una nuova piattaforma che sembra… rivoluzionaria.

In pratica è una azienda che che capito un trucco segreto per far funzionare un modello di linguaggio (come GPT) a una velocità estremamente superiore, grazie a una architettura che chiamano LPU.

Cos’è una LPU?

🟢 Sta per “Language Processing Unit”

🟢 Permette una velocità di scrittura di 500 Token/s (!)

🟢 Si può provare con l’ultimo modello di Mistral

🟢 Mixtral 54B, una delle AI più potenti (compete con GPT4)

🟢 GPT4 che dentro ChatGPT va a….15 Token/s