smartlearn
virtual learning. real success.

CHALLENGE-01-VERIFICARE IL POC

ICT-Formazione professionale Svizzera

Indice

Situazione iniziale **3**

Consegna **6**

Domanda D1 6

Domanda D2 7

Domanda D3 7

Domanda D4 7

Domanda D5 8

Domanda D6 9

Soluzione di riferimento **9**

Contesto di base 10

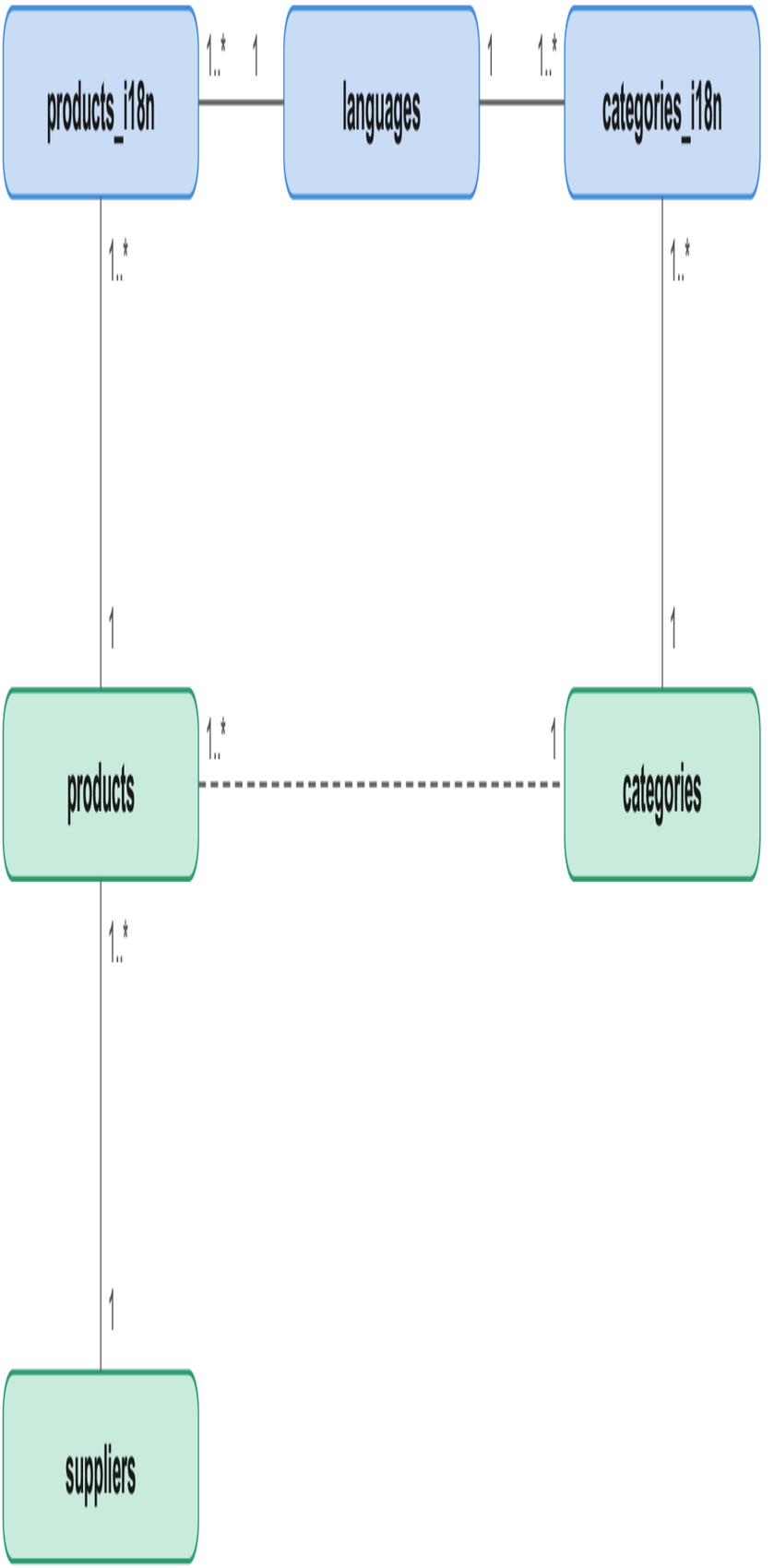
Situazione iniziale

Un'azienda operante nel settore del commercio all'ingrosso di frutta e verdura mette già a disposizione un'API REST tramite la quale è possibile consultare i dati relativi all'assortimento. Attualmente si sta valutando se in futuro queste informazioni potranno essere rese accessibili anche tramite query testuali in linguaggio naturale.

Un approccio promettente in tal senso è il RAG (Retrieval-Augmented Generation): in questo caso, un LLM viene arricchito in modo mirato con dati provenienti da una fonte controllata. Il modello LLM si occupa esclusivamente dell'elaborazione del linguaggio, mentre il database strutturato esistente dell'azienda rimane la fonte dei dati.

Per valutare questo approccio è stato sviluppato un Proof of Concept (PoC) che utilizza il modello linguistico `phi3:mini`. Il modello è stato scelto per le sue dimensioni ridotte e per la possibilità di farlo funzionare localmente su un server di prova.

Nel PoC sono stati utilizzati, a titolo esemplificativo, 4 endpoint dell'API REST esistente come fonte di dati, i quali restituiscono risposte JSON strutturate. La documentazione completa dell'API è accessibile tramite l'interfaccia utente Swagger dell'ambiente di test. Il database del PoC comprende le seguenti tabelle e relazioni:


products_i18n
languages
categories_i18n
products
categories
suppliers
1..*
1

Per l'integrazione dell'LLM sono stati implementati due approcci diversi, resi disponibili sotto forma di endpoint nell'interfaccia utente di Swagger:

- **`POST /llm/v1`**: L'LLM conosce le query SQL utilizzate anche negli endpoint REST. Sulla base di un prompt testuale, l'LLM seleziona una query SQL predefinita. Questo approccio è deterministico perché utilizza query SQL hardcoded.
- **`POST /llm/v2`**: L'LLM conosce la struttura del database, comprese le tabelle e le relazioni. Sulla base di un prompt testuale, l'LLM genera autonomamente la query SQL. Questo approccio è più flessibile, ma più soggetto a errori.

Swagger UI è un'interfaccia utente basata su browser per la descrizione e il test delle API REST. È disponibile sul server di prova `ubu-srv-exam-01` con indirizzo IPv4 `192.168.110.20` ed è accessibile dal browser di un client di prova all'indirizzo **`http://192.168.110.20/docs`**. Nell'ambiente di test sono disponibili a scelta un client con Windows 11 (`win-cli-exam-01`) e un client Ubuntu (`ubu-cli-exam-01`).

Nota: si prega di tenere presente che l'elaborazione delle richieste agli endpoint LLM può richiedere dai 30 ai 90 secondi a causa delle risorse limitate del server. Se non si riceve una risposta entro 180 secondi, l'endpoint interrompe la richiesta con un errore di timeout.

Ambienti e dati di accesso

Funzione	Host	Indirizzo IP	OS	Applicazioni & Strumenti	Utente	Password
Testserver	ubu-srv-exam-01	192.168.110.20/24	Ubuntu 24.04.2 LTS	Swagger UI su Webserver (Uvicorn) Ollama con phi3:mini Database MySQL contenente i dati dell'assortimento	kein Zugriff	
Testclient	ubu-cli-exam-01	192.168.110.yyy/24	Ubuntu 24.04.2 LTS	Browser Firefox	vmadmin	sml12345
Testclient	win-cli-exam-01	192.168.110.101/24	Windows 11 Client	Browser Edge	vmadmin	sml12345

Consegna

Accedi all'interfaccia utente Swagger nel browser di un client di prova all'indirizzo **`http://192.168.110.20/docs`**. Prova ad utilizzare gli endpoint disponibili e inserisci le risposte alle domande da F1 a F6 nei campi di risposta corrispondenti.

Domanda D1

Quante categorie diverse di prodotti sono presenti nel database? Inserisci il numero di categorie nel campo di risposta.

Numero di categorie diverse

Domanda D2

Qual è il prodotto più costoso della gamma? Inserisci il nome del prodotto nel campo di risposta.

Nome del prodotto più costoso

Domanda D3

Qual è la categoria con il maggior numero di prodotti? Inserisci il nome della categoria nel campo di risposta.

Nome della categoria

Domanda D4

Come è noto, le risposte fornite da un LLM non sono deterministiche.

In un test, l'endpoint **`POST /llm/v1`** ha risposto al prompt “`Show me all the products in the berries category.`” con la risposta JSON salvata nel seguente file:

- q4-response.json (https://ict-bb-ch.smartlearn.cloud/media/31dcb6f4033806444c012f.json)

Analizza la risposta fornita e seleziona l'affermazione corretta.

Single-Choice

- ☐ È stato selezionato l'endpoint JSON corretto e i prodotti sono stati riportati correttamente (HTTP 200).
- ☐ È stato selezionato l'endpoint JSON corretto, ma i prodotti riportati appartengono a un'altra categoria. (HTTP 200)
- ☐ È stato selezionato l'endpoint JSON corretto, ma la richiesta non è andata a buon fine a causa di un parametro mancante. (HTTP 400)
- ☐ È stato selezionato l'endpoint JSON corretto, ma il modello (LLM) ha generato nomi di prodotti che non esistono nel database. (HTTP 200)
- ☐ È stato selezionato un endpoint JSON errato e i dati restituiti sono incompleti o non pertinenti. (HTTP 200)
- ☐ La richiesta non è stata associata a nessuna query predefinita. (HTTP 400)

Domanda D5

Quali informazioni mancano nel prompt “`Show me all the products in the berries category.`” affinché l'endpoint `POST /llm/v1` possa eseguire la richiesta in modo affidabile? Motiva brevemente la tua risposta.

Suggerimento: testa la tua risposta direttamente all'endpoint `POST /llm/v1` in Swagger UI.

Domanda D6

In un test, l'endpoint **`POST /llm/v2`** ha risposto al prompt “`Show me the details of bananas.`” con la risposta JSON, che è stata salvata nel seguente file:

- q6-response.json (https://ict-bb-ch.smartlearn.cloud/media/03eaab029dd8fcb74ff28c.json)

Analizza la risposta salvata. Indica il nome della tabella immaginaria e spiega, sulla base dello schema del database nella situazione iniziale, perché tale tabella non esiste.

Soluzione di riferimento

Contesto di base

Modulo: 619 – Verificare la fattibilità di approcci risolutivi basati sull'IA

Obiettivo operativo: OO5 – Valuta approcci risolutivi basati sull'IA con l'aiuto dei dati, delle informazioni e delle esperienze raccolte, sulla base dei criteri definiti per la fattibilità e dei criteri generali relativi all'IA.

Competenza operativa: C4 – Verificare la fattibilità di una soluzione basata sull'IA

Criteri di prestazione: gli FAAIB sono in grado di…

- Definire criteri per la valutazione di soluzioni basate sull'IA, tenendo conto delle prospettive tecnica, organizzativa, giuridica e sociale;
- Definire la procedura più adatta per la verifica, nonché gli strumenti necessari, insieme a persone specializzate, ad esempio attraverso lo sviluppo di prototipi o lo svolgimento di test in laboratorio o sul campo;
- Analizzare e valutare la fattibilità di una soluzione basata sull'IA secondo i criteri stabiliti, coinvolgendo i gruppi di interesse e le persone specializzate rilevanti;
- Riconoscere segnali di possibili effetti di bias nei dati o nei modelli e, in collaborazione con persone specializzate, valutarne l'impatto sulla soluzione.