**Quy trình khuyến nghị (ổn định cho nhiều mẫu bảng)**

1. **p0 – Pre-processing nhẹ (trước OCR, rất nhanh)**
   * Render PDF ở **DPI 500** (giữ cả ảnh gốc).
   * **Deskew** nếu |góc| > 0.3°.
   * **Binarize nhẹ** (adaptive threshold) + **remove margins** rõ ràng.
   * Không làm line-detection/cắt ô phức tạp ở bước này.
2. **p1 – OCR toàn trang (bắt buộc)**
   * Tesseract **vie+eng**, --psm 6, preserve\_interword\_spaces=1.
   * Xuất **TSV** (bbox + dòng + từ) + \_text.txt + \_meta.json.
   * (Tuỳ chọn) **OCR pass số** cho nửa phải trang/ô số:  
     --psm 6 -c tessedit\_char\_whitelist=0123456789.,()%/-  
     → tăng độ chính xác chuỗi số.
3. **p1.5 – Tableizer (sau OCR, dựa TSV) — trọng tâm**
   * Gom token → dòng theo y-center.
   * Bắt **2 cụm số right-anchored** → Số cuối năm, Số đầu năm.
   * Phần trái tách Mã số (regex ^\d[\d.]\*$), Chỉ tiêu, Thuyết minh.
   * Chuẩn hoá số (ngăn nghìn, (123)→-123, loại chữ lẫn số).
   * Xuất CSV/JSONL (đưa vào vector) + text\_pretty.txt + checks.txt.
4. **Fallback (chỉ khi p1.5 thất bại) – Table detection trước OCR**
   * **Medium:** OpenCV line detection (morphology/Hough) để **crop vùng bảng**, rồi OCR lại vùng đó (numeric-whitelist).
   * **Heavy:** PP-Structure/LayoutParser để detect cell/structure.  
     Bật khi: không tìm được 2 cột số ổn định **hoặc** tỷ lệ dòng lệch > 10% **hoặc** các kiểm tra cộng dồn (110=111+112; 130=131+136+139; 100=I+II+…) bị fail.