RP-01 | Research protocol | VERSION 0.1 | AUGUST 2026

A Benchmark Protocol for Multi-Stage LLM Extraction from Heterogeneous Product Catalogs

Pablo Cardozo
Independent Applied AI Researcher
Buenos Aires, Argentina

Abstract

This research protocol turns an implemented three-stage catalog ingestion system into a falsifiable comparison. The system first extracts document structure, then reconstructs source rows, and finally normalizes products in bounded batches with schema validation and provenance checks. The proposed benchmark compares this decomposition with a single-pass baseline on field accuracy, row coverage, provenance integrity, review rate, latency, and cost. No comparative result is currently claimed.

Keywords: document extraction; structured generation; provenance; evaluation; Gemini

EVIDENCE STATUS	SOURCE REVISION
Implemented system; controlled benchmark not yet run	a5d3aadbfa04

1. Research question

Does separating document understanding, row reconstruction, and product normalization improve extraction reliability over a single-pass prompt?

2. System under study

A Convex action pipeline using Gemini Files API, structured JSON schemas, three retries per model call, 60-second timeouts, ten-row batches, source-row identifiers, confidence scores, forced review rules, and resumable ingestion state.

3. Method

1. Create a labelled corpus covering PDF, XLS, and XLSX catalogs with tables, multi-line products, packs, units, discounts, and ambiguous rows.
2. Compare the implemented three-stage pipeline with a pinned single-pass extraction baseline.
3. Measure exact and normalized field accuracy, row recall, hallucinated rows, source-row integrity, forced-review rate, latency, and estimated API cost.
4. Stratify results by document type, layout complexity, row count, and ambiguity.
5. Publish raw predictions, labels, scoring code, model configuration, and failure examples.

4. Current evidence

1. The public implementation separates metadata extraction, row reconstruction, and product normalization into explicit stages.
2. Every model response is schema-validated and stage two checks total row count and unique row identifiers.
3. Stage three verifies batch identity, source-row membership, page coverage, and review status before persistence.

4. The repository reports 409 automated tests across ingestion, validation, isolation, filtering, and interface behavior, but no labelled extraction benchmark.

5. Limitations

1. No labelled corpus, baseline output, or comparative result is published yet.
2. Application tests establish software behavior, not extraction accuracy.
3. Model versions are preview identifiers and may change behavior over time.
4. A fixed confidence threshold of 0.5 has not been calibrated against human review cost.

6. Next experiment

1. Select and redact a representative catalogue corpus.
2. Define annotation guidelines and double-label a subset for agreement measurement.
3. Implement the single-pass baseline and shared scorer.
4. Run repeated trials and release a versioned benchmark report.

7. Sources and reproducibility

1. Public repository
2. Three-stage ingestion implementation
3. Architecture document

Source revision: a5d3aadbfa049f4f6d77a2f144f557fa9bf7cffb

Suggested citation

Cardozo, Pablo. "A Benchmark Protocol for Multi-Stage LLM Extraction from Heterogeneous Product Catalogs." Research protocol, version 0.1, 2026. https://pablo.cardozo.com.ar/research/multi-stage-document-extraction.

Independent working document. Not peer reviewed.