RATIO QUOTIDIANO
Il commento e la critica delle più importanti notizie del momento

CONSULENZA AZIENDALE, COMMERCIALE E MARKETING

di ALESSANDRO MATTAVELLI

La temperatura nei modelli linguistici: trovare il giusto equilibrio

La temperatura è un parametro fondamentale negli LLM (Large Language Model) che controlla la casualità e la creatività delle risposte generate. Regolare correttamente la temperatura è essenziale per ottenere output appropriati ed evitare problemi come allucinazioni o banalità.

Cos'è la temperatura - In termini tecnici, la temperatura regola la distribuzione di **probabilità** da cui le parole vengono campionate ad ogni step durante la generazione del testo. Temperature basse rendono l'output più deterministico e conservativo, mentre temperature alte lo rendono più creativo e imprevedibile.
Ad esempio, con una temperatura di 0.2, l'AI sceglierà principalmente le parole con più alta probabilità, risultando in risposte ripetitive e prevedibili. Al contrario, con una temperatura di 1.5, l'AI esplorerà combinazioni di parole improbabili, generando output molto creativi e originali, ma potenzialmente incoerenti.

Trovare il valore ottimale - La scelta della temperatura ottimale dipende dal compito specifico e dalle preferenze dell'utente. Per compiti che richiedono risposte precise e fattuali, come rispondere a domande o riassumere testi, sono più adatte temperature basse intorno a 0.2-0.4. Questo assicura che l'AI si attenga strettamente ai fatti e generi risposte prevedibili.
Per conversazioni generiche o compiti che richiedono un po' di varietà, una temperatura media di circa 0.7 è spesso un buon punto di partenza. Questo bilancia probabilità e casualità, risultando in risposte pertinenti e coerenti ma non troppo ripetitive.
Infine, per compiti altamente creativi come *storytelling,* poesia o *brainstorming* di idee, temperature più alte intorno a 1-1.5 possono essere appropriate. Questo incoraggia l'AI a esplorare combinazioni insolite di parole e concetti, portando a output molto originali e sorprendenti.

I pericoli degli estremi - Tuttavia, bisogna stare attenti a non esagerare con temperature troppo alte o troppo basse. Temperature eccessivamente basse possono portare a risposte banali e ripetitive che non aggiungono molto valore. L'AI semplicemente ripeterà le frasi più probabili senza introdurre nuove informazioni o prospettive.
D'altra parte, temperature troppo alte spesso causano *"allucinazioni",* ovvero affermazioni incoerenti, illogiche o fattuali errate, generando una sorta di delirio creativo sia nel tipo di risposte che nei termini utilizzati. L'AI genererà liberamente combinazioni improbabili di parole senza curarsi della coerenza o veridicità. Questo può essere problematico per applicazioni che richiedono accuratezza e affidabilità.

Sperimentare e iterare - La regolazione ottimale della temperatura richiede spesso un processo di sperimentazione e iterazione. È una buona idea testare diversi valori e valutare la qualità dell'output per il compito specifico. Con un po' di pratica, diventa più facile intuire quale valore di temperatura funzionerà meglio.
Inoltre, è importante tenere presente che la temperatura ottimale può variare anche all'interno dello stesso compito o conversazione. Ad esempio, si potrebbe usare una temperatura più bassa per le parti fattuali di una storia e una temperatura più alta per le parti creative e descrittive.

Conclusione - In conclusione, la temperatura è un parametro cruciale da considerare quando si utilizzano modelli linguistici. Regolarla in modo appropriato è fondamentale per ottenere output di alta qualità ed evitare problemi come risposte banali o allucinazioni.
Non ci credete? Provate a copiare questo articolo in un qualsiasi Gpt e chiedete di riscriverlo con temperatura superiore a 1, ne vedrete delle belle o forse riuscirete a riscriverlo in maniera più creativa!

Riproduzione vietata - venerdì 26 luglio 2024