LLM Apps

Latency and Speed of LLM Apps

Latency and Speed of LLM apps

- Key: Can you afford for the user not to have a fast experience?
  - You can usually afford it if the users are employees of a company.
  - You usually CANNOT afford it if the users are customers.

- Apps that need low latency (high speed):
  - Conversational agents, virtual agents, and chatbots.
  - Content personalization and recommendation systems.

- Apps that can work well with high latency (low speed):
  - Research (legal, market, etc).
  - Creative writing and content generation.