

DeepSeek rilascia DeepInfer: il software open source per ridurre i costi dell'AI

DeepSeek rende disponibile **DeepInfer**, il framework interno utilizzato per servire **DeepSeek-V4.1-Flash** su larga scala.

Il software è stato pubblicato come open source e, per il momento, supporta soltanto V4.1-Flash. L'azienda ha però annunciato che in futuro verrà esteso anche ad altri modelli.

“Condividiamo il nostro stack di produzione per aiutare la comunità a ridurre i costi di calcolo, contenere i consumi energetici e rendere l'AI più accessibile”, ha spiegato DeepSeek.

Come funziona DeepInfer

DeepInfer interviene nella fase di inferenza, cioè quella in cui un modello già addestrato genera le risposte agli utenti. Il framework combina tecniche di inferenza distribuita, ottimizzazione dei carichi di lavoro e gestione più efficiente dell'hardware, con l'obiettivo di produrre più output in meno tempo e sfruttare meglio la capacità dei chip.

Per aziende, provider cloud e società di hosting, questo può tradursi in costi inferiori nell'esecuzione dei modelli DeepSeek sulle proprie infrastrutture. Restano però da verificare con benchmark indipendenti i reali vantaggi in termini di efficienza.

DeepSeek e il nuovo scontro tra

Stati Uniti e Cina

Il rilascio di DeepInfer arriva inoltre in una fase di crescente tensione tecnologica tra Washington e Pechino. Il governo statunitense ha recentemente vietato l'utilizzo dei modelli DeepSeek all'interno dei sistemi informatici delle agenzie federali civili. Il Dipartimento della sicurezza interna ha inoltre vietato alle agenzie di acquistare prodotti DeepSeek o servizi di terze parti sviluppati utilizzando il modello cinese.

Il ministero del Commercio di Pechino ha criticato la decisione, sostenendo che le misure statunitensi danneggino gli interessi delle aziende cinesi e chiedendo a Washington di revocare quelle che ha definito restrizioni discriminatorie. Lo scontro si è esteso anche alla proprietà intellettuale. Il governo cinese ha contestato le sanzioni statunitensi contro cinque entità accusate di essere coinvolte nel furto di segreti commerciali nel settore dell'intelligenza artificiale.

Tra le società colpite figura **Shanghai Zhisuan Technology**, fondata dall'ex ingegnere informatico di Google Linwei Ding, insieme a una società di venture capital che l'ha finanziata e ad altre entità collegate. Pechino ha respinto le accuse, sostenendo che gli Stati Uniti stiano enfatizzando il tema del furto di segreti commerciali per colpire le aziende cinesi.

Novità su Google, per aggiungere Key4Biz tra le tue fonti preferite, clicca qui

Leggi le altre notizie sull'home page di Key4biz

<https://www.key4biz.it/deepseek-rilascia-deepinfer-il-software-open-source-per-ridurre-i-costi-dellai/591606/>