tôi gửi bạn file code trước đó để xuất ra file a1\_text + metadate, bây giờ tôi cần clean để đưa vào vector store nên cần xuât file txt , metajson, json1 và output ở trong D:\1.TLAT\3. ChatBot\_project\1\_Insurance\_Strategy\outputs\a2\_text\_only\_prevector, nó tạo ra các folder con hay file tương ứng với ở input là: D:\1.TLAT\3. ChatBot\_project\1\_Insurance\_Strategy\outputs\a1\_text\_only\_outputs

- tôi gửi file mà nó chạy ra từ code a1 text trên:

\_ tôi gửi file code a2\_clean\_text hiện tại tôi có, bạn hãy xem tất cả data tôi gửi, và file code này rồi chỉnh cho tôi nhé.

- vẫn câu hỏi Y/N/A = Y (xóa file cũ, chạy mới)/ Skip / Append (chỉ xóa file, không xóa folder).

- chạy với lênh đơn giản là có thể in ra trang : python -m src.a2\_clean\_text\_only\_prevector --start 26 --end 28, thì in ra trang đó + metason+ jsonl, còn ko có start end thì in hết.

src/a1\_mixed\_runner.py — đọc hỗn hợp (TEXT + TABLE)

Giữ nguyên logic & hiệu năng hiện tại (đọc DOCX/PDF/IMG/TXT/Excel).

Tách TEXT vs TABLE, bảng đẩy qua GPT định dạng TSV “vector-ready”, có validator & chuẩn cột (financial/uw/kpi/client…) theo YAML bảng universal bạn gửi.

9ef1c91c-8f16-4298-b58b-520f27b…

Xuất: <name>\_text.txt, <name>\_meta.json, (tuỳ chọn) <name>\_vector.jsonl (gồm cả TEXT/TABLE, metadata đầy đủ).

Hai script đều hỏi chỉ 1 lần cho toàn bộ lượt chạy:

Câu 1 (gộp): “xóa file đầu ra nếu đã tồn tại? (Y/N/A)” — áp dụng cho tất cả định dạng nó sinh (\_text.txt, \_meta.json, \_vector.jsonl) trong toàn bộ cây thư mục chỉ định.

Câu 2: “Có dùng GPT cho lần chạy này không? (Y/N)” — và nếu bật Y thì nó cũng chỉ chạy với bảng thôi, ko chạy text văn bản.

Lần chạy đầu nếu thư mục OUTPUT chưa tồn tại thì mới tạo; những lần sau không xoá thư mục gốc, chỉ làm việc bên trong.

P1A \_GPT

Khuyến nghị pipeline tối ưu

OCR → văn bản thô (giữ để tham chiếu).

Tách bảng → PIPE (đặt trong <<<PIPE>>> … <<<END>>>).

GPT numbers-only (đối chiếu PNG) chỉ cho PIPE.

YAML validator/autofix (cha–con, tổng cộng) cho PIPE.

(Tuỳ chọn) Narrator tạo đoạn tóm tắt từ PIPE đã chuẩn.

Lưu:

\*\_text.txt: văn bản gốc + PIPE (và narrative nếu bật)

\*\_table.json: dữ liệu cấu trúc từng dòng (code, name, end, begin, unit…)

\*\_meta.json: company, period, statement, unit\_multiplier…

Nếu hiện tại bạn mới có đoạn văn OCR như ảnh (không có PIPE), mình khuyên bổ sung lưu PIPE. Narrative chỉ là “đọc cho dễ”, không thay thế PIPE/JSON/validator.

# STREAMLIT

1. THIẾT LẬP TRỤC TRUNG TÂM, X VÀ Y.
2. THIẾT LẬP MÀU NỀN CHÍNH
3. THIẾT LẬP KHỐI CHÍNH, Ở ĐÂY CHIA LÀM 2 ĐÓ LÀ KHỐI SIDEBAR BÊN TRÁI VÀ KHỐI PHẢI LÀ CHO LAYOUT CỦA PHẦN CHAT, CÓ LINE NGĂN GIỮA 2 KHỐI.
4. THANH SIDE BAR VÀ BÊN PHẢI GỒM CÁC ĐỐI TƯỢNG NHƯ GIAO DIỆN CHAT GPT. VÀ THIẾT LẬP ĐỂ CÁC ĐỐI TƯỢNG NÀY DỊCH CHUYỂN BẰNG VIỆC ĐIỀU CHỈNH TRUNG TÂM ĐỐI TƯỢNG SO VỚI TRỤC CHÍNH.
5. CHI TIẾT CỦA CÁC ĐỐI TƯỢNG DỊCH CHUYỂN SO VỚI CHÍNH TRỤC X/Y TRUNG TÂM CỦA CHÍNH NÓ.

- MÀU CỦA ĐỐI TƯỢNG. GỒM MÀU NỀN/ĐƯỜNG VIỀN/CHỮ.

- CHỈNH BO GÓC CỦA ĐỐI TƯỢNG.

**META**

* **VERSION: 1.0** → phiên bản cấu hình.
* **NOTES:** “Chỉnh mọi thứ ở YAML; code chỉ đọc & render.” → ghi chú để nhớ: không hard-code số trong code.

**CENTER\_AXIS (Thiết lập “trục trung tâm”)**

* **ENABLED: true** → bật cơ chế neo giao diện theo trục trung tâm.
* **MODE: viewport** → mốc căn giữa lấy theo **toàn màn hình** (viewport). container thì lấy theo khung giữa.
* **X\_PCT: 50** → **(X)** vị trí mốc dọc theo ngang, 0=trái, **50=giữa**, 100=phải.
* **Y\_PCT: 50** → **(Y)** vị trí mốc theo dọc, 0=đỉnh, **50=giữa**, 100=đáy.
* **SNAP\_TO\_GRID: false** → bật thì các đối tượng sẽ “hít” (snap) vào lưới.
* **GRID\_STEP\_PX: 8** → bước lưới (px) khi snap bật.

**COLORS (Bảng màu tổng)**

* **BG: "#FFFFFF"** → nền tổng thể app.
* **TEXT: "#0B1520"** → màu chữ mặc định.
* **CARD: "#FFFFFF"** → nền “khối/thẻ” (ví dụ panel chat).
* **LINE: "#E5E7EB"** → màu đường kẻ/đường phân cách.

**TYPOGRAPHY (Chữ)**

* **BASE\_SIZE\_PX: 15** → cỡ chữ gốc (làm base cho đơn vị rem).
* **TITLE\_SIZE\_PX: 20** → cỡ chữ tiêu đề chính (px).
* **FONT\_FAMILY:** … → danh sách font ưu tiên.

**LAYOUT (Khung chính & đường ngăn)**

* **MAIN\_MAX\_WIDTH\_PCT: 96** → bề rộng tối đa của **khung trung tâm** theo % viewport (căn giữa ngang).
* **GUTTER\_PX: 0** → rãnh/đệm 2 bên khung (px).
* **SIDEBAR\_WIDTH\_PCT: 30** → bề rộng sidebar theo % khung trung tâm.
* **CHAT\_WIDTH\_PCT: 82** → bề rộng cột chat theo % khung trung tâm.
* **DIVIDER.SHOW: true** → bật đường ngăn (ví dụ giữa sidebar & chat).
* **DIVIDER.COLOR: "#E5E7EB"** → màu đường ngăn.
* **DIVIDER.STYLE: "solid"** → kiểu nét: solid/dashed/dotted.
* **DIVIDER.WIDTH\_PX: 1** → độ dày đường ngăn (px).
* **DIVIDER.OFFSET\_X\_PX: 0** → **(X)** đẩy đường ngăn **trái/→ phải** (px).

**OBJECTS (Thiết lập từng đối tượng)**

**SIDEBAR**

* **ENABLED: true** → bật/tắt sidebar.
* **OFFSET\_X: "-28%"** → **(X)** dịch **trái/→ phải** (âm = sang trái, dương = sang phải).
* **OFFSET\_Y: "0px"** → **(Y)** dịch **lên/↓ xuống** (âm = lên, dương = xuống).
* **BG: "#0F172A"** → nền sidebar.
* **TEXT: "#E2E8F0"** → màu chữ sidebar.
* **BORDER\_COLOR: "#0F172A"** → màu viền sidebar.
* **BORDER\_PX: 0** → độ dày viền sidebar (px).
* **RADIUS\_PX: 18** → bo góc sidebar (px).
* **SHADOW: true** → có đổ bóng.
* **PADDING\_PX: 10** → đệm trong sidebar (px).

**CHAT\_PANEL (khu vực chat: lịch sử/nội dung)**

* **ENABLED: true** → bật phần chat panel.
* **OFFSET\_X: "22%"** → **(X)** dịch **trái/→ phải** (dương = sang phải).
* **OFFSET\_Y: "0px"** → **(Y)** dịch **lên/↓ xuống**.
* **BG: "#FFFFFF"** → nền chat panel (khung lịch sử).
* **TEXT: "#0B1520"** → màu chữ trong panel.
* **BORDER\_COLOR: "#FFFFFF"** → màu viền panel.
* **BORDER\_PX: 0** → độ dày viền panel.
* **RADIUS\_PX: 0** → bo góc panel.
* **SHADOW: false** → tắt bóng cho panel.
* **PADDING\_PX: 0** → đệm bên trong panel (px).
* **CHAT\_MAX\_WIDTH\_PX: 760** → **độ rộng cột giữa kiểu ChatGPT** (px).
* **CHAT\_HEIGHT\_VH\_DESKTOP: 72** → **(Y)** chiều cao vùng chat trên desktop theo **% màn hình** (vh).
* **CHAT\_HEIGHT\_VH\_MOBILE: 68** → **(Y)** chiều cao vùng chat trên mobile (vh).

**CHAT\_INPUT (ô nhập kiểu ChatGPT)**

* **ENABLED: true** → bật ô nhập.
* **WIDTH\_PCT: 100** → bề rộng ô nhập theo % khung chat/center.
* **OFFSET\_X: "0px"** → **(X)** dịch **trái/→ phải** của ô nhập.
* **OFFSET\_Y: "0px"** → **(Y)** dịch **lên/↓ xuống**; đặt **âm** để kéo **lên gần giữa**.
* **BG: "#F7F7F8"** → nền ô nhập (xám rất nhạt).
* **BORDER\_COLOR: "#E5E7EB"** → màu viền ô nhập (xám).
* **BORDER\_PX: 1** → độ dày viền ô nhập.
* **RADIUS\_PX: 999** → bo tròn dạng **pill** (giống ChatGPT).
* **STICKY\_BOTTOM\_PX: 20** → ghim ô nhập cách **đáy** 20px; đặt **0** để **không sticky** (cho phép kéo lên bằng OFFSET\_Y).

**BUBBLE\_USER (bong bóng người dùng)**

* **ENABLED: true** → bật bong bóng user.
* **BG: "#E5F1FF"** → nền bong bóng user (xanh nhạt).
* **TEXT: "#0B1520"** → màu chữ user.
* **BORDER\_COLOR: "#DCEBFF"** → màu viền user.
* **BORDER\_PX: 1** → độ dày viền user.
* **RADIUS\_PX: 18** → bo góc bong bóng user.

**BUBBLE\_BOT (bong bóng bot)**

* **ENABLED: true** → bật bong bóng bot.
* **BG: "#FFFFFF"** → nền bong bóng bot.
* **TEXT: "#0B1520"** → màu chữ bot.
* **BORDER\_COLOR: "#E5E7EB"** → màu viền bot (xám nhạt).
* **BORDER\_PX: 1** → độ dày viền bot.
* **RADIUS\_PX: 18** → bo góc bong bóng bot.

**COMPONENTS (Bật/tắt phần giao diện & nội dung phụ)**

* **HIDE\_STREAMLIT\_CHROME: true** → ẩn menu/header/footer mặc định của Streamlit (giao diện sạch).
* **EMPTY\_STATE.SHOW: true** → hiện màn hình chào rỗng khi chưa có chat.
* **EMPTY\_STATE.TITLE\_HTML:** “Tôi có thể giúp được gì, huy?” → tiêu đề trạng thái rỗng.
* **EMPTY\_STATE.SUB\_HTML:** “Hỏi bất kỳ điều gì…” → mô tả phụ trạng thái rỗng.
* **SIDEBAR\_CONTENT.SHOW: true** → hiển thị các control demo trong sidebar.
* **SIDEBAR\_CONTENT.HEADER:** “Sidebar” → tiêu đề khối sidebar.
* **SIDEBAR\_CONTENT.ITEMS:** → danh sách control (slider/toggle/number):
  + **Temperature / Top-p** → tham số mô hình.
  + **Fallback** → bật fallback tổng quát.
  + **Retriever k** → số lượng tài liệu truy hồi.
  + **Min relevance** → ngưỡng điểm liên quan tối thiểu.

**Cheat-sheet “kéo trái/phải/lên/xuống” (đổi số là xong)**

* **Kéo sidebar**: OBJECTS.SIDEBAR.OFFSET\_X (**X**), OFFSET\_Y (**Y**).
* **Kéo toàn khối chat panel**: OBJECTS.CHAT\_PANEL.OFFSET\_X (**X**), OFFSET\_Y (**Y**).
* **Kéo ô nhập riêng**: OBJECTS.CHAT\_INPUT.OFFSET\_Y (**Y**, âm = kéo **lên**), OFFSET\_X (**X**).  
  Nếu đang **sticky** (STICKY\_BOTTOM\_PX > 0), ưu tiên ghim xuống đáy → muốn kéo **lên** thì đặt **STICKY\_BOTTOM\_PX = 0**.
* **Đặt rộng hẹp kiểu ChatGPT**: OBJECTS.CHAT\_PANEL.CHAT\_MAX\_WIDTH\_PX (cột giữa), CHAT\_INPUT.WIDTH\_PCT (ô nhập).
* **Màu/viền/nền**: COLORS.\* (tổng), OBJECTS.\*.BG / BORDER\_COLOR / BORDER\_PX / RADIUS\_PX (theo đối tượng).

p1a\_clean10\_ocr\_bctc\_GPT\_80\_2 🡪 chỉ in png ra văn bản.