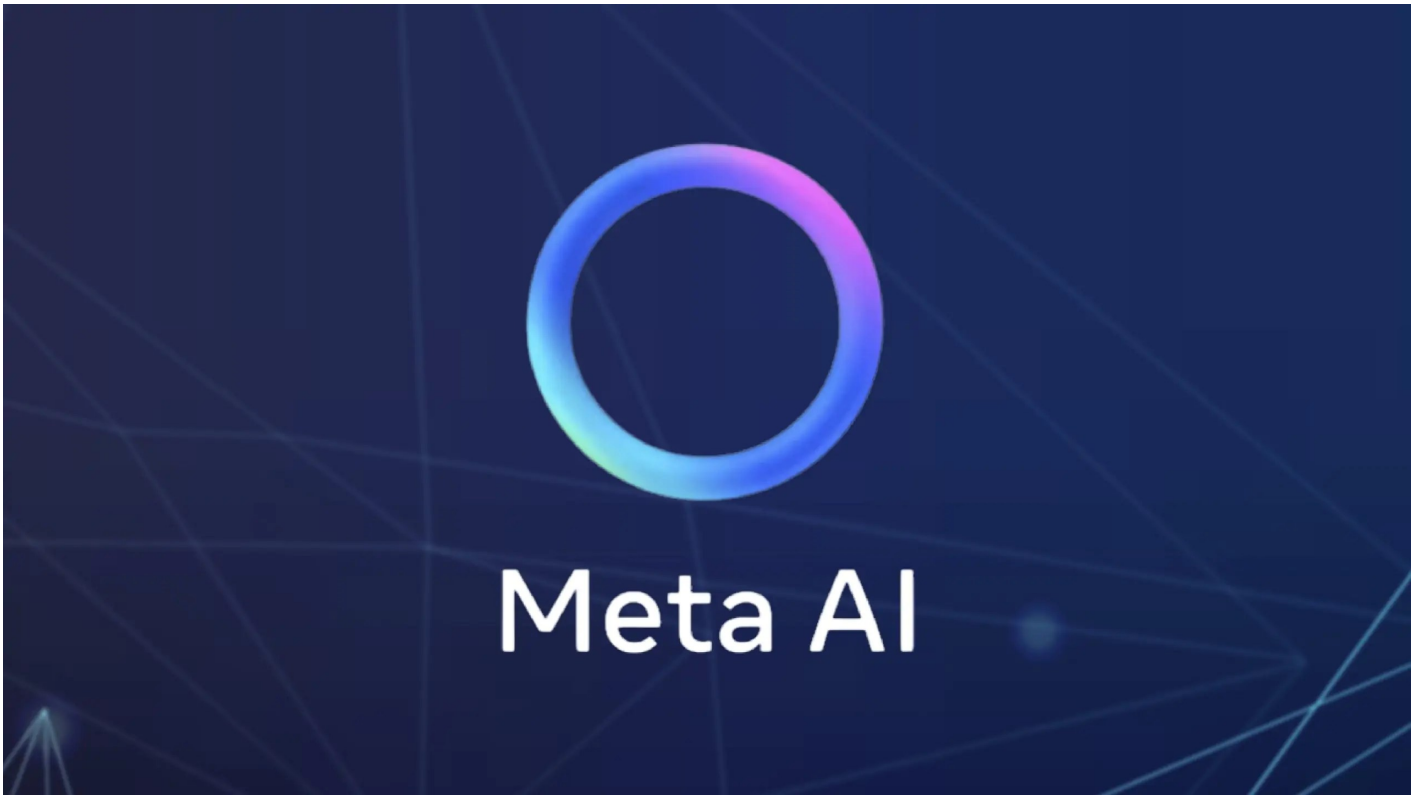
Addio all’errore umano? Meta presenta l’IA che si auto-valuta

Autore: Redazione Innovation Island

Data: 22 Ottobre 2024



Il team di ricerca sull’intelligenza artificiale di **Meta** (FAIR) ha rilasciato un nuovo strumento chiamato “**Self-Taught Evaluator**“, in grado di verificare l’accuratezza di altri modelli di IA senza bisogno di intervento umano. Questa innovazione potrebbe aprire la strada a un processo di sviluppo dell’IA meno dipendente dall’intervento umano.

Il metodo “Chain of Thought”

Il concetto di Self-Taught Evaluator si basa sul metodo “**chain of thought**” (catena di pensiero), lo stesso impiegato da OpenAI per il suo modello più recente, **o1.** Questo metodo consente all’IA di scomporre domande complesse in passaggi logici più piccoli, “pensando” prima di fornire risposte più accurate e affidabili.

Ecco come funziona esattamente:

Principio di base

Il CoT consente al modello di scomporre un problema in passaggi logici più piccoli, facilitando un processo di ragionamento più chiaro e strutturato. Questo approccio è stato introdotto da Wei et al. nel 2022 e si è dimostrato efficace in vari contesti, tra cui matematica e ragionamento.

Attivazione del metodo

Per attivare il CoT, si include nel prompt una frase che invita il modello a “pensare passo dopo passo”. Ad esempio, se si pone una domanda come: “Se oggi è lunedì, che giorno sarà tra tre giorni?”, il prompt potrebbe essere formulato come: Prompt: “Se oggi è lunedì, che giorno sarà tra tre giorni? Pensiamo passo dopo passo.”
Esecuzione del ragionamento

Il modello risponde decomponendo il problema in fasi:

- “Oggi è lunedì”.
- “Il giorno successivo a lunedì è martedì”.
- “Due giorni dopo lunedì è mercoledì”.
- “Tre giorni dopo lunedì è giovedì”.

Questa sequenza di pensieri consente al modello di arrivare a una risposta più accurata e coerente.

Vantaggi del CoT

- Miglioramento delle prestazioni: Il metodo ha dimostrato di migliorare significativamente le prestazioni dei modelli in compiti complessi rispetto ai metodi tradizionali;
- Riduzione degli errori: Scomponendo i problemi, il CoT aiuta a ridurre gli errori legati all’interpretazione errata delle domande;
- Versatilità: Può essere applicato a vari tipi di problemi, rendendolo utile in diversi contesti, dalla matematica alla programmazione.

Varianti del metodo

Esistono diverse varianti del CoT che ne ampliano l’efficacia:

- Auto-CoT: Genera automaticamente catene di pensiero utilizzando un approccio zero-shot;
- Plan-and-Solve Prompting: Combina la pianificazione della soluzione con l’esecuzione passo dopo passo;
- Recursion-of-Thought: Affronta problemi complessi inviando sotto-problemi a un modello separato per la risoluzione.
-

Addestramento basato su dati generati dall’IA

L’aspetto rivoluzionario è che Meta ha addestrato il modello esclusivamente su dati generati dall’IA, eliminando la necessità di intervento umano. I risultati mostrano che Self-Taught Evaluator ha prestazioni superiori rispetto ai modelli che si basano su dati etichettati da esseri umani, come ad esempio GPT-4.

Verso modelli di IA auto-miglioranti

Questa scoperta potrebbe portare allo sviluppo di modelli di IA in grado di auto-migliorarsi, riducendo la dipendenza da processi umani costosi e inefficienti. Attualmente, esperti umani devono etichettare correttamente i dati e verificare manualmente le risposte, con il rischio di errori e inaccuratezze.

Riferimento articolo: https://innovationisland.it/meta-self-taught-evaluator-ia/

Generato il 16/04/2025