Aplicaciones LLM

Latencia y velocidad de las LLM Apps

© 2023 Julio Colomer, Aceleradora AI

Latencia: velocidad de la LLM App

- Clave: ¿puedes permitirte que el usuario no tenga una experiencia rápida?
  - Normalmente puedes permitírtelo si los usuarios son los empleados de una empresa.
  - Normalmente NO puedes permitírtelo si los usuarios son clientes.

- Apps que necesitan low latency (alta velocidad):
  - conversational agents, virtual agents y chatbots.
  - content personalization y recommendation systems.

- Apps que pueden funcionar bien con high latency (baja velocidad):
  - research (legal, market, etc).
  - creative writing y content generation.

© 2023 Julio Colomer, Aceleradora AI