

Evidencia documental trazable con RAG: un asistente para auditorías GMP en la industria farmacéutica

María Eliana Pighin, Juan Manuel Fernández

Universidad Nacional de Luján, Buenos Aires, Argentina

me.pighin@hotmail.com, jmfernandez@unlu.edu.ar

Resumen Las auditorías de Buenas Prácticas de Manufactura (GMP) exigen localizar evidencia documental precisa en tiempo real, en plantas farmacéuticas cuya documentación activa supera los cientos de procedimientos. Este trabajo presenta un asistente documental basado en Retrieval-Augmented Generation (RAG) que permite consultar en lenguaje natural procedimientos operativos estándar (SOP) y directivas de calidad (QMS) reales de una planta en operación, y devuelve respuestas fundamentadas en evidencia trazable. El prototipo incorpora metadata específica del dominio, un filtro de diversidad documental y una expansión bilingüe de consultas, y se evaluó sobre 12 documentos reales mediante 20 consultas representativas de escenarios de auditoría. Los resultados muestran la viabilidad de aplicar grandes modelos de lenguaje (LLM) sobre documentación GMP controlada —el sistema alcanza un MRR de 0.935 y un Recall@5 de 0.822— y permiten caracterizar el principal desafío del enfoque en este dominio: distinguir el documento correcto dentro de series documentales temáticamente homogéneas.

Keywords: RAG · LLM · GMP · Auditoría farmacéutica · Trazabilidad documental

1. Introducción

En la industria farmacéutica, las auditorías de organismos reguladores como la Administración Nacional de Medicamentos, Alimentos y Tecnología Médica (ANMAT) exigen respuestas precisas y rápidas sobre procedimientos, controles y evidencias distribuidos en cientos de documentos: procedimientos operativos estándar (SOP), directivas del sistema de gestión de calidad (QMS), registros de validación y normativas de Buenas Prácticas de Manufactura (GMP) [1].

El modelo operativo actual depende de la presencia de personal experto en cada área auditada: la ausencia de un participante clave, o la dificultad para localizar el fragmento correcto en tiempo real, puede comprometer la defensa ante el organismo regulador. La limitación se agrava con el volumen documental: la planta de referencia cuenta con más de 730 SOP y 300 directivas QMS activas.

Este trabajo presenta un asistente conversacional que permite consultar la base documental GMP en lenguaje natural y devuelve evidencia trazable junto

con respuestas sintetizadas para contextos de auditoría, implementado como un *pipeline* de Retrieval-Augmented Generation (RAG) sobre documentos reales exportados desde Veeva Vault.

Las contribuciones de este trabajo son cuatro: (1) un *pipeline* RAG con metadata enriquecida específica del dominio farmacéutico; (2) una estrategia de diversidad de fuentes orientada a favorecer la cobertura multidocumental; (3) una evaluación por configuraciones que cuantifica el aporte de cada decisión de diseño sobre un corpus de idioma mixto; y (4) una interfaz orientada a usuarios no técnicos, con respuestas estructuradas para auditoría.

2. Estado del arte y antecedentes

Los grandes modelos de lenguaje (LLM), construidos sobre la arquitectura Transformer [2] y entrenados sobre corpus masivos [3], permiten responder consultas en lenguaje natural con notable fluidez, pero presentan dos limitaciones críticas para entornos regulados: pueden generar información no fundamentada (alucinaciones) y no ofrecen trazabilidad sobre el origen de sus respuestas [4]. RAG [5] aborda ambas limitaciones condicionando la generación a fragmentos documentales recuperados mediante representaciones densas [6,7] desde una base vectorial, de modo que cada respuesta queda anclada a fuentes identificables. Sobre este esquema se han propuesto extensiones como el *reranking*, la búsqueda híbrida y la expansión de consultas [8].

En dominios de documentación técnica especializada, RAG ha mostrado resultados competitivos en áreas como el derecho [9] y la medicina [10], donde la fidelidad a la fuente resulta tan relevante como la calidad de la respuesta. En el ámbito farmacéutico existen antecedentes vinculados al cumplimiento regulatorio [11], las desviaciones de manufactura [12] y la búsqueda sobre guías de la FDA [13]. Este trabajo se inscribe en esa línea, sobre documentación procedimental controlada de una planta en operación —SOP y directivas QMS— y con un corpus de idioma mixto: consultas en español sobre documentación mayormente en inglés.

3. Metodología

3.1. Base documental

Los documentos utilizados se exportaron desde Veeva Vault, el sistema de gestión documental de la planta, en formato PDF. Dado el volumen de la base documental, el prototipo se acotó a 12 documentos que permitieran validar el enfoque antes de escalarlo: 6 de sistemas computarizados GxP (QMS 2002–2006 y SOP G0108) y 6 de producción y calidad (QMS 131, 209, 268 y SOP P0002, P0251, P0258). La selección abarca dos áreas de naturaleza distinta e incorpora deliberadamente una serie temática completa —QMS 2002–2006—, un caso adverso para la recuperación por similitud. Por tratarse de documentación controlada, no se exponen contenidos sensibles: los resultados se presentan sobre códigos documentales.

3.2. Indexación y metadata enriquecida

En la Figura 1 se presenta la arquitectura del sistema: un *pipeline* de indexación (*offline*), que genera y persiste el índice vectorial, y uno de consulta (*online*), que lo reutiliza para responder en tiempo real.

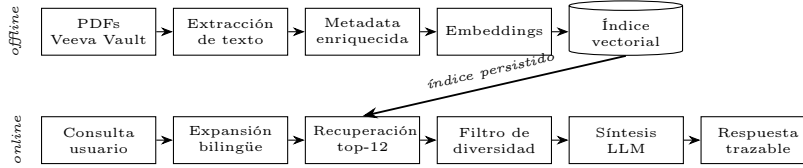


Figura 1: Arquitectura del sistema RAG para auditorías GMP.

En la etapa *offline* (fila superior), los PDF se procesan con PyMuPDF y cada fragmento se indexa con cuatro campos de metadata: código documental (inferido del nombre del archivo), tipo de documento (QMS o SOP), título y área funcional. El título se extrae mediante una regla basada en el *layout* homogéneo de los documentos de Veeva Vault, donde aparece inmediatamente antes de la línea IQMS-XXXXX; esta regla resulta más robusta que los enfoques basados en palabras clave o en LLM. Estos campos alimentan el filtro de diversidad y sustentan la citación de fuentes por parte del modelo.

3.3. Recuperación y generación

En la etapa *online* (Figura 1, fila inferior), la consulta se expande con términos técnicos en inglés antes de la búsqueda vectorial: los documentos QMS están redactados en inglés y los usuarios consultan en español, lo que genera *drift* semántico (“retiro” y “retirement” producen *embeddings* distintos). Se recuperan 12 fragmentos candidatos y se aplica un filtro de diversidad (máximo 2 fragmentos por documento) que favorece la cobertura multidocumental.

El sistema se construyó sobre LlamaIndex [14], cuyo motor de citación (CitationQueryEngine) obliga al modelo a citar sus fuentes explícitamente¹. Dos decisiones de diseño refuerzan la trazabilidad. En primer lugar, las etapas de recuperación y síntesis se mantienen separadas, en lugar de emplear CitationQueryEngine de forma monolítica, lo que permite aplicar el filtro de diversidad antes de la generación y garantiza que la síntesis se apoye exactamente en los fragmentos seleccionados. En segundo lugar, el *prompt* inyecta la metadata de cada fragmento como campos etiquetados que el modelo debe reproducir literalmente, en lugar de inferirla.

3.4. Protocolo de evaluación

Se definieron 20 consultas representativas de escenarios de auditoría real, cada una con sus documentos de referencia esperados, sobre las que se evaluaron tres

¹ El prototipo se implementó en Python 3.10; los modelos utilizados son `text-embedding-3-small` (*embeddings*) y `gpt-4o-mini` (síntesis).

configuraciones incrementales del sistema: (C1) búsqueda vectorial simple; (C2) C1 más expansión bilingüe; y (C3) C2 más filtro de diversidad. El desempeño del *retriever* se midió mediante dos métricas complementarias: **MRR** (*Mean Reciprocal Rank*), calculado a nivel de documento sobre el *ranking* de recuperación como el promedio de $1/r$, donde r denota la posición del primer documento esperado; y **Recall@5**, la fracción de los documentos esperados presentes en el contexto final entregado al modelo.

4. Resultados

La Tabla 1 ilustra el comportamiento del sistema sobre cinco consultas representativas del conjunto de evaluación²: en cuatro de los cinco casos el sistema recuperó la evidencia esperada y produjo una respuesta correcta. En todos ellos la respuesta conservó el formato estructurado previsto para el contexto de auditoría: documentos fuente con su metadata, evidencia documental específica y una observación sobre la suficiencia de la documentación recuperada.

Tabla 1: Casos de prueba evaluados sobre documentos reales GMP.

Consulta	Docs esperados	Resultado
¿Qué evidencias se requieren para retirar un sistema GxP?	2005, G0108	Correcto
¿Qué controles en proceso se realizan durante producción?	P0258, P0251, 131	Correcto
¿Qué información contiene un Master Batch Record?	P0002	Correcto
¿Qué documentación para migración de sistema GxP?	2006, G0108	Parcial
¿Requisitos de calidad para cleanrooms?	268	Correcto

Sobre el conjunto completo de 20 consultas, la Tabla 2 desagrega el desempeño del *retriever* en las tres configuraciones del protocolo, de acuerdo a las métricas MRR y Recall@5.

Tabla 2: Desempeño del retriever por configuración (20 consultas).

Configuración	MRR	Recall@5
C1: búsqueda vectorial simple	0.931	0.755
C2: C1 + expansión bilingüe	0.931	0.735
C3: C2 + filtro de diversidad	0.935	0.822

El sistema alcanza un MRR de 0.935 y un Recall@5 de 0.822 en su configuración final. El MRR, sin embargo, se mantiene prácticamente constante entre configuraciones: con doce documentos indexados y consultas que refieren de manera directa a su tema, la búsqueda vectorial simple ya sitúa un documento esperado en las primeras posiciones, con lo que la métrica se satura. Las diferencias entre configuraciones se manifiestan en Recall@5.

El filtro de diversidad (C3) es el componente de mayor incidencia: eleva el Recall@5 de 0.735 a 0.822 al evitar que varios fragmentos de un mismo documento

² Una demostración del prototipo se encuentra disponible en <https://asistente-gmp-pighin.streamlit.app/>.

ocupen posiciones del *top-5* que deberían corresponder a documentos distintos. La mejora resulta coherente con el escenario de auditoría, en el que numerosas consultas requieren evidencia de dos o tres documentos de forma simultánea.

La expansión bilingüe (C2), en cambio, no produce un efecto apreciable: deja el MRR sin cambios y la diferencia en Recall@5 respecto de C1 equivale a un documento en una consulta, dentro de la variabilidad esperable para un conjunto de veinte consultas. Dos factores concurren a explicarlo. El corpus no es uniformemente inglés: incorporar terminología en inglés aproxima la consulta a los documentos QMS, pero la aleja de los redactados en español —como el SOP G0108—, que ya presentaban coincidencia con la consulta original. A ello se suma que el modelo de *embeddings* posee capacidad multilingüe parcial, de modo que el *drift* se manifestaría principalmente en consultas con terminología muy específica, poco representadas en este conjunto. Determinar en qué condiciones resulta beneficiosa exige una evaluación ampliada y estratificada según el idioma del documento esperado.

El caso con resultado parcial de la Tabla 1 ilustra la segunda dificultad identificada: distinguir el documento más específico dentro de una serie temáticamente relacionada. Los documentos QMS 2002–2006 conforman una serie sobre el ciclo de vida de sistemas GxP y sus *embeddings* resultan semánticamente similares, por lo que una consulta sobre migración (QMS 2006) también recupera documentos sobre operación (QMS 2004) o implementación (QMS 2003). No se trata de un error del sistema, sino de una limitación inherente al enfoque basado en similitud: el *retriever* selecciona los fragmentos cuyos vectores son más cercanos al de la consulta y, dado que la serie comparte temática y vocabulario, esos vectores resultan naturalmente próximos. Una vía de mejora consiste en incorporar una etapa de *reranking* que evalúe la relevancia procesando consulta y fragmento de manera conjunta.

5. Conclusiones y trabajo futuro

Los resultados confirman la viabilidad de asistir auditorías GMP mediante un sistema RAG sobre documentación controlada de una planta en operación, con respuestas en lenguaje natural trazables a evidencia documental real. La evaluación por configuraciones cuantificó el aporte de cada decisión de diseño y evidenció que no todas contribuyen en igual medida: el filtro de diversidad mejora la cobertura de la evidencia, mientras que el efecto de la expansión bilingüe no resulta apreciable sobre un corpus de idioma mixto. Quedó caracterizada, además, la principal limitación del enfoque basado en similitud: distinguir el documento correcto dentro de series documentales temáticamente homogéneas.

Como trabajo futuro se propone incorporar una etapa de *reranking* con *cross-encoder* que resuelva la confusión entre documentos de una misma serie; resolver las consultas por código documental mediante filtrado por metadata, aprovechando que el código constituye un campo indexado; adoptar *embeddings* multilingües; integrar la API de Veeva Vault; y escalar el sistema al volumen completo de la planta, que supera los 1.000 documentos activos.

Referencias

1. ISPE: GAMP 5: A Risk-Based Approach to Compliant GxP Computerized Systems, 2nd edn. International Society for Pharmaceutical Engineering (2022)
2. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is All You Need. *Advances in Neural Information Processing Systems* 30, 5998–6008 (2017)
3. Brown, T., Mann, B., Ryder, N., et al.: Language Models are Few-Shot Learners. *Advances in Neural Information Processing Systems* 33, 1877–1901 (2020)
4. Ji, Z., Lee, N., Frieske, R., et al.: Survey of Hallucination in Natural Language Generation. *ACM Computing Surveys* 55(12), 1–38 (2023)
5. Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Kiela, D.: Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. *Advances in Neural Information Processing Systems* 33, 9459–9474 (2020)
6. Karpukhin, V., Oğuz, B., Min, S., et al.: Dense Passage Retrieval for Open-Domain Question Answering. In: *Proceedings of EMNLP 2020*, pp. 6769–6781 (2020)
7. Reimers, N., Gurevych, I.: Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In: *Proceedings of EMNLP-IJCNLP 2019*, pp. 3982–3992 (2019)
8. Gao, Y., Xiong, Y., Gao, X., et al.: Retrieval-Augmented Generation for Large Language Models: A Survey. *arXiv:2312.10997* (2023)
9. Louis, A., Spanakis, G.: A Statutory Article Retrieval Dataset in French. *Proceedings of ACL 2023* (2023)
10. Xiong, G., et al.: Benchmarking Retrieval-Augmented Generation for Medicine. *arXiv:2402.13178* (2024)
11. Kim, J., Hur, M., Min, M.: From RAG to QA-RAG: Integrating Generative AI for Pharmaceutical Regulatory Compliance Process. In: *Proceedings of the 40th ACM/SIGAPP Symposium on Applied Computing (SAC '25)*, pp. 1293–1295 (2025)
12. Salami, H., Smith-Goettler, B., Yadav, V.: How can language models assist with pharmaceuticals manufacturing deviations and investigations? *International Journal of Pharmaceutics* 669, 125100 (2025)
13. Proestel, S., Jeng, L.J.B., Smith, C., Deady, M., Amer, O., Ahmed, M., Rodgers, S.: Semantic Search of FDA Guidance Documents Using Generative AI. *Therapeutic Innovation & Regulatory Science* 59(5), 1148–1159 (2025)
14. LlamaIndex: LlamaIndex Documentation (2024). Disponible en: <https://docs.llamaindex.ai>, último acceso 2026/07/08