1. Khi mở rộng phạm vi của Chatbot, cần promt lớn hơn và logic hơn. Do đó khi query cần nhiều thời gian hơn. Trong khi API Endpoint time limit đang bị hardcode 30s.
2. Việc giữ kết nối HTTP liên tục sẽ gây tốn kém tài nguyên, vài khi đạt đến giới hạn time limit, kết nối sẽ bị ngắt. Từ đó gây trải nghiệm không tốt cho người dùng.

Ví dụ: người dùng thực hiện truy vấn phức tạp, app cần nhiều hơn 30s để trả lời 🡪 khí quá 30s, kết nối bị ngắt 🡪 trải nghiệm không tốt.

* Đổi sang Asyn HTTP: Async API cho phép hệ thống gửi và nhận dữ liệu theo cách không đồng bộ, giúp người dùng nhận phản hồi gần như ngay lập tức mà không phải chờ đợi các yêu cầu khác hoàn tất. Điều này làm cải thiện trải nghiệm người dùng và giảm độ trễ.

Cụ thể: Khi người dùng truy vấn, app sẽ trả về query\_id ngày lập tức. Sau đó worker lambda function sẽ truy vấn query\_id này từ database và thực hiện generate câu trả lời.

1. Quá trình xử lý truy vấn

A screen shot of a computer

Description automatically generated