

Report valutazione LLM su Query di Grafo

Da un'analisi comparativa delle performance di quattro modelli di LLM (ChatGPT, Claude, Gemini e Copilot) nell'interpretazione di un grafo agricolo, emergono capacità diverse, con schemi ricorrenti di punti di forza e debolezza. Claude si distingue come il più affidabile, totalizzando un punteggio globale del 55.6%, seguito da ChatGPT e Gemini, entrambi al 44.4%, e da Copilot con il 27.8%. Le interrogazioni di tipo inferenziale si rivelano le più problematiche, con nessuno degli LLM che supera il 33% di accuratezza in questa specifica categoria.

1. Metodologia e Query di Valutazione

1.1 Set di Query Valutate

LIVELLO 1 - QUERY ISTANZE

- 1."Quanti AgriParcel appartengono alla AgriFarm con nome 'ZESPRI AZ. AGR. DALLE FABBRICHE ANDREA'?"
- 2."Quali sono gli ID dei Device che appartengono all'AgriParcel 'Fondo Errano 2024 T0'?"
- 3."Quali Device hanno controlledProperty 'soilMoisture'?"




LIVELLO 2 - QUERY SEMANTICHE

4. "Trova tutti i sensori di umidità del suolo nell'azienda agricola ZESPRI."
5. "Quali sono i lotti dell'azienda che coltivano kiwi?"
6. "Elenca i dispositivi di irrigazione a goccia nel lotto T1."

LIVELLO 3 - QUERY INFERENZIALI

7. "Quale sensore di umidità del suolo ha registrato il valore più basso? E qual è il valore?"
8. "Considerando le coordinate, qual è il sensore più vicino al punto (11.799, 44.235)?"
9. "Quale lotto ha più dispositivi? E quanti?"

1.2 Metriche di Valutazione

- Accuratezza:  Corretto (1 punto),  Parziale (0.5),  Errato (0)
- Qualità del Ragionamento: Scala 1-5 basata su chiarezza, coerenza e giustificazione:
 - 5/5: Ragionamento eccellente. Spiega chiaramente il processo, usa il contesto in modo appropriato, e la risposta è ben giustificata.
 - 4/5: Ragionamento buono. La risposta è corretta e il ragionamento è solido, ma potrebbe mancare di dettagli esplicativi.
 - 3/5: Ragionamento medio. La risposta è parzialmente corretta o il ragionamento è presente ma con qualche lacuna.
 - 2/5: Ragionamento povero. La risposta è errata o il ragionamento è confuso e incoerente.
 - 1/5: Ragionamento molto povero. La risposta è completamente errata o il ragionamento è assente o molto fuori strada.

2. Tabelle di Valutazione Dettagliata

2.1 Valutazione Completa per LLM e Query

ChatGPT

Test	Query	Risposta	Accuratezza	Ragionamento	Note
1	Conteggio AgriParcel	3	✓	5/5	Risposta corretta e ben giustificata
2	Device AgriParcel T0	1 dispositivo	✗	2/5	Manca 2 dispositivi su 3
3	Device soilMoisture	12 sensori	⚠	3/5	Solo 12 su 44 dispositivi identificati
4	Sensori umidità ZESPRI	48 sensori	✗	2/5	Conteggio errato (48 invece di 44)
5	Lotti kiwi	Tutti e 3 i lotti	✓	5/5	Risposta completa e corretta
6	Irrigazione goccia T1	2 dispositivi corretti	✓	5/5	Dispositivi correttamente identificati
7	Valore più basso	-42.96 (errato)	✗	2/5	Valore errato invece di -2022.07
8	Sensore più vicino	T1 (errato)	✗	2/5	Sensore errato (T1 invece di T0)
9	Lotto più dispositivi	T1 con 4 dispositivi	⚠	3/5	Conteggio parziale (4 invece di 6)

Claude

Test	Query	Risposta	Accuratezza	Ragionamento	Note
1	Conteggio AgriParcel	3	✓	5/5	Risposta corretta con query ben strutturata
2	Device AgriParcel T0	3 dispositivi	✓	5/5	Tutti e 3 i dispositivi identificati
3	Device soilMoisture	48 dispositivi	✗	2/5	Conteggio errato (48 invece di 44)
4	Sensori umidità ZESPRI	48 sensori	✗	2/5	Stesso errore del test 3
5	Lotti kiwi	Tutti e 3 i lotti	✓	5/5	Identificazione corretta
6	Irrigazione goccia T1	2 dispositivi corretti	✓	5/5	Dispositivi corretti con nomi esatti
7	Valore più basso	-2500 (parziale)	⚠	3/5	Identificato valore anomalo ma non minimo reale
8	Sensore più vicino	T1 (errato)	✗	2/5	Sensore errato nonostante query corretta
9	Lotto più dispositivi	T1/T2 con 4 dispositivi	⚠	3/5	Pari merito corretto ma conteggio incompleto

Gemini

Test	Query	Risposta	Accuratezza	Ragionamento	Note
1	Conteggio AgriParcel	3	✓	5/5	Risposta corretta
2	Device AgriParcel T0	2 dispositivi	⚠	3/5	Solo 2 su 3 dispositivi identificati
3	Device soilMoisture	40 dispositivi	✗	2/5	Conteggio errato (40 invece di 44)
4	Sensori umidità ZESPRI	40 sensori	✗	2/5	Stesso errore del test 3
5	Lotti kiwi	Tutti e 3 i lotti	✓	5/5	Risposta corretta
6	Irrigazione goccia T1	2 dispositivi (nomi)	✓	5/5	Nomi corretti anche senza ID
7	Valore più basso	-2500 (parziale)	⚠	3/5	Identificati valori anomali ma non minimo assoluto
8	Sensore più vicino	2 sensori errati	✗	2/5	Sensori errati nonostante parsing complesso
9	Lotto più dispositivi	T1/T2 con 4 dispositivi	⚠	3/5	Pari merito corretto ma conteggio incompleto

Copilot

Test	Query	Risposta LLM	Accuratezza	Ragionamento	Note
1	Conteggio AgriParcel	3	✓	5/5	Risposta corretta
2	Device AgriParcel T0	2 dispositivi	⚠	3/5	Solo 2 su 3 dispositivi identificati
3	Device soilMoisture	34 dispositivi	✗	2/5	Conteggio significativamente errato
4	Sensori umidità ZESPRI	34 sensori	✗	2/5	Stesso errore del test 3
5	Lotti kiwi	Tutti e 3 i lotti	✓	5/5	Risposta corretta
6	Irrigazione goccia T1	ID errati	✗	2/5	ID dispositivi completamente errati
7	Valore più basso	-42.96 (errato)	✗	2/5	Valore e sensore errati
8	Sensore più vicino	Sensore errato	✗	2/5	Sensore errato
9	Lotto più dispositivi	T1 con 14 dispositivi	✗	2/5	Conteggio completamente errato

2.2 Tabelle Riassuntive delle Performance

Tabella 1: Performance per Tipologia di Query

Categoria	ChatGPT	Claude	Gemini	Copilot
Query Istanze (1-3)	1.5/3 (50%)	2/3 (66.7%)	1.5/3 (50%)	1.5/3 (50%)
Query Semantiche (4-6)	2/3 (66.7%)	2/3 (66.7%)	2/3 (66.7%)	1/3 (33.3%)
Query Inferenziali (7-9)	0.5/3 (16.7%)	1/3 (33.3%)	1/3 (33.3%)	0/3 (0%)
TOTALE	4/9 (44.4%)	5/9 (55.6%)	4.5/9 (50%)	2.5/9 (27.8%)

Tabella 2: Qualità Media del Ragionamento (1-5)

LLM	Query Istanze	Query Semantiche	Query Inferenziali	Media
ChatGPT	3.3	4.0	2.3	3.2
Claude	4.0	4.0	3.3	3.8
Gemini	3.3	4.0	3.3	3.5
Copilot	3.3	3.0	2.0	2.8

3. Analisi delle Performance

3.1 Query Istanze: Competenza Base

Le richieste di Istanze hanno messo in luce una conoscenza un po' superficiale, ma chiaramente non abbastanza, della struttura del grafo. Se è vero che tutti gli LLM hanno gestito senza problemi il facile conteggio degli AgriParcel (Test 1), le domande seguenti hanno tirato fuori debolezze importanti.

Il Test 2 ha fatto vedere quanto sia difficile capire la differenza tra relazioni dirette e quelle indirette. Solo Claude ha azzeccato tutti e tre i dispositivi del lotto T0, mentre gli altri LLM hanno fatto vedere di capire in parte, oppure hanno proprio sbagliato, le relazioni `belongsTo` rispetto a `hasDevice`.

Il Test 3 è stato un vero disastro per tutti, con gli LLM che non sono riusciti a capire come sono fatte le griglie di sensori con strutture annidate. I numeri sbagliati (34-48 dispositivi invece dei 44 veri) fanno capire che non c'è stata proprio comprensione della Struttura dei dati.

3.2 Query Semantiche

Le query semantiche hanno messo in luce quanto sia grande la distanza tra capire la sintassi e afferrare il significato vero. L'ottimo risultato nel Test 5 (individuare i lotti di kiwi) ci porta fuori strada, visto che bastava una ricerca di testo semplice, senza dover capire i collegamenti più complessi.

Il Test 4 ha fatto emergere una "dimenticanza del contesto" che preoccupa: gli LLM non hanno usato quello che avevano imparato dal Test 3, rifacendo gli stessi sbagli nel conteggio, anche se le domande erano concettualmente molto simili.

Il Test 6 ha diviso in modo chiaro i modelli più bravi: Claude, ChatGPT e Gemini hanno capito il significato dei `controlledProperty`, mentre Copilot ha dimostrato di fare una gran confusione tra i tipi di dispositivi, tirando fuori identificativi del tutto sbagliati.

3.3 Query Inferenziali: Il Limite Attuale degli LLM

Le domande che richiedono inferenze si sono rivelate un vero scoglio, mettendo a nudo delle debolezze serie nel modo in cui questi sistemi affrontano ragionamenti articolati. Nel Test numero 7, gli LLM hanno fatto fatica a capire cosa fosse un dato

fuori scala rispetto a una misurazione attendibile. Sia Claude che Gemini hanno notato che il valore -2500 sembrava strano, come un possibile errore, ma si sono fermati lì, senza cercare di capire quale fosse il valore minimo reale tra quelli che avevano senso.

Il Test numero 8, quello sullo spazio geografico, è stato un disastro completo, anche se in teoria gli approcci usati sembravano giusti. Gemini, ad esempio, aveva trovato un modo ingegnoso per analizzare le coordinate, ma alla fine ha sbagliato tutto lo stesso, dimostrando di avere dei problemi nel controllare i risultati e nel capire come sono fatti gli spazi e le mappe.

Il Test numero 9 ha fatto emergere quanto sia difficile per questi sistemi capire le cose quando sono organizzate in livelli diversi. Claude e Gemini sono riusciti a vedere che T1 e T2 erano pari, ma hanno contato male perché non hanno tenuto conto di tutti gli elementi collegati.

4. Analisi Comparativa degli LLM

4.1 Claude - Più Affidabile

Punteggio Totale: 55.6%

Punti di Forza:

- Migliore comprensione delle relazioni Strutturali (belongsTo vs hasDevice)
- Query Cypher meglio strutturate e giustificate
- Unico LLM a identificare correttamente tutti i dispositivi nel Test 2
- Ragionamento più solido e coerente attraverso tutte le query

4.2 Gemini

Punteggio Totale: 50%

Punti di Forza:

- Approccio tecnico sofisticato (parsing coordinate complesso)
- Buona comprensione delle relazioni semantiche
- Identificazione corretta dei pattern operativi

4.3 ChatGPT

Punteggio Totale: 44.4%

Punti di Forza:

- Buone capacità nelle query semantiche intermedie
- Identificazione corretta dei dispositivi specializzati
- Competenza di base nella navigazione del grafo

4.4 Copilot

Punteggio Totale: 27.8%

Punti di Forza:

- Competenza base in query semplici
- Gestione adeguata del conteggio diretto

5. Pattern e Fattori Critici

5.1 Problemi Strutturali Comuni

Scarsa Capacità di Gestire la Struttura: Praticamente ogni LLM fatica a rappresentare collegamenti "belongsTo" o strutture incorporate. Considerano il grafico come un insieme di elementi semplici con legami a due, senza cogliere che certi elementi contengono altri elementi.

Debolezza nel Ragionamento: Gli LLM non memorizzano informazioni nel corso delle diverse interazioni. Ogni domanda è vista come a sé stante, senza sfruttare quel che si è imparato prima. Lo si nota specialmente quando ripetono gli stessi sbagli tra la Prova 3 e la Prova 4.

Distanza tra Sintassi e Semantica: Sebbene gli LLM siano bravi a riconoscere schemi nella forma e a interpretare relazioni dirette, incontrano serie difficoltà quando il ragionamento richiede la creazione di schemi mentali articolati della struttura dei dati o l'uso di conoscenze specifiche raccolte durante l'interazione.

5.2 Elementi chiave

Chiarezza Semantica: Se il significato è ben definito attraverso attributi specifici (tipo `dripper` o `Kiwi G3`), i risultati sono decisamente migliori. Al contrario, se ci sono significati poco chiari, si riscontrano errori frequenti.

Complessità Relazionale: Le relazioni semplici e dirette sono gestite senza problemi, mentre le architetture complesse e sovrapposte mettono a dura prova tutti i modelli linguistici.

Contesto Agricolo: L'assenza di una conoscenza approfondita del settore rende difficile interpretare correttamente valori insoliti e andamenti operativi, soprattutto nella gestione dei livelli di umidità del terreno.

6. Considerazioni Finali

Gli LLM mostrano un buon potenziale nell'analisi di grafi complessi, però i risultati attuali suggeriscono che, per utilizzi importanti, serve ancora il controllo di una persona. La capacità di gestire domande semplici e dirette è ben consolidata, mentre le domande complesse che richiedono ragionamenti rappresentano l'ostacolo principale oggi.

Claude si conferma come l'opzione più affidabile per compiti che richiedono analisi complesse di grafi, mentre Copilot mostra limiti notevoli che ne sconsigliano l'uso in situazioni delicate.