Quanto sono informati i LLM?

Uno studio intrapreso da Kai Sun, Yifan Ethan Xu, Hanwen Zha, Yue Liu e Xin Luna Dong, affronta tematiche nel contesto delle recenti evoluzioni nel campo dei Large Language Models (LLMs). Questi modelli linguistici negli ultimi tempi, hanno suscitato numerose discussioni e domande sulla loro effettiva precisione ed accuratezza delle risposte da loro fornite. Inoltre viene valutata un'eventuale sostituzione dei Knowledge Graphs (KGs) i quali conservano la loro conoscenza in forma simbolica.
Nel corso della relazione analizzeremo gli obiettivi posti dagli autori, il metodo utilizzato ed eventuali conclusioni.

I principali obiettivi posti sono:

- **Valutare la conoscenza dei Large Language Models**: Gli autori desiderano conoscere quanto i LLMs siano in grado di comprendere e fornire risposte accurate.
- **Come variano le performance di un LLMs**: Gli autori si sono interrogati se esistevano variazioni nelle capacità delle LLMs, fornendogli quesiti basati su popolarità differenti.
- **Quanto sono più efficaci i LLMs più grandi**: Gli autori hanno effettuato diversi test su più LLMs, così da poter notare quanto effettivamente è la differenza di accuratezza tra i diversi LLMs.

Per raggiungere questi obiettivi gli autori hanno utilizzato una metodologia di ricerca, composta da:

1. **Creazione del Benchmark Head-to-tail**: È stato sviluppato un benchmark composto da 18.000 domande-risposte che ricoprono diversi domini. In particolare il benchmark va da argomenti di maggior popolarità (head) fino ai meno popolari (tail).
2. **Definizione di metriche**: Gli autori hanno introdotto metriche specifiche per valutare le performance dei LLMs.
3. **Valutazione dei LLMs**: La valutazione è stata presa coinvolgendo 14 diversi tipi di LLMs diversi e confrontando i vari risultati ottenuti.

Gli autori hanno valutato la capacità dei LLMs di fornire risposte accurate a domande fattuali. La loro analisi ha rivelato che la percentuale di risposte corrette da parte dei LLMs è stata sorprendentemente bassa. In particolare, Chat GPT e LLaMA-33B, che rappresentano alcuni dei modelli di punta, hanno fornito risposte corrette solo per circa il 20% delle domande. Questo indica che nonostante le capacità impressionanti dei LLMs, la loro conoscenza di dettagli specifici e fattuali è ancora limitata.
L'insieme delle domande-risposte è stato suddiviso, come detto in precedenza, in Head, Torso e Tail in base alla loro popolarità. Si è osservato come la precisione delle risposte fornite dai LLMs ha una tendenza decrescente a seconda della popolarità della domanda. inoltre anche per le domande contenute nella categoria Head, quindi le più popolari, la percentuale di correttezza delle risposte è pari a circa il 51% per Chat GPT. Gli autori sono quindi arrivati alla conclusione, dopo una serie di ottimizzazioni attraverso istruzioni

specifiche, che il semplice aumento delle dimensioni del modello, non si traduce in un aumento della correttezza delle risposte.

Gli autori sono giunti infine alle seguenti conclusioni:
Uno dei risultati più significativi è stato il rivelarsi dei limiti della conoscenza dei LLMs. Nonostante la loro enorme conoscenza dei dati, questi modelli hanno dimostrato di non essere fonti pienamente affidabili, soprattutto per informazioni meno popolari e specifiche. Quest'ultimo punto apre un'altra conclusione, a seconda della popolarità dei dati, il LLMs ha risultati differenti, questo probabilmente poiché durante l'addestramento, sono stati forniti principalmente dati di maggior popolarità poiché reperibili con più facilità. Potrebbe essere necessario quindi includere nell'addestramento anche dati di popolarità minori nei dataset di addestramento.

In futuro una prospettiva interessante è l'integrazione di Knowledge Graphs (KGs) tradizionali con Dual Neural KGs. Questa combinazione potrebbe sfruttare la rappresentazione simbolica e quella neurale per migliorare la comprensione delle informazioni. La coesistenza delle due forme di conoscenza potrebbe portare a un'implementazione più efficace dei dati. Inoltre bisognerà pensare a come arricchire i dati di addestramento con informazioni più dettagliate e meno popolari, magari con strategie di addestramento mirate. Potrebbe essere una soluzione sviluppare dei LLMs specifici, che integrano dati di addestramento più raffinati, questi modelli potrebbero collaborare in modo sinergico in modo da fornire risposte più accurate.