# Modelado e Implementación de un Sistema GraphRAG para Consultas en Lenguaje Natural sobre la Base de Datos de Hugging Face

Construcción de conocimiento sobre Hugging Face usando GraphRAG

Juan E, Gamba A.

Universidad Escuela Colombiana de Ingeniería Julio Garavito

Septiembre, 2025

## Introducción

La revolución en el procesamiento del lenguaje natural (NLP) impulsada por los Modelos de Lenguaje Grandes (LLMs, por sus siglas en inglés) ha transformado radicalmente la capacidad de las máquinas para comprender, generar y razonar con el lenguaje humano, impactando campos como la sumarización, la búsqueda semántica y el razonamiento autónomo (Roll, D. S et al 2025). Sin embargo, a pesar de sus ~~impresionantes~~ capacidades, los LLMs exhiben limitaciones fundamentales, incluyendo la propensión a **generar "alucinaciones"** (información plausible pero factual y contextualmente incorrecta). Debido a la dependencia de una base de conocimiento estática y a menudo desactualizada, y la falta de transparencia o trazabilidad en sus procesos de razonamiento (Roll, D. S et al 2025), estas deficiencias se acentúan en dominios especializados y en la gestión de conocimiento de "cola larga", donde la precisión factual y la consistencia son críticas (Yang, Z et al 2025).

Los grandes modelos de lenguaje exhiben limitaciones fundamentales, incluyendo la propensión a generar “alucinaciones” (información plausible pero factual y contextualmente incorrecta). Debido a la dependencia de una base de conocimiento estática, con la que fue entrenado el modelo, es decir no tiene en cuenta información actualizada (*debido a la dependencia a un conocimiento actualizado hasta cierta fecha, cualquier consulta posterior a esa fecha puede dar lugar a una respuesta equivocada. La data de los LLM no es en tiempo real*), y debido también a la falta de transparencia o trazabilidad en sus procesos de razonamiento (Roll, D. S et al 2025), estas alucinaciones se acentúan en dominios especializados y en conocimientos poco representados, pero potencialmente muy valiosos, donde la precisión factual y la consistencia son criticas (Yang, Z et al 2025)

Una solución prominente para mitigar estas limitaciones es la **Generación Aumentada por Recuperación (RAG, por sus siglas en inglés)**.( Yunfan Gaoa et al 2024; Roll, D. S et al 2025). En el esquema canónico de RAG, el sistema tiene acceso a un amplio corpus externo de registros textuales y recupera un subconjunto de ellos que sean individualmente relevantes para la consulta que se le haga al LLM y que, en conjunto, sean lo suficientemente pequeños para caber en la ventana de contexto del modelo (Darren Edge et al 2025). La **ventana de contexto** de un LLM es la cantidad máxima de texto (en tokens) que el modelo puede procesar y “recordar” simultáneamente durante una interacción.

Aunque el RAG convencional ha demostrado ser eficaz en mejorar la precisión factual y reducir las alucinaciones para preguntas que pueden responderse con información localizada dentro de un conjunto reducido de registros (Darren Edge et al 2025). Esta técnica presenta sus propias limitaciones a la hora de realizar consultas de comprensión o interpretación global (sensemaking queries) (Darren Edge et al 2025), es decir, aquellas que requieren una comprensión global del conjunto de datos. Las tareas de sensemaking requieren razonamiento sobre conexiones entre personas, lugares y eventos con el fin de anticipar sus trayectorias y actuar de forma efectiva (Klein et al., 2006). Además, RAG tiene una dependencia de la calidad y organización del corpus externo, así como una debilidad con los datos dispersos o estructurados jerárquicamente (Xie, X et al 2025). Los sistemas RAG tradicionales, al basarse en la recuperación de documentos planos, a menudo pasan por alto las jerarquías contextuales y las relaciones entre entidades, lo que reduce la interpretabilidad y la capacidad de los LLMs para "conectar los puntos" a través de piezas de información dispares (Jonathan & Steven, 2024; Knollmeyer, S 2025) y responder a preguntas de comprensión global y contextual especialmente en dominios específicos u complejos.

*(Poner una pregunta de ejemplo de sensemaking – evolución publicaciones de países latam, cuanto dinero a recibido china para estar contribuyendo a HuggingFace*)

Para abordar estas deficiencias, ha surgido el **GraphRAG**, un paradigma avanzado que integra la fortaleza de los grafos de conocimiento (KGs, por sus siglas en inglés) con las capacidades generativas de los LLMs (Xie, X et al 2025). GraphRAG utiliza un LLM para construir un grafo de conocimiento, donde los nodos representan entidades clave del corpus y los arcos (edges) representan relaciones entre ellas (Darren Edge et al 2025). Los KGs ofrecen una representación semántica rica e interpretable (Roll, D. S et al 2025) del corpus. Al aprovechar los KGs, GraphRAG no solo mejora la comprensión global y contextual, sino que también fomenta la exhaustividad, diversidad y relevancia en las respuestas de los LLMs (Darren Edge et al 2025), reduciendo drásticamente la ocurrencia de alucinaciones al anclar las generaciones en una base factual estructurada y verificable (Xie, X et al 2025).

La eficacia de GraphRAG se ha demostrado en diversos dominios, desde el diagnóstico de fallas de aeronaves (Xie, X et al 2025) y la identificación de desechos orbitales (Roll, D. S et al 2025) hasta la búsqueda inteligente en la industria minera (Li, B 2025) y sistemas de gobierno electrónico (Papageorgiou, G 2025). En particular, en un estudio se demostró que el enfoque de GraphRAG superó significativamente al **RAG convencional** tanto en los criterios de **comprension** como de **diversidad** *(Entendiendo por compresión ---- y por diversidad ----)* a través de distintos conjuntos de datos, mostrando una victoria del GraphRAG en el 65% - 80% de los casos (Darren Edge et al 2025).

enfoques como HybridRAG han logrado reducir las tasas de alucinación en más del 7% y mejorar la puntuación F1 en al menos un 4% en comparación con los *baselines* de RAG convencionales (Xie, X et al 2025).

Además, la implementación de GraphRAG habilitado con Neo4j ha mostrado una **reducción notable en las alucinaciones**, llegando a "cero alucinaciones" en ciertas tareas cualitativas (Roll, D. S et al 2025). Estos sistemas suelen emplear **Neo4j** como base de datos de grafos, gracias a su robustez, madurez y capacidades para modelar y consultar datos interconectados a través de su modelo de grafo de propiedades y el lenguaje de consulta Cypher (Xie, X et al 2025).

La plataforma **Hugging Face (HF)** se ha consolidado como un centro crucial para el desarrollo y la colaboración en proyectos de *Machine Learning*, facilitando el intercambio de modelos preentrenados, conjuntos de datos y *espacios* (Ait et al 2024). **HFCOMMUNITY** es una herramienta que recopila y organiza información sobre los repositorios y discusiones de la comunidad en el HFH en una base de datos relacional, ofreciendo conceptos específicos del dominio y permitiendo la exploración mediante lenguajes tipo SQL (Ait et al 2024). Aunque HFCOMMUNITY facilita el análisis de datos mediante su estructura relacional, la vasta y compleja red de interconexiones entre modelos, conjuntos de datos, usuarios, discusiones y sus metadatos intrínsecos (91,616,030 de entidades y 115,051,895 de relaciones) no se explota completamente en un formato que permita un razonamiento semántico profundo y consultas en lenguaje natural transparentes. Según Ait et al (2024), Existe la posibilidad de trabajar en la aplicación de técnicas basadas en NLP para extraer información adicional de las descripciones de los repositorios, lo que podría generar "datos de anotación de grano fino para análisis de explicabilidad", un área que se alinea perfectamente con las capacidades de GraphRAG.

Esta tesis aborda la creciente necesidad de sistemas de inteligencia artificial que no solo generen respuestas fluidas, sino que también demuestren precisión factual, trazabilidad y una comprensión contextual y global profunda en dominios especializados. **En esta investigación diseño e implemento un sistema GraphRAG** que permite consultas en lenguaje natural sobre la base de datos de Hugging Face, implementando grafos de conocimiento en Neo4j, una base de datos orientada a grafos que almacena y gestiona información mediante nodos, relaciones y propiedades.

A diferencia de las aproximaciones existentes que utilizan métodos RAG tradicionales, en este trabajo me centro en la construcción de un grafo de conocimiento estructurado en Neo4j a partir de los datos de HFCOMMUNITY y este lo integro como insumo principal del sistema de GraphRAG. La originalidad de esta investigación radica en una de las primeras aplicaciones en Colombia de un *framework* GraphRAG específicamente diseñado para capitalizar y explotar el potencial en la naturaleza interconectada de los modelos, datasets y aplicaciones en el Hugging Face Hub. Al hacer esta integración, no solo logro la ejecución de consultas complejas en lenguaje natural con una mayor precisión contextual, un entendimiento global y una reducción significativa de las alucinaciones, sino que también ofreceré una trazabilidad inherente en el razonamiento de las respuestas, algo crucial para la confianza y la auditabilidad de las respuestas del modelo.

La metodología de esta tesis incluirá la extracción y estructuración de entidades y relaciones clave de la base de datos HFCOMMUNITY para construir un grafo de conocimiento detallado en Neo4j. Se implementarán estrategias de recuperación híbridas que combinarán la traversión del grafo con técnicas de incrustación densa (utilizando modelos como Sentence-BERT o MiniLM) para contextualizar las consultas en lenguaje natural (Roll, D. S et al 2025). Posteriormente, se detallará el proceso de aumento del *prompt* de un LLM (e.g., LLaVA, GPT-4.1-mini o Qwen2-72b) con los subgrafos relevantes recuperados y se evaluará la capacidad del sistema para generar respuestas precisas, explicables y factuales. Finalmente, se presentarán los resultados de las evaluaciones cuantitativas y cualitativas que demuestren la eficacia del sistema propuesto en la reducción de alucinaciones y la mejora de la calidad de las respuestas en comparación con *baselines* de RAG convencionales.

## Referencias

[1] Roll, D. S., Kurt, Z., Li, Y., & Woo, W. L. (2025). Augmenting Orbital Debris Identification with Neo4j-Enabled Graph-Based Retrieval-Augmented Generation for Multimodal Large Language Models. *Sensors*, *25*(11), 3352. <https://doi.org/10.3390/s25113352>

[2] Ait, A., Cánovas Izquierdo, J. L., & Cabot, J. (2024). HFCommunity: An extraction process and relational database to analyze Hugging Face Hub data. *Science of Computer Programming*, *234*, 103079. <https://doi.org/10.1016/j.scico.2024.103079>

[3] Li, B., Wang, Y., Ding, Z., Wang, B., & Wen, S. (2025). Intelligent search technology for Jiaodong gold mines based on large models and GraphRAG. *Earth Science Frontiers*. [https://doi.org/10.13745/j.esf.sf.2025.4.77](https://www.google.com/search?q=https://doi.org/10.13745/j.esf.sf.2025.4.77)

[4] Xie, X., Tang, X., Gu, S., & Cui, L. (2025). An intelligent guided troubleshooting method for aircraft based on HybirdRAG. *Scientific Reports*, *15*, 17752. <https://doi.org/10.1038/s41598-025-02643-2>

[5] Papageorgiou, G., Sarlis, V., Maragoudakis, M., & Tjortjis, C. (2025). Hybrid Multi-Agent GraphRAG for E-Government: Towards a Trustworthy AI Assistant. *Applied Sciences*, *15*(11), 6315 . <https://doi.org/10.3390/app15116315>

[6] Knollmeyer, S., Caymazer, O., & Grossmann, D. (2025). Document GraphRAG: Knowledge Graph Enhanced Retrieval Augmented Generation for Document Question Answering Within the Manufacturing Domain. *Electronics*, *14*(11), 2102 <https://doi.org/10.3390/electronics14112102>

[7] Edge, D., Trinh, H., Cheng, N., Bradley, J., Chao, A., Mody, A., Truitt, S., Metropolitansky, D., Ness, R. O., & Larson, J. (2024). From Local to Global: A Graph RAG Approach to Query-Focused Summarization. <https://doi.org/10.48550/arXiv.2404.16130>.