SotaAgents Tooling

SotaAgents: Kế hoạch Tích Hợp Tool và MCP cho LLM Agents

Tổng quan

Trình bày kế hoạch tích hợp các tool và MCP vào SotaAgents để tăng cường khả năng của LLM Agents, đặc biệt là trong việc tìm kiếm thông tin, truy vấn dữ liệu và tương tác với các dịch vụ khác nhau. SotaAgents sẽ sử dụng kiến trúc ReAct (Reasoning and Acting) để phối hợp các tool/MCP một cách hiệu quả, thực hiện từng bước và quan sát kết quả cho đến khi đạt được mục tiêu cuối cùng.

Các Tool Hiện Có

- **Semantic Search:**
  - Sử dụng vector database (Qdrant) để tìm kiếm tài liệu liên quan và lọc metadata.
- **SQL Query:**
  - Sử dụng LLM để sinh câu truy vấn SQL.
  - Thực thi câu truy vấn trên database.
  - Trả về kết quả cho người dùng.
  - Hiện tại hỗ trợ:
    - Database: PostgreSQL, SQL Lite
    - Read Action Only

Các Tool Dự Kiến Tích Hợp

Các tool được phân loại theo mục đích sử dụng và sẽ được tích hợp theo các phương pháp khác nhau. Ưu tiên lựa chọn một tool đại diện cho mỗi category (trừ khi có nhu cầu đặc biệt).

Web Crawler

- Firecrawl
- Puppeteer
- Trafilatura
- Playwright
- Tavily

Lựa chọn ưu tiên: Cần đánh giá và chọn một crawler tốt nhất về hiệu năng, khả năng trích xuất thông tin và dễ tích hợp.

Search Engine

- Google Search
- Bing Search
- Brave Search
- Yandex
- Zhipu
- Baidu
- Kagi

Lựa chọn ưu tiên: Google Search (vì độ phổ biến và chất lượng tìm kiếm) hoặc có thể tích hợp thêm một search engine khác như Bing hoặc Brave cho đa dạng kết quả.

Map

- Google Map
- Apple Map
- Baidu Map

Lựa chọn ưu tiên: Google Map (vì độ phổ biến và dữ liệu phong phú).

Database

- MCP database toolbox (created by Google)
- Google BigQuery
- MySQL
- SQL Lite
- Neon
- PostgreSQL
- Clickhouse
- MongoDB
- ElasticSearch
- DuckDB

Lựa chọn ưu tiên: Cần xác định các database phổ biến mà người dùng mục tiêu sử dụng để ưu tiên tích hợp. Ví dụ:
MySQL, PostgreSQL: Các database quan hệ phổ biến.

MongoDB: Database NoSQL.
* Google BigQuery: Dành cho phân tích dữ liệu lớn.

File System

- Obsidian
- Google Drive
- Local File System
- Excel
- Notion

Lựa chọn ưu tiên: Google Drive (vì phổ biến và dễ tích hợp). Các lựa chọn khác tùy thuộc vào nhu cầu người dùng.

Communication Service

- Discord
- Slack
- Email

Lựa chọn ưu tiên: Tùy thuộc vào kênh giao tiếp mà người dùng mục tiêu sử dụng nhiều nhất.

GenAI Service

- Minimax
- Flux
- ElevenLabs

Lựa chọn ưu tiên: Cần đánh giá và chọn các dịch vụ GenAI có khả năng tốt nhất cho các tác vụ cụ thể (ví dụ: tạo văn bản, tạo giọng nói).

Phương Pháp Tích Hợp

Có hai phương pháp chính để tích hợp các tool:

1. **Tích hợp qua MCP (Managed Cloud Platform):**
    - Kết nối với các MCP server được build và host sẵn (ví dụ: Google Search, Web Crawler, Map).
    - Tự build và host MCP Server (ví dụ: Database, Communication Service,...).
    - Kết nối với server do client tự build và host (có thể build SotaAgents có options build sẵn và kết nối bằng OAuth) (ví dụ: Google Drive, Notion,...).

2. **Tool Calling:**
   Sử dụng tool calling để giảm số lượng tool gửi lên server. Phù hợp với các service define sẵn.
   ReAct Agent:
   Sử dụng kiến trúc ReAct để điều phối các tool/MCP.
   Agent thực hiện tool calling/MCP từng bước, quan sát kết quả và đưa ra quyết định tiếp theo.
   Cần define các tool/MCP được sử dụng trước khi chạy query Agent.

Lộ Trình Tích Hợp

1. **Ưu tiên:** Các tool cốt lõi như:
   - Semantic Search (Qdrant)
   - SQL Query (PostgreSQL, SQL Lite)
   - Google Search (qua MCP/Tool calling)
   - Google Drive (qua MCP)
2. **Tiếp theo:** Dựa trên phản hồi của người dùng và phân tích nhu cầu, tích hợp các tool khác, ưu tiên các tool có tính ứng dụng cao và dễ tích hợp.
3. **Liên tục:** Thường xuyên cập nhật và đánh giá hiệu quả của các tool đã tích hợp. Nghiên cứu và thử nghiệm các tool mới để cải thiện khả năng của SotaAgents.