



Basi di Dati (16AFQPL, 16AFQPI)

Anno Accademico 2024-2025

Politecnico di Torino

Utilizzo di Large Language Models con Reasoning per la risoluzione di query complesse

Obiettivo

La finalità di questa esercitazione è quella di utilizzare Large Language Models (LLMs) dotati di capacità di “*reasoning*” per la risoluzione di query in linguaggio SQL di livello intermedio-avanzato (livello esame).

L’esercitazione prevede di risolvere diverse query testando le abilità di tre differenti Large Language Models analizzando sia la soluzione che il processo di reasoning, composto solitamente da più passaggi (*reasoning steps*), prodotti da ogni modello.

Query

Le seguenti 10 query sono tratte da 5 temi d’esame del corso, per ogni coppia di query è fornito lo schema logico della base dati corrispondente.

Nota: Poiché i Large Language Models sono allenati prevalentemente su contenuti in lingua inglese, generalmente hanno performance migliori se interrogati in questa lingua. Per questo motivo, i prompt da fornire ai modelli devono essere scritti unicamente in **lingua inglese**.

Tema d'esame 1 (query Q1 e Q2)

- *Database schema:*
BOOK_PUBLISHED(ISBN, Title, AuthorCode, Genre)
AUTHOR(AuthorCode, AuthorName, DateOfBirth, City)
SALE(SaleCode, ISBN, Date, Time, Price)
- **[Q1]** Display the name of the authors who have published at least one book in the "fantasy" genre but have never published a book in the "adventure" genre that has been sold in more than 5,000 copies.
- **[Q2]** Display the title and genre of books for which at least 4,000 copies have been sold overall, and that in April 2025 had a revenue higher than the one obtained in January 2025.

Tema d'esame 2 (query Q3 e Q4)

- *Database schema:*
STUDENT(StudentCode, StudentName, SchoolType, SchoolYear, Section)
TEACHER(TeacherCode, TeacherName, Profession)
CONSULTATION(TeacherCode, Date, StartTime, Duration, Topic, ConsultationCost, StudentCode)
- **[Q3]** Display the teacher's name and the student's name for all teacher-student pairs who have carried out consultations together for a total of more than 10 hours and who have never had more than one consultation together on the same day.
- **[Q4]** Display the teacher's name and the maximum daily amount earned for consultations given to first-year students (specified in the SchoolYear field), considering only those teachers who have conducted consultations on a number of different days greater than the average.

Tema d'esame 3 (query Q5 e Q6)

- *Database schema:*
CLASSROOM(ClassroomCode, ClassroomName, Capacity)
COURSE(CourseCode, CourseName, ECTSCredit, EnrolledCount)
LECTURE(ClassroomCode, Date, StartTime, CourseCode)
- **[Q5]** Display the code, name, and number of ECTS credits of the courses for which, in January 2024, more than two lectures were held on the same day.
- **[Q6]** Display the code and name of the classrooms in which all the lectures of (at least) one course were held.

Tema d'esame 4 (query Q7 e Q8)

- *Database schema:*
ROUTE(RouteCode, OriginAirport, DestinationAirport, Duration)
FLIGHT(RouteCode, Date, DepartureTime, AircraftID)
TICKET(TicketCode, RouteCode, Date, DepartureTime, PassengerName)
BOARDING_PASS(TicketCode, SeatNumber)
- **[Q7]** Display the names of the passengers for flights on August 28, 2024, departing from Turin airport, for whom a ticket was issued but no boarding pass was issued.
- **[Q8]** Considering the month of August 2024, for routes on which more than 30 flights were operated and no flight had fewer than 200 passengers, display the route code, date, departure time, origin airport, and destination airport of each flight on the route.

Tema d'esame 5 (query Q9 e Q10)

- *Database schema:*
CUSTOMER(CustomerCode, CustomerName, DocumentType, DocumentNumber)
HOTEL(HotelCode, HotelName, Municipality, Province, StarRating)
HOTEL_ROOM(RoomNumber, HotelCode, RoomType)
BOOKING(RoomNumber, HotelCode, StartDate, EndDate, CustomerCode)
- **[Q9]** Display the code and name of customers who have made bookings at hotels that have the fewest number of rooms within their municipality.
- **[Q10]** Display the names of pairs of customers who, in at least one hotel, have made the same number of bookings, and who, in at least one hotel, have booked rooms for the same total number of days.

Modelli

I tre LLMs da utilizzare, previa registrazione alle rispettive piattaforme, sono i seguenti:

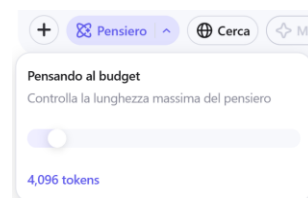
- **Gemini:** <https://gemini.google.com/>

Selezionare dalla tendina in alto a sinistra il modello **2.5 Flash (experimental)**. Non è necessario abilitare esplicitamente il reasoning.


- **Qwen:** <https://chat.qwen.ai/>

Selezionare dalla tendina in alto a sinistra il modello **Qwen3-235B-A22B**.

Abilitare il reasoning cliccando sul pulsante “Pensiero” e impostare dalla tendina un budget di 4096 thinking tokens.



- **ChatGPT:** <https://chatgpt.com/>

Abilitare il reasoning cliccando sull'icona  (“Pensa prima di rispondere”).

Nota: Ad oggi, ChatGPT mette a disposizione un numero limitato (circa 10) di chiamate con la funzionalità reasoning. Si consiglia quindi di utilizzarle con parsimonia e, in caso di esaurimento delle chiamate a disposizione:

- Se si svolge l'esercitazione in coppia, continuare utilizzando l'account ChatGPT del/della vostro/vostra collega oppure creare un altro account.
- oppure
- Continuare l'esercitazione utilizzando gli altri modelli.
- In ogni caso, il numero di chiamate con reasoning a disposizione si resetta ogni 3-4 ore, quindi è possibile completare le parti mancanti anche in un secondo momento.

Svolgimento

Per ognuna delle query, svolgere i seguenti passi e compilare i campi corrispondenti nei form predisposti per l'esercitazione:

1. Scrivere la propria soluzione attesa (senza utilizzare un LLM) in linguaggio SQL. Riportarla nel campo *Soluzione attesa*.
2. Modificare il prompt (in lingua inglese) fornito come esempio inserendo lo schema logico della base dati e la query in linguaggio naturale che si sta risolvendo. Riportarlo nel campo *Prompt*.
3. Per ognuno dei tre Large Language Models da utilizzare:
 - Andare sull'interfaccia web corrispondente e selezionare il modello da utilizzare, abilitando se necessario dall'interfaccia la funzionalità reasoning.
 - Inviare al modello il prompt preparato al punto 2.
 - Copiare l'output del modello nei campi *Reasoning* (dove riportare i passaggi di ragionamento intermedi effettuati dal modello) e *Soluzione SQL* (dove riportare la soluzione finale in linguaggio SQL prodotta dal modello).

Nota: Gemini e Qwen restituiscono in output una sezione separata per il reasoning, mentre ChatGPT decide quando farlo in base alla richiesta.

Copiare nel campo *Reasoning* tutto quello che descrive il ragionamento del modello e nel campo *Soluzione SQL* la soluzione in linguaggio SQL generata.

Nota: Utilizzare i campi *Reasoning (continua)* se l'output non dovesse entrare per intero nel campo *Reasoning*.

- Analizzare il reasoning del modello rispondendo alle seguenti domande:
 - Il reasoning è corretto?
 - Ci sono passi mancanti?
 - Ci sono passi errati?
 - Ci sono passi in ordine errato?
 - Ci sono passi ridondanti o inefficienti?
 - Analizzare con un commento il reasoning prodotto.

 - Analizzare la soluzione finale in linguaggio SQL del modello rispondendo alle seguenti domande:
 - La soluzione è corretta?
 - La soluzione segue il reasoning generato in precedenza?
 - Analizzare con un commento la soluzione prodotta.

 - Alla luce delle analisi del reasoning e della soluzione prodotta, proporre eventuali modifiche o integrazioni del prompt per mitigare gli errori e migliorare la risposta del modello. Se la soluzione del modello è già corretta, passare al punto 4. Riportare il nuovo prompt nel campo *Prompt aggiornato*.
 - Inviare al modello il prompt aggiornato.
 - Copiare l'output del modello nei campi *Reasoning (aggiornato)* (dove riportare i passaggi di ragionamento intermedi effettuati dal modello) e *Soluzione SQL (aggiornata)* (dove riportare la soluzione finale in linguaggio SQL prodotta dal modello).

 - Analizzare il reasoning e la soluzione finale in linguaggio SQL del modello ottenuta con il prompt aggiornato rispondendo alle seguenti domande:
 - Il reasoning (aggiornato) è corretto?
 - Analizzare con un commento il reasoning (aggiornato) prodotto.
 - La soluzione (aggiornata) in linguaggio SQL è corretta?
 - Analizzare con un commento la soluzione (aggiornata) in linguaggio SQL prodotta.
4. Confrontare la qualità del reasoning e delle soluzioni in linguaggio SQL dei tre Large Language Models utilizzando i campi *Confronto modelli (reasoning)* e *Confronto modelli (soluzione SQL)*.

Nota: Oltre all'utilizzo di modelli dotati di reasoning, per le query **Q4** e **Q10**, è anche richiesto di risolverle utilizzando ChatGPT disabilitando la funzionalità reasoning. In questo caso:

- Inviare al modello lo stesso prompt (eventualmente aggiornato) utilizzato in precedenza per lo svolgimento con ChatGPT nella variante con reasoning.

- Copiare l'output del modello nei campi *Output* (dove riportare la risposta in linguaggio naturale generata dal modello) e *Soluzione* (dove riportare la soluzione finale in linguaggio SQL prodotta dal modello).
- Analizzare l'output in linguaggio naturale e la soluzione finale in linguaggio SQL del modello rispondendo alle seguenti domande:
 - L'output in linguaggio naturale è corretto?
 - Analizzare con un commento l'output in linguaggio naturale prodotto.
 - La soluzione in linguaggio SQL è corretta?
 - Analizzare con un commento la soluzione in linguaggio SQL prodotta.
- Confrontare la qualità del reasoning e delle soluzioni in linguaggio SQL prodotte da ChatGPT con e senza funzionalità di reasoning abilitata utilizzando i campi *Confronto modelli GPT (reasoning)* e *Confronto modelli GPT (soluzione SQL)*.

Al fondo del form 5 (da svolgere a casa), per ogni LLM è presente un campo *Analisi finale* nel quale lo studente può riportare un commento generale sulle principali risultanze sperimentali evidenziate tra cui, ad esempio:

- Punti forti e punti deboli di ciascun LLM
- Confronti tra LLM

A ciascuno studente è richiesto di:

- Scrivere le soluzioni attese per le query proposte.
- Formulare i prompt testuali da inviare agli LLM.
Occorre partire dalla struttura del prompt fornito in calce e completarlo per ognuna delle query proposte. I prompt devono essere scritti in **lingua inglese** e devono includere lo schema logico delle tabelle e il testo in linguaggio naturale della query da risolvere.
- Interrogare i **tre Large Language Models** utilizzando i prompt generati.
- Inserire l'output fornito dai vari LLM all'interno dei campi di testo delle corrispondenti domande dei form.
- **Analizzare** gli output (reasoning e soluzione in SQL) forniti dai modelli rispondendo alle relative domande nei campi predisposti.

Lo svolgimento dell'esercitazione prevede **due fasi distinte**, ciascuna delle quali consente di ottenere i punti associati all'homework. Il punteggio totale massimo assegnabile al primo homework è **1 punto, diviso equamente tra le due fasi**.

Fase 1 – in laboratorio/aula e completamento a casa (0.5 punti)

- Seguendo le indicazioni degli esercitatori di laboratorio, collegarsi alle interfacce web di ciascun LLM e riportare nei form, per ciascuna query, la relativa soluzione attesa, prompt, output dei modelli (reasoning e soluzione in SQL) e analisi.
- Prima del termine dell'esercitazione, inviare ciascuno dei form adibiti al laboratorio compilati arrivando all'ultima pagina e premendo **INVIA**.
 - **Salvare** le risposte inviate cliccando sul pulsante "Salva la mia risposta per modificare" mostrato dopo l'invio del modulo. Prestare molta attenzione al salvataggio delle risposte, altrimenti non sarà possibile continuare la compilazione in un secondo momento.
 - Per continuare la compilazione, accedere a Microsoft Forms da forms.office.com. Si troverà il modulo inviato nella scheda "Recente" o "Moduli compilati".
- La consegna è ritenuta valida **se e solo se**:
 - I **form 1 e 2** sono stati inviati;
 - Contengono le soluzioni attese e i relativi prompt, output del modello (reasoning e soluzione in SQL) e analisi delle query **Q1, Q2 e Q3**, tratte dai **Temì d'esame 1 e 2**;
 - Data e orario di consegna non superano l'orario di conclusione del proprio slot di laboratorio.
- La **consegna dei form 1 e 2** può essere effettuata in modo **individuale** o a **coppie**.
 - **Individuale**: compilare i campi "Nome/Cognome/Matricola/Email (1)", lasciando vuoti i campi "Nome/Cognome/Matricola/Email (2)".
 - **A coppie**: compilare sia i campi "Nome/Cognome/Matricola/Email (1)" che i campi "Nome/Cognome/Matricola/Email (2)".
- **Completare i form 1 e 2** (contenenti le query **da Q1 a Q4**) a casa entro la scadenza prefissata qualora non già completati in laboratorio/aula **con la stessa modalità utilizzata in laboratorio/aula** (se svolti in modo individuale continuare in modo individuale, se svolti a coppie continuare a coppie).

Fase 2 – a casa (0.5 punti)

- Compilare i **form 3, 4 e 5** (contenenti le query **da Q5 a Q10**) adibiti allo svolgimento a casa inserendo anche i **commenti finali**.
- La **consegna dei form 3, 4 e 5** può essere effettuata esclusivamente in modo **individuale**.
- **Entro la scadenza prefissata** inviare i form 3, 4 e 5 compilati.
 - Scadenza: **26 maggio 2025 ore 23:59 CEST**
- La consegna è ritenuta valida **se e solo se**:
 - I **form 3, 4 e 5** sono stati inviati;
 - Contengono le soluzioni attese e i relativi prompt, output dei modelli (reasoning e soluzione in SQL) e analisi delle query **da Q5 a Q10**, tratte dai **Temì d'esame 3, 4 e 5**;
 - La data di consegna non supera la scadenza prefissata.

Piattaforma

Per svolgere l'homework, è necessario compilare i seguenti form, accessibili esclusivamente attraverso l'indirizzo email istituzionale fornito dal Politecnico (SMATRICOLA@studenti.polito.it):

- Form 1 e 2 (da svolgere in laboratorio/aula e da completare a casa):
 - Form 1 (Tema d'esame 1): <https://forms.cloud.microsoft/e/r4r38fM7Pt>
 - Form 2 (Tema d'esame 2): <https://forms.cloud.microsoft/e/kQqbXm9XNz>

Scadenza per l'invio: termine slot laboratorio

Scadenza per il completamento: **26 maggio 2025 ore 23:59 CEST**

- Form 3, 4 e 5 (da svolgere a casa):
 - Form 3 (Tema d'esame 3): <https://forms.office.com/e/agiZ9dKV4u>
 - Form 4 (Tema d'esame 4): <https://forms.office.com/e/4GUcYcgNVj>
 - Form 5 (Tema d'esame 5): <https://forms.office.com/e/VdCsebKRxD>

Scadenza per l'invio: **26 maggio 2025 ore 23:59 CEST**

In caso di problemi nell'accesso ai form, segnalarlo all'esercitatore presente durante lo svolgimento dell'esercitazione o inviando un'email all'indirizzo aurora.gensale@polito.it.

Per poter inviare e sottomettere correttamente le risposte, occorre arrivare all'ultima pagina e premere il pulsante **INVIA**.

Sottomissioni precedenti all'orario di inizio del proprio slot di laboratorio non saranno considerate.

Note

- Assicurarsi di **inviare e salvare** correttamente i form 1 e 2 al termine dell'esercitazione, altrimenti non sarà possibile completarli.
- Sottomissioni parziali saranno valutate con eventuali punteggi parziali.
- Sottomissioni largamente incomplete non riceveranno punti.

Esempio di prompt

You are an expert AI assistant for solving SQL language queries given a question in natural language.

I will provide in the following the database schema and the question in natural language.

Database schema:

PERSON(FiscalCode, Name, DateOfBirth, Gender)

Primary key: {FiscalCode}

TENNIS_CLUB(ClubCode, ClubName, Address, City)

Primary key: {ClubCode}

COURT_BOOKING(ClubCode, Date, Time, Court, FiscalCode)

Primary key: {ClubCode, Date, Time, Court}

Foreign keys: {ClubCode} references TENNIS_CLUB{ClubCode}; {FiscalCode} references PERSON{FiscalCode}

Question:

For the tennis clubs where more than 50 different female individuals booked in June 2009, display the club code, name, and city, the total number of different courts used, the total number of different individuals (including both males and females) who made bookings, and the total number of bookings in June 2009.

SQL code: