

Por Qué el RAG Tradicional Falla en Manufactura Regulada: Citación Determinista a Nivel de Página para 50,000 PNOs

Cómo diseñamos flujos híbridos de recuperación dispersa y densa con coordenadas OCR exactas a través de bibliotecas masivas de PNOs, registros de lote y Certificados de Análisis.

AUTHOR Dr. Carlos Perez Tobie

CATEGORY Inteligencia Documental

PUBLISHED 2026-08-28 / 7 min de lectura

Por Qué el RAG Tradicional Falla en Manufactura Regulada: C

En el desarrollo de software convencional, "RAG" (Generación Aumentada por Recuperación) suele limitarse a tomar archivos, dividirlos en fragmentos de 500 tokens, calcular incrustaciones (*embeddings*) con un modelo genérico y pasar los cinco resultados más cercanos al prompt.

En manufactura regulada -como la industria farmacéutica, alimentaria o médica- este enfoque representa un riesgo inaceptable.

La investigación de una desviación en un lote de producción no puede fundamentarse en un texto que sólo "resume" el protocolo de limpieza. Un auditor de calidad o un inspector sanitario exige el código exacto del Procedimiento Normalizado de Operación (PNO), el número de revisión vigente, la página precisa y el párrafo de referencia. Si el sistema afirma que una autoclave requiere 121°C durante 25 minutos pero su cita apunta a un párrafo sobre secado de material de vidrio, la auditoría se derrumba.

A continuación explicamos cómo construimos motores de inteligencia documental con citación determinista a nivel de página para archivos confidenciales de gran volumen.

1. La Canalización de Ingesta Multi-Escala

Las bibliotecas documentales en plantas productivas son altamente heterogéneas. Un solo repositorio combina archivos PDF vectoriales recién exportados de SAP con registros de lote escaneados hace 12 años con sellos de matriz de puntos y firmas autógrafas.

Nuestro motor de ingesta opera en tres fases rigurosas:

- **Extracción Nativa Vectorial:** Para documentos digitales limpios, las cajas delimitadoras (*bounding boxes*) de cada carácter y los pesos tipográficos se preservan directamente del flujo PDF, construyendo un mapa espacial exacto.
- **OCR de Alta Resolución con Visión Computacional:** Para documentos escaneados, las páginas se rasterizan a 300 DPI y se procesan con modelos locales de reconocimiento óptico de caracteres calibrados para nomenclatura química y terminología farmacéutica.
- **Reconstrucción Geométrica de Tablas:** Los Certificados de Análisis (CoAs) están compuestos por tablas con números de lote, especificaciones y resultados analíticos. La fragmentación arbitraria destruye las relaciones de filas y columnas. Nuestro motor reconstruye cada celda como una estructura JSON vinculada a sus coordenadas físicas exactas [x0, y0, x1, y1] en la página original.

2. Búsqueda Híbrida Léxica y Semántica

Las incrustaciones semánticas densas son excelentes para comprender conceptos generales, pero fallan cuando un operador busca un número de lote exacto, un código SKU o el número CAS de un reactivo químico.

Implementamos una arquitectura híbrida equilibrada:

- Búsqueda Dispersa (BM25): Indexación léxica de alta velocidad para códigos exactos de equipo, números de control y especificaciones numéricas.
- Búsqueda Semántica Densa: Modelos de vectores locales entrenados en documentación técnica industrial.
- Fusión por Rango Recíproco (RRF) y Re-clasificación con Cross-Encoder: Un modelo secundario que re-evalúa los 30 mejores candidatos en GPU local y entrega al modelo principal únicamente los fragmentos verificados.

3. El Contrato de Atribución Obligatoria

Nuestros agentes operan bajo una regla inquebrantable de procedencia:

> *Toda afirmación técnica, tolerancia de proceso o parámetro de control debe incluir un token de citación auditable que referencie [ID_Documento : Número_Página : Coordenadas_Espaciales]. Si una respuesta no puede vincularse directamente a una evidencia recuperada, el sistema tiene prohibido generarla.*

En la interfaz de usuario, los directores de planta no leen texto abstracto: hacen clic sobre cualquier dato y visualizan de inmediato el PDF original en pantalla dividida, con la línea exacta resaltada sobre la página fuente.