RP-01 | Protocolo de investigación | VERSION 0.1 | AUGUST 2026

Protocolo de benchmark para extracción multi-etapa con LLMs desde catálogos heterogéneos

Pablo Cardozo
Independent Applied AI Researcher
Buenos Aires, Argentina

Resumen

Este protocolo convierte un sistema implementado de ingesta de catálogos en tres etapas en una comparación falsable. El sistema primero extrae la estructura documental, luego reconstruye filas fuente y finalmente normaliza productos en lotes acotados con validación de esquemas y controles de procedencia. El benchmark propuesto compara esta descomposición con un baseline de una sola pasada usando exactitud por campo, cobertura de filas, integridad de procedencia, tasa de revisión, latencia y costo. Actualmente no se afirma ningún resultado comparativo.

Palabras clave: extracción documental; generación estructurada; procedencia; evaluación; Gemini

ESTADO DE LA EVIDENCIA	REVISIÓN FUENTE
Sistema implementado; benchmark controlado todavía no ejecutado	a5d3aadbfa04

1. Pregunta de investigación

¿Separar comprensión documental, reconstrucción de filas y normalización de productos mejora la confiabilidad frente a un prompt de una sola pasada?

2. Sistema bajo estudio

Un pipeline de actions en Convex que usa Gemini Files API, esquemas JSON estructurados, tres reintentos por llamada, timeouts de 60 segundos, lotes de diez filas, identificadores de fila fuente, scores de confianza, reglas de revisión forzada y estado de ingesta reanudable.

3. Método

1. Crear un corpus etiquetado con catálogos PDF, XLS y XLSX que incluya tablas, productos multilínea, packs, unidades, descuentos y filas ambiguas.
2. Comparar el pipeline implementado de tres etapas con un baseline fijado de extracción en una sola pasada.
3. Medir exactitud exacta y normalizada por campo, recall de filas, filas alucinadas, integridad de procedencia, tasa de revisión forzada, latencia y costo estimado de API.
4. Estratificar resultados por tipo de documento, complejidad de layout, cantidad de filas y ambigüedad.
5. Publicar predicciones crudas, etiquetas, código de scoring, configuración del modelo y ejemplos de fallos.

4. Evidencia actual

1. La implementación pública separa extracción de metadata, reconstrucción de filas y normalización de productos en etapas explícitas.
2. Cada respuesta del modelo se valida contra un esquema y la segunda etapa verifica cantidad total e identificadores únicos.

3. La tercera etapa verifica identidad del lote, pertenencia de filas fuente, páginas cubiertas y estado de revisión antes de persistir.
4. El repositorio reporta 409 tests automatizados sobre ingesta, validación, aislamiento, filtros e interfaz, pero no un benchmark etiquetado de extracción.

5. Limitaciones

1. Todavía no se publicó un corpus etiquetado, salida de baseline ni resultado comparativo.
2. Los tests de aplicación establecen comportamiento del software, no exactitud de extracción.
3. Los modelos usan identificadores preview y su comportamiento puede cambiar.
4. El umbral fijo de confianza de 0,5 no fue calibrado contra el costo de revisión humana.

6. Próximo experimento

1. Seleccionar y anonimizar un corpus representativo de catálogos.
2. Definir una guía de anotación y etiquetar por duplicado un subconjunto para medir concordancia.
3. Implementar el baseline de una sola pasada y un scorer compartido.
4. Ejecutar ensayos repetidos y publicar un informe versionado del benchmark.

7. Fuentes y reproducibilidad

1. Repositorio público
2. Implementación de ingesta en tres etapas
3. Documento de arquitectura

Revisión fuente: a5d3aadbfa049f4f6d77a2f144f557fa9bf7cffb

Cita sugerida

Cardozo, Pablo. "Protocolo de benchmark para extracción multi-etapa con LLMs desde catálogos heterogéneos." Protocolo de investigación, version 0.1, 2026.
https://pablo.cardozo.com.ar/es/research/multi-stage-document-extraction.

Documento de trabajo independiente. Sin peer review.