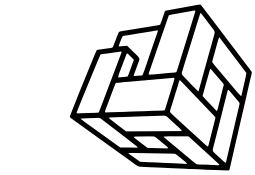
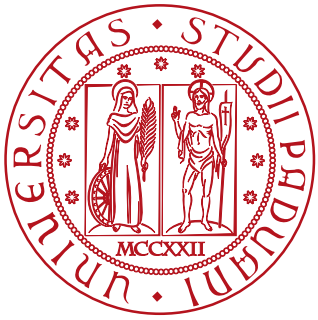


PEBKAC
Email: pebkacswe@gmail.com
Gruppo: 11



Università degli Studi di Padova
Corso di Laurea: Informatica
Corso: Ingegneria del Software
Anno Accademico: 2024/2025

Verbale Esterno
23 ottobre 2024

Informazioni sul documento:

Responsabile	Tommaso Zocche
Verificatore	Alessandro Benin
Redattori	Derek Gusatto
	Matteo Piron
Uso	Esterno
Destinatari	Tullio Vardanega
	Riccardo Cardin

Abstract:

Verbalizzazione delle domande con le relative risposte fornite da Mariano Sciacco (Vimar S.p.A.) attraverso documento Google condiviso.

Registro delle modifiche

Versione	Data	Autore	Ruolo	Descrizione
2.0.0	05/11/2024	Tommaso Zocche	Responsabile	Approvazione
1.1.0	05/11/2024	Alessandro Benin	Verificatore	Verifica
1.0.1	05/11/2024	Derek Gusatto	Amministratore	Correzioni
1.0.0	05/11/2024	Tommaso Zocche	Responsabile	Approvazione
0.1.0	05/11/2024	Derek Gusatto	Verificatore	Verifica
0.0.2	04/11/2024	Tommaso Zocche	Amministratore	Templating del verbale
0.0.1	04/11/2024	Matteo Piron	Amministratore	Prima stesura

Indice

1 Informazioni generali

- **Tipo riunione**: Esterna
- **Luogo**: Telematico, Google Docs
- **Data**: 23/10/2024
- **Ora inizio**: /
- **Ora fine**: /
- **Presenti**:
  - Alessandro Benin
  - Matteo Gerardin
  - Ion Bourosu
  - Derek Gusatto
  - Davide Martinelli
  - Matteo Piron
  - Tommaso Zocche
  - Mariano Sciacco (Vimar S.p.A.)
- **Assenti**:
  - Nessun Assente

2 Riassunto della riunione

L'azienda, in questa prima fase, ha preferito gestire le domande che il gruppo aveva riguardo il capitolato C2 - VIMAR GENIALE di Vimar S.p.A.
(reperibile al link Capitolato C2)
attraverso un Google Docs condiviso, nella quale abbiamo fornito le seguenti domande ottenendo le rispettive risposte:

1. **È possibile fornire ulteriori informazioni riguardo ai controlli che il componente di interrogazione dovrà effettuare relativamente agli argomenti proibiti?**
   Il componente di interrogazione deve prevedere un controllo per delimitare gli argomenti che l'utente può trattare, evitando argomenti proibiti come politica e finanza durante la conversazione. Dunque:

   - INPUT (= utente fa la domanda): dovete verificare con un filtro cosa sta chiedendo l'installatore. Per fare questa parte, ad esempio, potreste sfruttare proprio un "secondo" LLM di controllo per chiedergli se l'argomento trattato menziona argomenti proibiti. Questo filtro non lo riteniamo essenziale per il completamento del progetto quindi potete decidere voi quale soluzione è più opportuna in termini di costi / beneficio (i.e. una blacklist di argomenti o parole potrebbe essere accettabile).
   - OUTPUT (= LLM fornisce la risposta): è opzionale, ma il funzionamento è lo stesso, ossia un filtro prima che la risposta arrivi all'utente, così da verificare anche l'assenza di allucinazioni o di tentativi di jailbraking.

2. **Relativamente alle tecnologie suggerite, ci sono raccomandazioni più stringenti o la scelta può essere effettuata liberamente tra quelle consigliate senza rischiare, successivamente, di incorrere in dei problemi?**
   Le tecnologie suggerite sono solo consigliate per facilitare lo sviluppo e garantire la compatibilità con le esigenze e i tempi del progetto. Tuttavia, la scelta può essere effettuata liberamente tra quelle consigliate, purché si rispettino i vincoli tecnologici obbligatori, come l'uso di Docker e Git per la gestione del codice. Il mio consiglio è di esplorare e di proporre se avete tecnologie che vi piacerebbe provare, così poi possiamo consigliarvi se il rapporto tra difficoltà di apprendimento / fattibilità / tempo è abbastanza buono.

3. **Riguardo all'utilizzo di AWS, il suo impiego porta particolari funzionalità oltre al completamento di un requisito opzionale?**
   L'utilizzo di AWS è opzionale, ma offre un duplice vantaggio: vi permette di lavorare su un cloud provider richiestissimo nel mondo del lavoro e vi dà modo di "hostare" la vostra soluzione - una volta realizzata - già dopo la fase PoC (e ciò vi torna utile ai fini di test sul campo / test E2E). Chiaramente, di tutto il mondo AWS assaggerete solo una piccola parte della torta, perché AWS ha una quantità infinita di servizi e richiede un'esperienza di almeno 1-2 anni per saperlo adoperare bene.

4. **Il chatbot deve poter essere interrogato anche offline?**
   Domanda lecita. La risposta breve è potenzialmente sì. Rispetto ad altri capitolati, l'idea alla base del progetto è di riuscire ad avere un prodotto che, se tirato su con Docker Compose su un PC locale, mi attiva una serie di servizi (vedi i componenti / moduli dell'architettura proposta) che sono ready-to-use, senza la necessità di risorse "esterne" che richiedono un costante uso di internet. In altre parole, ci aspettiamo che il modello LLM riesca a eseguire localmente nel PC e non ci sia un servizio esterno verso cui appoggiarsi (esempio: Ollama). Nella fase successiva all'assegnazione dei capitolati possiamo comunque discuterne direttamente coi vincitori aprendoci volentieri ad altre proposte, e qualora sorgano difficoltà se ne può parlare.

5. **Verranno fornite una lista di domande di prova con annesse risposte desiderate?**
   Sì. Per agevolarvi con i test E2E vi possiamo fornire una lista di domande con opportune risposte desiderate.

6. **Contesto di deploy, ci sono vincoli per quanto riguarda browser e relative versioni da supportare?**
   L'applicativo deve essere responsive e funzionare via browser su smartphone, tablet e desktop. So che siete esperti di Tecnologie Web, pertanto, per aiutarvi a formalizzare un requisito, ci aspettiamo una buona compatibilità con le versioni recenti (fino a 6 mesi fa) dei browser più famosi: Edge (versione Chromium-based), Google Chrome, Firefox e Safari. A livello di usabilità prendete pure spunto dalla bozza nel capitolato ma siete liberi di proporre qualcosa di (sicuramente) meglio. In termini di accessibilità, nessun vincolo, ma bene se volete pensarci.

7. **Qual è la preferenza dell'azienda riguardo la tecnologia AI? C'è già un modello LLM consigliato da utilizzare?**
   Trovate tutto nel capitolato. Per questo progetto è obbligatorio l'uso

di modelli AI (LLM) Open Source. Sono suggeriti modelli come Llama 3.1, Mistral, Bert o Phi. L'approccio RAG (Retrieval Augmented Generation) è richiesto ed è essenziale per il completamento del progetto. Il mio consiglio è esplorare questi modelli e capire bene quale di questi è adeguato alla soluzione richiesta (valutate lingua, peso in gigabyte del modello, performance comparison, risorse richieste, numero di token per conversazione, ecc.).

8. **La funzionalità di OCR è necessaria per tutti i prodotti o solo per determinati documenti?**
   La funzionalità di OCR è suggerita per reperire dati non strutturati dai PDF dei manuali istruzioni (o, volendo, dalle pagine prodotto). Su questo fronte potete valutare voi come e se impiegarla (potenzialmente potreste usare anche la tecnica text2pdf, ma su questo fronte possiamo discuterne insieme). I dati strutturati (ossia il codice HTML / tabelle presenti nella pagina prodotto) potete ricavarli facilmente e non richiedono OCR.

9. **Per l'implementazione Cloud, il capitolato suggerisce AWS. È necessario utilizzare AWS specificatamente o possiamo utilizzare alternative?**
   L'uso di AWS è opzionale e non obbligatorio. È possibile utilizzare alternative, purché l'infrastruttura Cloud rispetti i principi infrastructure as code (IaC) e uso dei container. Su questo fronte, se avete proposte possiamo parlarne insieme.

Firma del referente Vimar S.p.A.: _____