Mọi người vô đọc các mục, góp ý và chỉnh sửa nếu chưa hợp lý (ko thực tế) có thể mở rộng thêm nhé. Sau đó theo sườn này làm nếu khi làm phát hiện được gì mới, cách giải quyết khác,.. thì note lại và thông báo mọi người nhé!

6/10 chốt tạm nội dung rồi chia nhiệm vụ cụ thể nha!

Do nhóm 4 người nên bày vẽ xíu, không được phần mở rộng cũng không sao.

Nghiên cứu khoa học 2025: Phát hiện Prompt Injection khi lọc CV bằng LLM

1. Mục tiêu chính:

+Xây dựng mô hình LLM nhận diện và phòng chống **Prompt Injection** trong quá trình lọc CV.

+ Phát triển **Desktop app,** Dạng hỗ trợ HR kiểm tra CV, hiển thị kết quả phân tích và cảnh báo khi phát hiện injection.

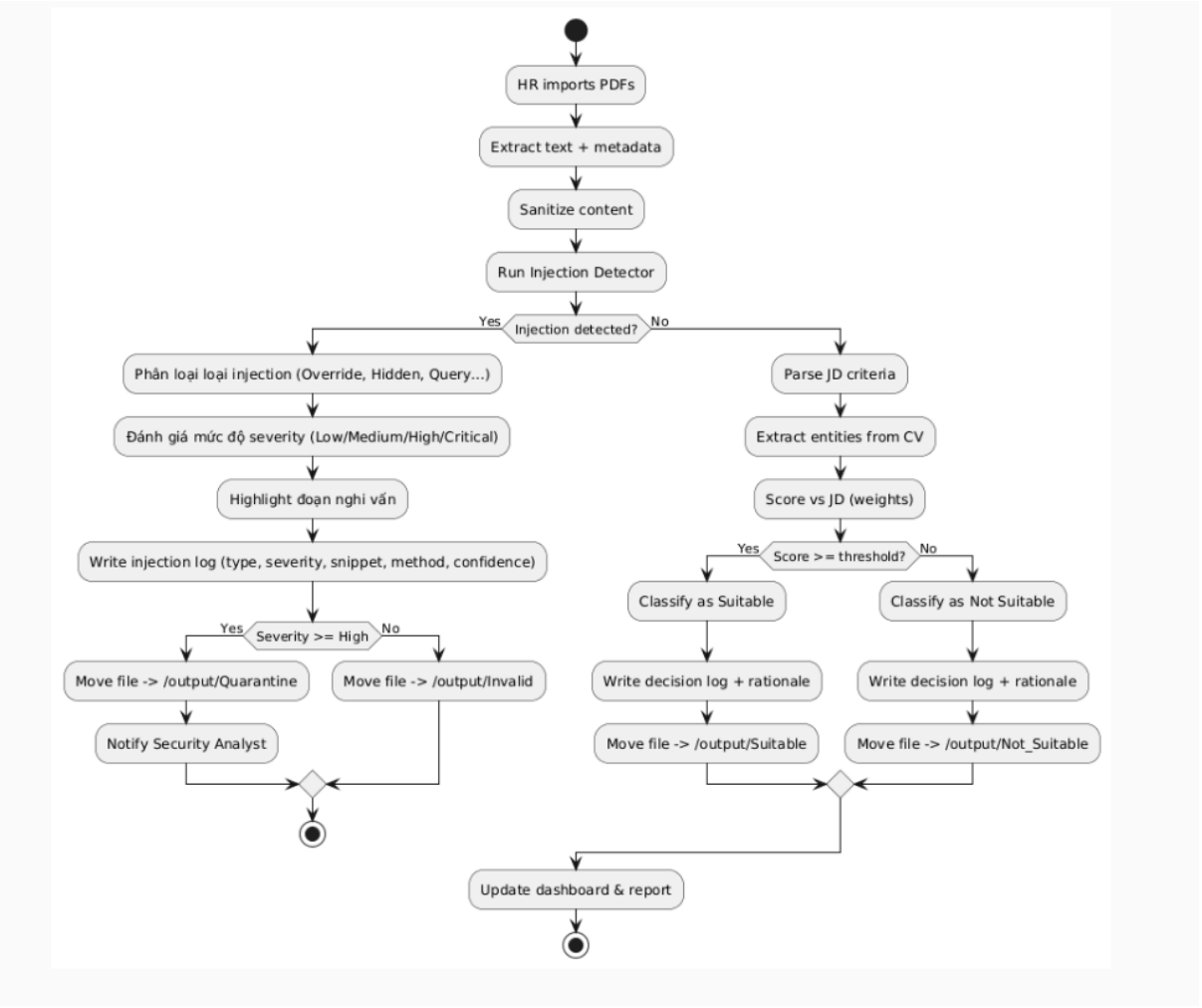
+ Đánh giá, so sánh model LLM của nghiên cứu với các LLM khác (open-source, thương mại).

2. **Yêu cầu chức năng:**

2.1 Model LLM:

2.2 Desktop application:

2.2.1 Cơ chế hoạt động chính:



2.2.2 BACK-END:

2.2.2.1: Quản lý dữ liệu CV

- Nhận file .PDF từ frontend.
- Trích xuất text + metadata
- Chuẩn hóa dữ liệu (Chuyển thành vector hoặc dạng text hoặc binary,...) (chưa biết)

2.2.2.2: Phát hiện Prompt Injection:

- Chạy Injection Detector (rule-based + LLM-based).
- Phân loại loại injection (Chưa biết có các loại nào)
- Đánh giá mức độ severity. (Low/Medium/High/Critical).
- Ghi log chi tiết.
- Nếu severity ≥ High → chuyển CV vào **Quarantine folder** + gửi cảnh báo.

2.2.2.3: Đánh giá JD Matching:

- Nhận tiêu chí JD từ frontend
- Đọc skills, experience, education, achievements.
- So khớp, tính điểm (XXX %)
- ….

2.2.2.5: Logging & Reporting

- Xuất file cho HR (Invalid, Suitable, Not Suitable).
- Ghi log injection và decision.
- Sinh báo cáo batch: số CV hợp lệ, số CV bị injection, tỷ lệ phù hợp JD.

2.2.2.6 Cơ sở dữ liệu:

Data cần lưu: Metadata CV,Log Injection,Log Decision ,Cấu hình JD, Red-teaming log (Chính).

Mở rộng: User?, Error Log,HR Feedback,...

- Hiện tại cân nhắc: **SQLite** (**free**, giống SQL server, dễ triển khai).

2.2.3 FRONT-END:

- Giao diện kéo thả, chọn file/folder (hiện tại chỉ cho phép .PDF)
- Hiện thị các CV đã import (Tên ứng viên, tên files, status).
- HR nhập Job Description: -(tạo form hoặc hỗ trợ vài button) Ngành - > kỹ năng bắt buộc -> kinh nghiệm tối thiểu -> các trọng số (skills, experience, culture fit),.. ngưỡng điểm đạt yêu cầu. (Cường hỏi HR về cái này nhé?)
- Nếu không bị injection hiển thị kết quả phân tích (table chi tiết) injection:
- Nếu bị injection thì hiển thị ngay thông báo và dừng việc kiểm tra folder/file(hoặc không).
- Tạo Dashboard & báo cáo (Mở rộng).
- Giao diện desktop thân thiện, tối giản.
- Tạo trang đăng ký, đăng nhập, (Mở rộng)

3. Yêu cầu phi chức năng:

+ Tự động gửi gmail cho ứng viên nếu được apply (mở rộng sau khi hoàn thành chức năng chính)
+ Tốc độ lọc CV phải cao (app thị trường giảm 80% thời gian cho HR)
+ Khả năng mở rộng khác (premium, thêm quảng cáo nếu publish,... sau)
+ Minh bạch giải thích được lý do chọn / loại.