionario dati. Avendo a disposizione un database e una serie di tabelle, di campi e di relazioni tra le tabelle, qual è la soluzione migliore per scrivere il dizionario dati, in modo che, una volta generato il prompt, il modello restituisca la frase SQL corretta? Un ulteriore problema riguarda la lunghezza limitata del prompt. Qui subentra la ricerca semantica: è necessario analizzare la richiesta in linguaggio naturale per capire quali porzioni del dizionario dati inserire nel prompt.

2.1.6 Architettura

Domanda: Come pensate di strutturare l'architettura del progetto?

Risposta: La prima parte riguarda il caricamento del dizionario dati da parte di un utente con il ruolo di amministratore. Viene menzionata la possibilità di utilizzare JSON. Gli utenti finali, invece, hanno la possibilità di inserire la richiesta in linguaggio naturale all'interno di un'area di testo (textarea). L'output restituito dall'applicativo è un prompt da copiare e incollare su ChatGPT. Per migliorare notevolmente la user-experience, la Proponente suggerisce di sfruttare le API di ChatGPT, in modo da connettersi direttamente al LLM e visualizzare sotto al prompt la frase SQL generata. L'unico motivo per cui tale requisito è stato marcato come opzionale è l'eventualità di dover sottoscrivere un abbonamento a pagamento. Una soluzione proposta è quella di visitare e approfondire LM Studio [lmstudio.ai], che consente di eseguire LLM tramite un server locale compatibile con OpenAI. Sul sito web di LM Studio è linkato il repository GitHub di LLaMa C++, una libreria sviluppata da Georgi Gerganov come risposta alle richieste di hardware sofisticato da parte dei Large Language Models. Viene menzionato anche Hugging Face per la ricerca di modelli (tra cui Mistral, il Large Language Model di Mistral AI). Sempre su Mistral, cercando Mistral-GGUF di TheBloke, si ottiene il modello in formato GGUF (si parla di quantizzazione dei modelli di IA, un processo attuato per ridurre la complessità dei modelli stessi, spesso a vantaggio della velocità di esecuzione o della memoria richiesta). Come si può notare, il file più grande ha una dimensione di 7 GB. Tutto ciò di cui si è appena discusso può essere realizzato in locale ed è open source. Inoltre, l'interfaccia di LM Studio è molto semplice e intuitiva, sia per quanto riguarda la scelta dei modelli, sia per quanto concerne il server.

2.1.7 Prompt engineering

Domanda: È corretto dire che il punto focale del progetto è il prompt engineering?

Risposta: In un certo senso sì, ma in realtà c'è anche una fase abbastanza complessa di ricerca semantica, per questo è consigliato txtai. Come è stato sottolineato in precedenza, il prompt non può contenere l'intero dizionario dati. Bisogna prima capire quali tabelle sono coinvolte in una richiesta. Solo a quel punto si costruisce il prompt per produrre una frase SQL sensata. È preferibile un sistema con tanta recall, piuttosto che un sistema con grande precisione. Con recall si intende la capacità del sistema di estrarre tutte le informazioni rilevanti presenti nel dizionario dati. Un siste-

ma con tanta recall è in grado di carpire un'ampia gamma di informazioni pertinenti, anche se ciò potrebbe comportare l'inserimento di elementi non rilevanti.

2.1.8 Riflessioni sul capitolato

Domanda: Avendo seguito gli altri gruppi, la vostra idea sul capitolato è cambiata? Quali parti pensate possano risultare più complesse?

Risposta: Le idee dell'azienda sulla fattibilità del progetto sono rimaste pressoché invariate. Si pensava che lavorare in locale fosse più difficile; invece, i gruppi del primo lotto non hanno riscontrato problemi al riguardo. L'ostacolo maggiore è rappresentato dalla selezione delle porzioni del dizionario dati, in quanto è necessario carpire solo i pezzi pertinenti. Può essere interessante inserire dei sinonimi o delle similitudini all'interno del dizionario dati. Nulla vieta di chiedere aiuto a un LLM, per esempio:

- Domanda: "Dammi un sinonimo di cliente";
- Risposta: Committente.

2.1.9 Reperibilità dell'azienda

Domanda: L'organizzazione degli incontri è a discrezione del gruppo? Con quale frequenza è possibile pianificare riunioni?

Risposta: Non ci sono particolari vincoli o limitazioni, l'importante è avvisare con un po' di anticipo il referente dell'azienda. I meeting vengono organizzati via mail e si svolgono da remoto su Zoom. In attesa delle revisioni di avanzamento, la Proponente consiglia di svolgere riunioni con frequenza maggiore. Viene infine menzionata la sperimentazione come buona pratica per affrontare il progetto didattico.

Luogo e Data:
Padova (PD) 11/03/2024

Firma: _____

Responsabile: Sebastiano Lewental

Per approvazione:

Firma: _____

Referente: Gregorio Piccoli (Zucchetti S.p.A.)