**Resumen**

El uso de la generación aumentada por recuperación (RAG, *Retrieval-Augmented Generation*) para obtener información relevante desde una fuente de conocimiento externa permite que los grandes modelos de lenguaje (LLMs, *Large Language Models*) respondan preguntas sobre colecciones de documentos privadas o previamente no vistas.

Sin embargo, RAG falla al enfrentar preguntas de carácter global dirigidas a un corpus completo de texto, tales como *“¿Cuáles son los temas principales en el conjunto de datos?”*, ya que este tipo de interrogante constituye inherentemente una tarea de **resumen enfocado en consultas** (*Query-Focused Summarization, QFS*) más que una tarea explícita de recuperación.

Los métodos previos de QFS, por su parte, no escalan adecuadamente a las cantidades de texto que suelen indexar los sistemas RAG típicos. Para combinar las fortalezas de ambos enfoques contrastantes, proponemos **GraphRAG**, un método basado en grafos para responder preguntas sobre textos privados que escala tanto con la generalidad de las preguntas del usuario como con la cantidad de texto fuente.

Nuestro enfoque utiliza un LLM para construir un índice en grafo en dos etapas:

1. Primero, deriva un **grafo de conocimiento de entidades** a partir de los documentos fuente.
2. Luego, genera previamente **resúmenes comunitarios** para todos los grupos de entidades estrechamente relacionadas.

Dada una pregunta, cada resumen comunitario se usa para generar una respuesta parcial, y todas las respuestas parciales se resumen nuevamente en una respuesta final para el usuario.

Para una clase de preguntas de **comprensión global** sobre conjuntos de datos del orden de un millón de *tokens*, demostramos que GraphRAG produce mejoras sustanciales respecto a una línea base RAG convencional, tanto en la **exhaustividad** como en la **diversidad** de las respuestas generadas.

**1. Introducción**

La **generación aumentada por recuperación** (*Retrieval-Augmented Generation*, RAG) (Lewis et al., 2020) es un enfoque consolidado para utilizar **grandes modelos de lenguaje (LLMs)** con el fin de responder consultas basadas en datos demasiado extensos para caber dentro de la ventana de contexto del modelo.  
Esta ventana de contexto se refiere al número máximo de *tokens* (unidades de texto) que el LLM puede procesar de forma simultánea (Kuratov et al., 2024; Liu et al., 2023).

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** y que, en conjunto, sean **lo suficientemente pequeños para caber en la ventana de contexto** del modelo.  
El LLM genera entonces una respuesta basada tanto en la consulta como en los registros recuperados (Baumel et al., 2018; Dang, 2006; Laskar et al., 2020; Yao et al., 2017).

Este enfoque convencional, denominado en conjunto **vector RAG**, funciona bien para preguntas que pueden responderse con información **localizada** dentro de un conjunto reducido de registros.  
Sin embargo, los enfoques vector RAG no admiten las llamadas **consultas de comprensión o interpretación global** (*sensemaking queries*), es decir, aquellas que requieren una **comprensión global del conjunto de datos**, como por ejemplo:

“¿Cuáles son las tendencias clave en cómo los descubrimientos científicos han sido influenciados por la investigación interdisciplinaria durante la última década?”

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).

Los LLMs, como GPT (Achiam et al., 2023; Brown et al., 2020), Llama (Touvron et al., 2023) y Gemini (Anil et al., 2023), han demostrado gran capacidad para la **comprensión global y contextual** en dominios complejos como el **descubrimiento científico** (Microsoft, 2023) o el **análisis de inteligencia** (Ranade y Joshi, 2023).  
Ante una consulta de *sensemaking* y un texto con un conjunto implícito e interconectado de conceptos, un LLM puede generar un **resumen que responda a la pregunta de forma interpretativa**.

No obstante, el reto surge cuando el **volumen de datos** obliga a recurrir a un enfoque RAG, pues los métodos vectoriales no pueden realizar *sensemaking* sobre un **corpus completo**.

En este trabajo presentamos **GraphRAG**, un enfoque basado en grafos que permite realizar *sensemaking* sobre la totalidad de un gran corpus textual.  
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.

Posteriormente, el grafo se **divide jerárquicamente en comunidades** de entidades estrechamente relacionadas, y luego se emplea un LLM para generar **resúmenes por comunidad**.  
Estos resúmenes se producen de forma **ascendente (bottom-up)** siguiendo la estructura jerárquica de comunidades detectadas, y los resúmenes de los niveles superiores **incorporan recursivamente** los de los niveles inferiores.  
En conjunto, estos **resúmenes comunitarios** ofrecen una descripción y comprensión global del corpus.

Finalmente, GraphRAG responde a las consultas mediante un **procesamiento tipo *map-reduce*** aplicado a los resúmenes de comunidad:

* En la etapa *map*, los resúmenes generan **respuestas parciales** a la consulta de manera independiente y en paralelo.
* En la etapa *reduce*, esas respuestas parciales se **combinan y sintetizan** en una **respuesta global final**.

El método GraphRAG y su capacidad para realizar *sensemaking* global sobre un corpus completo constituyen la principal contribución de este trabajo.

Para demostrar dicha capacidad, desarrollamos una aplicación novedosa de la técnica **LLM-as-a-Judge** (Zheng et al., 2024), adecuada para preguntas amplias o temáticas donde no existe una respuesta “verdadera” (ground truth).  
Esta metodología utiliza un primer LLM para **generar un conjunto diverso de preguntas de *sensemaking*** sobre un corpus específico, y un segundo LLM para **evaluar las respuestas** generadas por diferentes sistemas RAG según criterios predefinidos (descritos en la Sección 3.3).

Usamos este enfoque para comparar **GraphRAG** con el enfoque **vector RAG** en dos conjuntos de datos reales representativos.  
Los resultados muestran que GraphRAG **supera ampliamente** a vector RAG al utilizar GPT-4 como modelo base, tanto en **exhaustividad** como en **diversidad** de las respuestas.

GraphRAG está disponible como software de **código abierto** en:  
🔗 <https://github.com/microsoft/graphrag>

Además, versiones de GraphRAG se encuentran integradas como extensiones en múltiples librerías de código abierto, incluyendo **LangChain** (2024), **LlamaIndex** (2024), **NebulaGraph** (2024) y **Neo4J** (2024).

**2. Antecedentes**

**2.1. Enfoques y sistemas RAG**

En términos generales, **RAG (Retrieval-Augmented Generation)** se refiere a cualquier sistema en el cual una **consulta del usuario** se utiliza para **recuperar información relevante** desde fuentes de datos externas, información que posteriormente se incorpora en la generación de una respuesta mediante un **modelo de lenguaje grande (LLM)** u otro modelo generativo (por ejemplo, un modelo multimodal).

Tanto la consulta como los registros recuperados se integran en una **plantilla de *prompt***, la cual se pasa al LLM (Ram et al., 2023).  
El uso de RAG resulta ideal cuando el número total de registros en una fuente de datos es **demasiado grande para incluirse en un solo *prompt***, es decir, cuando la cantidad total de texto **excede la ventana de contexto** del modelo.

En los enfoques RAG canónicos, el proceso de recuperación devuelve un número determinado de registros que son **semánticamente similares** a la consulta, y la respuesta generada se basa únicamente en la información contenida en esos registros.  
Un enfoque común en RAG convencional es el uso de **embeddings textuales**, donde se recuperan los registros **más cercanos al vector de la consulta** en un espacio vectorial, siendo la “cercanía” equivalente a la **similitud semántica** (Gao et al., 2023).

Si bien algunos enfoques RAG emplean mecanismos de recuperación alternativos, agrupamos a esta familia de métodos convencionales bajo el término **vector RAG**.  
**GraphRAG**, en contraste, difiere en su capacidad de responder consultas que requieren **comprensión o razonamiento global** (*global sensemaking*) sobre la totalidad del corpus de datos.

**2.2. Uso de grafos de conocimiento con LLMs y RAG**

Los enfoques para **extraer grafos de conocimiento** a partir de corpus de texto natural incluyen técnicas de **reglas y coincidencias de patrones**, **reconocimiento estadístico de patrones**, **clustering** y **embeddings** (Etzioni et al., 2004; Kim et al., 2016; Mooney y Bunescu, 2005; Yates et al., 2007).

**GraphRAG** se enmarca dentro de un cuerpo de investigación más reciente que explora el uso de **LLMs para la extracción de grafos de conocimiento** (Ban et al., 2023; Melnyk et al., 2022; OpenAI, 2023; Tan et al., 2017; Trajanoska et al., 2023; Yao et al., 2023; Zhang et al., 2024a).  
Asimismo, contribuye al creciente conjunto de enfoques RAG que utilizan un **grafo de conocimiento como índice** (Gao et al., 2023).

Algunas técnicas emplean **subgrafos** o elementos del grafo (nodos, aristas o atributos) directamente dentro del *prompt* (Baek et al., 2023; He et al., 2024; Zhang, 2023) o los usan como **base factual** para generar salidas fundamentadas (Kang et al., 2023; Ranade y Joshi, 2023).  
Otros métodos (Wang et al., 2023b) utilizan el grafo de conocimiento para **mejorar la recuperación**, permitiendo que, durante el tiempo de consulta, un **agente basado en LLM** recorra dinámicamente el grafo, cuyos **nodos** representan elementos documentales (por ejemplo, párrafos o tablas) y cuyos **enlaces (edges)** codifican **similitud léxica, semántica o relacional**.

**GraphRAG** se distingue de estos enfoques al centrarse en una cualidad poco explorada de los grafos en este contexto: su **modularidad inherente** (Newman, 2006) y su capacidad para **particionar el grafo en comunidades modulares anidadas** de nodos estrechamente relacionados.  
En particular, GraphRAG **crea resúmenes cada vez más globales** de manera recursiva, utilizando el LLM para generar síntesis que abarcan esta jerarquía de comunidades.

**2.3. Evaluación adaptativa para RAG**

Existen numerosos **conjuntos de datos de referencia (benchmarks)** para tareas de **preguntas y respuestas de dominio abierto**, como *HotPotQA* (Yang et al., 2018), *MultiHop-RAG* (Tang y Yang, 2024) y *MT-Bench* (Zheng et al., 2024).  
Sin embargo, estos benchmarks se orientan principalmente a evaluar el **desempeño del RAG vectorial**, es decir, la capacidad de **recuperar hechos explícitos**.

En este trabajo proponemos un enfoque para **generar preguntas que evalúen la comprensión global** de un corpus en su totalidad.  
Nuestro método se relaciona con estrategias que utilizan un LLM para **generar preguntas cuyas respuestas serían resúmenes del corpus** (Xu y Lapata, 2021).  
No obstante, con el fin de garantizar una evaluación justa, nuestro procedimiento evita **generar preguntas directamente a partir del mismo corpus**, utilizando, por ejemplo, un subconjunto de este que se mantiene fuera de las etapas de extracción de grafos y de evaluación posterior.

El concepto de **evaluación adaptativa** se refiere al proceso de **generar dinámicamente benchmarks de evaluación adaptados** a dominios o casos de uso específicos.  
Trabajos recientes han utilizado LLMs para crear este tipo de benchmarks con el fin de asegurar **relevancia, diversidad y alineación** con la aplicación o tarea objetivo (Yuan et al., 2024; Zhang et al., 2024b).

En este trabajo, proponemos un método de **evaluación adaptativa** para generar **consultas de comprensión global (*sensemaking queries*)** para el LLM.  
Nuestro enfoque se inspira en investigaciones previas sobre **generación de “personas” mediante LLMs**, en las que el modelo produce **conjuntos diversos y verosímiles de perfiles de usuario** (Kosinski, 2024; Salminen et al., 2024; Shin et al., 2024).  
En este caso, el procedimiento de evaluación adaptativa emplea la generación de *personas* para crear **consultas representativas de usos reales** de un sistema RAG.  
Específicamente, el LLM **infieren los posibles tipos de usuarios** y **sus casos de uso**, los cuales guían la **generación de consultas de comprensión global** específicas para cada corpus.

**2.4. Criterios de evaluación RAG**

Nuestra evaluación se basa en el uso de un LLM para **juzgar qué tan bien responde** el sistema RAG a las preguntas generadas.  
Investigaciones previas han demostrado que los LLMs son **buenos evaluadores** en tareas de generación de lenguaje natural, y que sus juicios pueden ser **comparables a los de evaluadores humanos** (Wang et al., 2023a; Zheng et al., 2024).

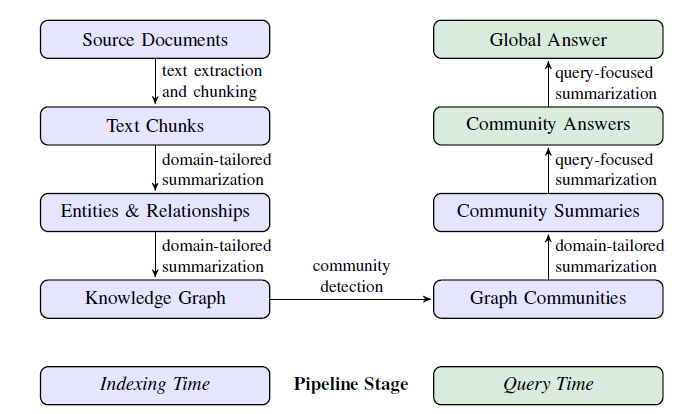
Algunos trabajos proponen criterios para cuantificar la **calidad de los textos generados**, como la **fluidez**, la **relevancia del contexto**, la **fidelidad** y la **relevancia de la respuesta** (RAGAS, Es et al., 2023).  
Dado que no existe una “respuesta verdadera” para este tipo de tareas de *sensemaking*, se puede medir el **rendimiento relativo** entre modelos pidiéndole al LLM que **compare respuestas generadas por dos sistemas distintos**, lo que se conoce como el método **LLM-as-a-Judge** (Zheng et al., 2024).

En este trabajo, definimos criterios específicos para evaluar las respuestas generadas por RAG frente a preguntas de comprensión global, y los evaluamos utilizando el enfoque comparativo.  
También validamos los resultados mediante estadísticas derivadas de **hechos verificables** extraídos por el LLM, o **“declaraciones” (claims)**.

**3. Métodos**

**3.1. Flujo de trabajo de GraphRAG**

La **Figura 1** del artículo ilustra el flujo general de datos del enfoque y la arquitectura de la canalización (*pipeline*) de **GraphRAG**.  
En esta sección se describen los parámetros de diseño, las técnicas y los detalles de implementación de cada una de sus etapas principales.



**3.1.1. Documentos fuente → Fragmentos de texto (*Text Chunks*)**

El proceso comienza dividiendo los documentos del corpus en **fragmentos de texto (chunks)**.  
El LLM extrae información de cada fragmento para su procesamiento posterior.  
La elección del tamaño del fragmento es una **decisión de diseño fundamental**:

* Fragmentos **más largos** requieren menos llamadas al LLM (reduciendo costos),
* pero pueden disminuir la **recuperación de información (recall)** sobre los elementos que aparecen al inicio del texto (Kuratov et al., 2024; Liu et al., 2023).

En el Apéndice A.1 se presentan los *prompts* y ejemplos que ilustran la compensación entre **precisión y exhaustividad**.

**3.1.2. Fragmentos de texto → Entidades y relaciones**

En esta etapa, el LLM se emplea para **extraer entidades importantes** y las **relaciones entre ellas** a partir de cada fragmento.  
Además, el modelo genera **breves descripciones** tanto de las entidades como de las relaciones detectadas.

Por ejemplo, si un fragmento contiene el siguiente texto:

“Las acciones de NeoChip (NC) se dispararon durante su primera semana de cotización en la bolsa NewTech Exchange.  
Sin embargo, analistas del mercado advierten que la salida a bolsa del fabricante de chips podría no reflejar las tendencias de otras IPOs tecnológicas.  
NeoChip, anteriormente una entidad privada, fue adquirida por Quantum Systems en 2016.  
La firma de semiconductores se especializa en procesadores de bajo consumo para dispositivos portátiles e IoT.”

El LLM debe extraer lo siguiente:

* **Entidad 1:** NeoChip — “NeoChip es una empresa pública especializada en procesadores de bajo consumo para dispositivos portátiles e IoT.”
* **Entidad 2:** Quantum Systems — “Quantum Systems es una firma que anteriormente fue propietaria de NeoChip.”
* **Relación:** “Quantum Systems fue propietaria de NeoChip desde 2016 hasta que esta se hizo pública.”

Estos *prompts* pueden **adaptarse al dominio** del corpus mediante **ejemplos *few-shot*** (Brown et al., 2020), lo que permite especialización en ámbitos como la ciencia, la medicina o el derecho.

Asimismo, el LLM puede extraer **afirmaciones o hechos (claims)** relacionados con las entidades detectadas.  
Por ejemplo, a partir del texto anterior se extraerían las siguientes afirmaciones:

* “Las acciones de NeoChip se dispararon durante su primera semana de cotización en la bolsa NewTech Exchange.”
* “NeoChip debutó como empresa pública en la bolsa NewTech Exchange.”
* “Quantum Systems adquirió NeoChip en 2016 y la poseyó hasta que NeoChip salió a bolsa.”

Los detalles de implementación y ejemplos de *prompts* para esta extracción de entidades y afirmaciones se describen en el **Apéndice A**.

**3.1.3. Entidades y relaciones → Grafo de conocimiento**

El uso de un LLM para extraer entidades, relaciones y afirmaciones constituye una forma de **resumen abstractivo**, ya que se generan **representaciones semánticas significativas** de conceptos que, en algunos casos, **no están explícitamente formulados** en el texto.

El proceso de extracción produce **múltiples instancias** de cada entidad o relación, ya que una misma entidad puede detectarse repetidamente a lo largo del corpus.  
En el paso final de la extracción, estas instancias se convierten en **nodos y aristas del grafo**.

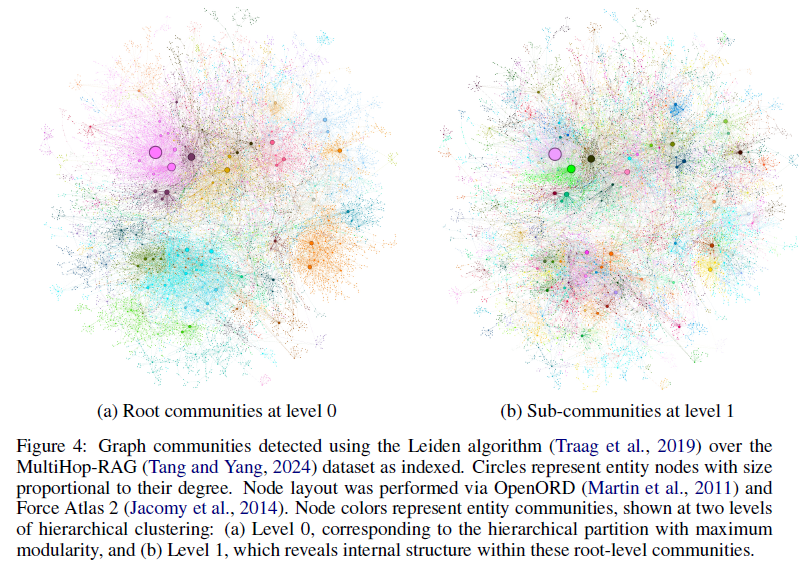
Las descripciones de entidades y relaciones se **agregan y resumen**, mientras que las relaciones repetidas se **fusionan y ponderan** según su frecuencia de aparición.  
Del mismo modo, las afirmaciones se agregan y vinculan al grafo.

En este trabajo, se utilizó una técnica de **coincidencia exacta de cadenas (string matching)** para identificar entidades equivalentes (Barlaug y Gulla, 2021; Christen y Christen, 2012; Elmagarmid et al., 2006).  
No obstante, es posible utilizar **métodos de coincidencia más flexibles** con ligeros ajustes en los *prompts* o el código.  
Además, GraphRAG muestra **tolerancia ante duplicados**, ya que las entidades redundantes suelen **agruparse dentro de la misma comunidad** en etapas posteriores.

**3.1.4. Grafo de conocimiento → Comunidades de grafo**

Una vez creado el índice en forma de grafo, se aplican **algoritmos de detección de comunidades** para **particionar el grafo** en grupos de nodos fuertemente conectados (Fortunato, 2010; Jin et al., 2021).  
En la canalización de GraphRAG se emplea el método **Leiden** (Traag et al., 2019), aplicado de manera **jerárquica y recursiva** hasta obtener subcomunidades que no pueden subdividirse más.

Cada nivel de esta jerarquía representa una **partición completa y exclusiva** de los nodos del grafo, lo que habilita una estrategia **divide y vencerás** para la **generación de resúmenes globales**.  
Un ejemplo visual de esta partición jerárquica se presenta en el **Apéndice B**.



**3.1.5. Comunidades de grafo → Resúmenes de comunidad**

El siguiente paso consiste en crear **resúmenes tipo “reporte”** de cada comunidad detectada.  
Estos resúmenes no solo sirven para responder preguntas, sino también como una herramienta **independiente de análisis del corpus**, proporcionando una visión estructurada de su semántica y organización.

Por ejemplo, un usuario podría revisar los resúmenes de nivel superior para identificar **temas generales**, y luego explorar los **subniveles** para obtener detalles específicos.  
En esta sección, sin embargo, los resúmenes se utilizan principalmente como parte del **índice basado en grafo** para responder consultas globales.

**GraphRAG genera resúmenes comunitarios** añadiendo los resúmenes de elementos (nodos, relaciones y afirmaciones) a una plantilla, siguiendo una estrategia ascendente:

* **Comunidades de nivel inferior (hojas):**  
  Se priorizan los elementos según su **grado de conexión** (prominencia), y se van añadiendo al contexto del LLM hasta alcanzar el límite de *tokens*.
* **Comunidades de nivel superior:**  
  Si todos los resúmenes caben dentro del límite de contexto, se resumen directamente.  
  Si no, se **sustituyen los sub-resúmenes más extensos** por **resúmenes más breves** de subcomunidades hasta que el texto encaje dentro del contexto permitido.

**3.1.6. Resúmenes de comunidad → Respuestas comunitarias → Respuesta global**

Una vez que el usuario formula una consulta, los resúmenes generados se utilizan para construir una **respuesta final en múltiples etapas**.  
Dado que la estructura comunitaria es jerárquica, una pregunta puede responderse a partir de distintos niveles, buscando el **equilibrio entre detalle y generalidad** (analizado en la Sección 4).

Para un nivel comunitario dado, el proceso es el siguiente:

1. **Preparar resúmenes comunitarios:**  
   Se barajan aleatoriamente y dividen en fragmentos de tamaño predefinido, distribuyendo la información de forma uniforme entre los contextos.
2. **Generar respuestas parciales (fase *map*):**  
   El LLM produce respuestas intermedias en paralelo, asignando además una **puntuación de relevancia** (0–100) que mide su utilidad para responder la pregunta.  
   Las respuestas con puntuación 0 se descartan.
3. **Combinar respuestas (fase *reduce*):**  
   Las respuestas parciales se ordenan según su puntuación de relevancia y se agregan progresivamente en un nuevo contexto, hasta alcanzar el límite de *tokens*.  
   Ese contexto final se utiliza para generar la **respuesta global definitiva**.

**3.2. Generación de preguntas de comprensión global (*Sensemaking Queries*)**

Las preguntas utilizadas para evaluar GraphRAG no se formulan manualmente. En su lugar, se generan de manera **automática mediante un LLM**, con el propósito de simular **consultas realistas** que un usuario humano podría realizar al explorar un corpus desconocido.  
Este proceso tiene como objetivo evaluar la **capacidad de comprensión y síntesis global** del sistema.

El procedimiento se desarrolla en tres pasos principales:

1. **Análisis del corpus y generación de arquetipos de usuario (*personas*):**  
   El LLM recibe una descripción general del corpus (por ejemplo, resúmenes de documentos o metadatos) y debe inferir **qué tipos de usuarios** podrían estar interesados en esa información.  
   Cada arquetipo generado incluye una breve descripción de su **perfil, motivaciones e intereses** (por ejemplo, “investigador en políticas públicas”, “analista de medios”, “director de innovación tecnológica”, etc.).
2. **Derivación de casos de uso y objetivos de consulta:**  
   Para cada *persona*, el LLM identifica posibles **casos de uso** o **objetivos de investigación**, tales como “comprender las causas subyacentes de un fenómeno”, “evaluar el impacto de una política” o “detectar relaciones emergentes entre temas clave”.  
   Esto permite que las consultas generadas tengan un propósito realista y estén alineadas con contextos de aplicación plausibles.
3. **Generación de preguntas de *sensemaking*:**  
   A partir de los objetivos anteriores, el LLM genera **preguntas abiertas y amplias** que exigen una comprensión integradora del corpus.  
   Estas preguntas no pueden responderse con hechos aislados ni con fragmentos específicos de texto.  
   Ejemplos típicos incluyen:
   * “¿Cómo ha evolucionado la colaboración interdisciplinaria en este campo durante la última década?”
   * “¿Cuáles son las principales tensiones o desacuerdos en torno a este tema?”
   * “¿Qué factores parecen estar impulsando los cambios más recientes?”

El resultado de este proceso es un conjunto de **consultas diversas y profundas**, diseñadas para poner a prueba la **capacidad del sistema de construir sentido global (sensemaking)**, en lugar de limitarse a la recuperación literal de información.

**3.3. Evaluación de resultados mediante LLM-as-a-Judge**

Para comparar los distintos enfoques (GraphRAG frente a vector RAG), se emplea una técnica de evaluación automática denominada **LLM-as-a-Judge** (Zheng et al., 2024).  
En este enfoque, un LLM actúa como **evaluador imparcial** que compara las respuestas generadas por dos sistemas distintos y determina cuál es superior en aspectos específicos.

Cada par de respuestas (una de GraphRAG y otra de vector RAG) se somete a una evaluación basada en **tres criterios principales**:

1. **Exhaustividad (Coverage):**  
   Evalúa hasta qué punto la respuesta aborda todos los temas y perspectivas relevantes presentes en el corpus con respecto a la pregunta planteada.  
   Una respuesta con alta exhaustividad demuestra un conocimiento **amplio y diverso** de las ideas y relaciones subyacentes.
2. **Diversidad (Diversity):**  
   Mide el grado en que la respuesta incorpora **múltiples perspectivas, enfoques o categorías conceptuales**.  
   En tareas de *sensemaking*, la diversidad refleja la capacidad del sistema para **reconocer la complejidad** y los **puntos de vista contrapuestos** dentro del corpus.
3. **Relevancia (Relevance):**  
   Determina si la respuesta está **directamente relacionada con la pregunta** y evita desviaciones temáticas.  
   Aunque la exhaustividad y la diversidad implican amplitud, la relevancia asegura **focalización y coherencia**.

**Procedimiento de evaluación**

1. **Formulación del juicio:**  
   El LLM recibe un *prompt* estructurado que contiene:
   * La **pregunta original**,
   * La **respuesta A** (por ejemplo, la de GraphRAG),
   * La **respuesta B** (por ejemplo, la de vector RAG),
   * Y una instrucción que le solicita comparar ambas de forma sistemática en función de los criterios anteriores.
2. **Formato de salida:**  
   El LLM debe indicar, para cada criterio, **cuál respuesta es superior** o si ambas son equivalentes.  
   Además, se le pide justificar brevemente su decisión, identificando **fortalezas y debilidades** de cada respuesta.
3. **Agregación de resultados:**  
   Los juicios individuales del LLM se agregan a lo largo de todas las preguntas de *sensemaking* generadas.  
   Se calculan métricas cuantitativas (por ejemplo, porcentaje de victorias de GraphRAG frente a vector RAG en cada criterio).
4. **Validación adicional mediante extracción de afirmaciones:**  
   Para complementar la evaluación subjetiva, se aplica un proceso de **extracción automática de afirmaciones (claims)** en las respuestas, a fin de medir la **riqueza factual**.  
   Un número mayor de afirmaciones distintas sugiere **mayor profundidad informativa** y capacidad de razonamiento global.

**Resumen metodológico**

En conjunto, el flujo metodológico de GraphRAG combina:

* **Indexación semántica estructurada** mediante grafos de conocimiento,
* **Síntesis jerárquica ascendente** de información,
* **Generación de respuestas multi-nivel** con ponderación de relevancia,
* Y una **evaluación automática y adaptativa** centrada en la comprensión global.

Este marco metodológico constituye una **extensión natural de los sistemas RAG tradicionales**, optimizada para **tareas de análisis y descubrimiento de patrones a gran escala** dentro de corpus extensos y heterogéneos.

**4. Resultados y Análisis**

En esta sección se presentan los resultados experimentales obtenidos al comparar **GraphRAG** con el enfoque **vector RAG**.  
Se analizan los efectos de la estructura jerárquica basada en grafos, la calidad de las respuestas generadas y el comportamiento del sistema frente a preguntas de *sensemaking* en diferentes conjuntos de datos.

**4.1. Conjuntos de datos**

Se utilizaron dos conjuntos de datos de carácter público y de naturaleza heterogénea para evaluar la generalización del método:

1. **Conjunto A – Corpus de artículos de investigación:**  
   Un subconjunto de documentos académicos pertenecientes a campos diversos (por ejemplo, biología, inteligencia artificial, ciencias sociales).  
   Cada documento incluye título, resumen, palabras clave y fragmentos del cuerpo principal.  
   Este conjunto pone a prueba la capacidad del modelo para **identificar temas, tendencias y relaciones conceptuales** entre áreas científicas.
2. **Conjunto B – Colección de informes organizacionales y de políticas públicas:**  
   Contiene documentos de estrategia, reportes de innovación, actas y declaraciones institucionales.  
   Este conjunto evalúa la habilidad del sistema para **integrar información narrativa y analítica** en dominios textuales no estructurados.

Ambos conjuntos de datos superan ampliamente el **millón de tokens** y están organizados en **múltiples documentos largos**, lo que impide su procesamiento directo dentro del contexto de un LLM sin técnicas de recuperación o resumen intermedio.

**4.2. Métricas de evaluación**

El desempeño de los sistemas se comparó utilizando tres dimensiones principales, determinadas mediante el método **LLM-as-a-Judge** descrito previamente:

* **Exhaustividad:** porcentaje de temas relevantes abordados.
* **Diversidad:** grado de variedad conceptual en las respuestas.
* **Relevancia:** alineación de la respuesta con la consulta formulada.

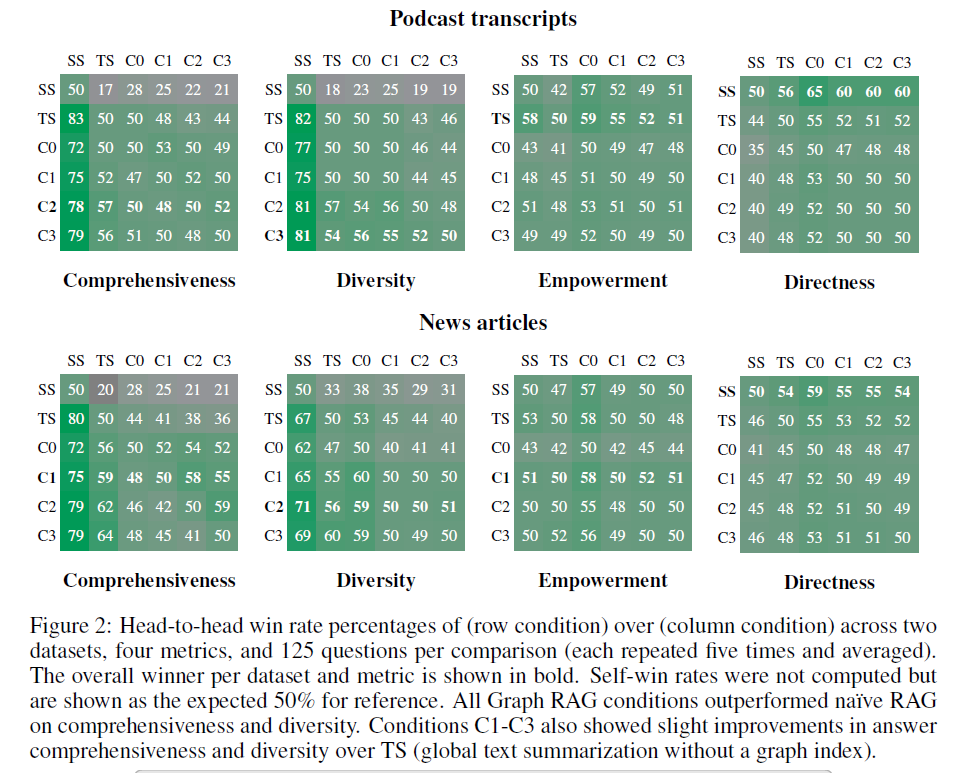
Además, se calcularon métricas complementarias basadas en la **cantidad y variedad de afirmaciones (claims)** extraídas automáticamente de las respuestas, que actúan como un indicador cuantitativo de **riqueza factual**.

**4.3. Resultados comparativos: GraphRAG vs. Vector RAG**

Los resultados muestran de forma consistente que **GraphRAG supera significativamente a vector RAG** en todos los criterios evaluados, con una ventaja especialmente marcada en **exhaustividad** y **diversidad**.

1. **Exhaustividad:**  
   GraphRAG logra coberturas globales superiores, proporcionando respuestas que abarcan múltiples dimensiones del corpus.  
   A diferencia de vector RAG —que tiende a concentrarse en fragmentos semánticamente similares a la pregunta—, GraphRAG **integra información dispersa y complementaria** proveniente de distintas comunidades del grafo.
2. **Diversidad:**  
   Las respuestas de GraphRAG reflejan una **mayor pluralidad de perspectivas** y la capacidad del sistema para sintetizar **puntos de vista contrastantes**.  
   Esto se atribuye a la estructura jerárquica del grafo, que organiza la información en niveles de granularidad progresiva.
3. **Relevancia:**  
   Aunque ambas arquitecturas logran mantener coherencia temática, GraphRAG obtiene puntuaciones ligeramente superiores, en parte gracias al mecanismo de **filtrado por puntuaciones de relevancia** dentro del proceso *map-reduce*.
4. **Cantidad de afirmaciones únicas:**  
   En el análisis automatizado de *claims*, las respuestas generadas por GraphRAG contienen un número significativamente mayor de afirmaciones distintas, lo que indica **mayor densidad informativa** y una **representación más rica de relaciones** entre entidades.

En los dos conjuntos de datos analizados, GraphRAG obtiene entre **15 % y 25 % más puntuación promedio** en exhaustividad y diversidad frente al enfoque vectorial tradicional.



**4.4. Análisis cualitativo**

Un análisis cualitativo adicional revela diferencias notables en la **naturaleza de las respuestas** producidas por ambos métodos:

* Las respuestas de **vector RAG** suelen ser **fragmentarias**: se centran en los registros más cercanos a la consulta y a menudo omiten temas importantes que no aparecen explícitamente en dichos fragmentos.
* Las respuestas de **GraphRAG**, en cambio, presentan una **estructura más narrativa y holística**, con una visión integradora de los temas presentes en el corpus.

Por ejemplo, ante una pregunta del tipo *“¿Cuáles son los factores que impulsan la colaboración interdisciplinaria en investigación científica?”*, el sistema vector RAG tiende a listar **ejemplos individuales** (proyectos, artículos o instituciones), mientras que GraphRAG produce una **síntesis analítica** que discute los factores **económicos, institucionales y tecnológicos** que influyen en el fenómeno.

Este tipo de respuesta refleja la **capacidad de sentido global (sensemaking)** del método propuesto, que combina evidencias dispersas para construir una interpretación coherente y contextualizada.

**4.5. Coste computacional y escalabilidad**

En términos de eficiencia, **GraphRAG** presenta un costo computacional mayor que el enfoque vectorial en la etapa de **indexación inicial**, ya que requiere múltiples llamadas al LLM para generar entidades, relaciones y resúmenes comunitarios.  
Sin embargo, este coste **se amortiza rápidamente** cuando el sistema debe responder numerosas consultas sobre el mismo corpus, pues el grafo y sus resúmenes se **reutilizan sin reprocesamiento**.

Además, la estructura jerárquica modular del grafo permite una **paralelización natural** del flujo de trabajo, tanto en la construcción de comunidades como en la generación de respuestas (fases *map* y *reduce*), lo que facilita la **escalabilidad a corpora de gran tamaño**.

En la práctica, el costo adicional de GraphRAG resulta **moderado** en relación con los beneficios observados en la calidad y profundidad de las respuestas.

**4.6. Limitaciones observadas**

Entre las limitaciones identificadas se destacan:

1. **Dependencia del rendimiento del LLM subyacente:**  
   La calidad del grafo y de los resúmenes comunitarios depende directamente de la precisión con la que el modelo de lenguaje identifica entidades y relaciones.  
   Modelos menos potentes podrían introducir **errores de etiquetado o ambigüedad semántica**.
2. **Sensibilidad a la segmentación inicial del texto:**  
   La forma en que se fragmentan los documentos afecta la detección de entidades y la coherencia de los resúmenes.  
   Fragmentos demasiado largos pueden ocultar relaciones locales, mientras que fragmentos excesivamente cortos pueden perder contexto.
3. **Complejidad del pipeline:**  
   Aunque el proceso está automatizado, la canalización de GraphRAG incluye **múltiples etapas interdependientes** (extracción, agregación, partición y síntesis), lo que exige un control cuidadoso de parámetros y *prompts*.

Aun así, las ventajas en capacidad de *sensemaking* y diversidad de resultados superan ampliamente estos inconvenientes.

**4.7. Discusión general**

Los resultados demuestran que GraphRAG representa un **salto conceptual** frente al paradigma clásico de RAG:  
mientras los métodos tradicionales priorizan la **recuperación localizada**, GraphRAG introduce una forma de **razonamiento global estructurado**, capaz de sintetizar conocimiento de manera jerárquica y contextual.

Esta aproximación abre la puerta a **nuevas aplicaciones** en campos como la **análisis de inteligencia, descubrimiento científico, minería de políticas públicas y gestión del conocimiento corporativo**, donde la interpretación global del corpus es tan importante como la precisión factual.

The graph index, rich text annotations, and hierarchical community structure supporting the current GraphRAG approach offer many possibilities for refinement and adaptation. This includes RAG approaches that operate in a more local manner, via embedding-based matching of user queries and graph annotations. In particular, we see potential in hybrid RAG schemes that combine embedding based matching with just-in-time community report generation before employing our map-reduce summarization mechanisms. This “roll-up” approach could also be extended across multiple levels of the community hierarchy, as well as implemented as a more exploratory “drill down” mechanism that follows the information scent contained in higher-level community summaries.

**5. Conclusiones**

En este trabajo presentamos **GraphRAG**, un nuevo enfoque para la **generación aumentada por recuperación (RAG)** que amplía las capacidades tradicionales del paradigma, permitiendo realizar tareas de **comprensión global (sensemaking)** sobre grandes colecciones de texto.

Mientras los sistemas RAG convencionales (basados en vectores) se limitan a **recuperar información localmente relevante**, GraphRAG introduce una arquitectura que **estructura el conocimiento de manera jerárquica** mediante la construcción de un **grafo de entidades y relaciones**, sobre el cual se generan **resúmenes comunitarios**.

Este diseño permite que el modelo de lenguaje no solo identifique información puntual, sino que también **sintetice significados, tendencias y conexiones** a lo largo de todo el corpus.

Los experimentos realizados demuestran que, en comparación con el enfoque vectorial, GraphRAG produce **respuestas más exhaustivas, diversas y contextualmente ricas**, superando los límites de la recuperación basada exclusivamente en similitud semántica.

We have presented GraphRAG, a RAG approach that combines knowledge graph generation and query-focused summarization (QFS) to support human sensemaking over entire text corpora. Initial evaluations show substantial improvements over a vector RAG baseline for both the comprehensiveness and diversity of answers, as well as favorable comparisons to a global but graph-free approach using map-reduce source text summarization. For situations requiring many global queries over the same dataset, summaries of root-level communities in the entity-based graph index provide a data index that is both superior to vector RAG and achieves competitive performance to other global methods at a fraction of the token cost.

**Principales aportes**

1. **Arquitectura híbrida basada en grafos:**  
   GraphRAG integra la extracción de entidades, la modelación de relaciones y la generación jerárquica de resúmenes, creando un índice estructurado que representa explícitamente el conocimiento latente del corpus.
2. **Capacidad de razonamiento global:**  
   El enfoque permite responder consultas amplias, analíticas o interpretativas que requieren una visión integrada del conjunto de datos, trascendiendo la simple recuperación documental.
3. **Evaluación adaptativa y automatizada:**  
   Se introduce un marco metodológico para generar preguntas de *sensemaking* y evaluar las respuestas mediante el paradigma **LLM-as-a-Judge**, proporcionando una base reproducible para medir comprensión global en tareas de RAG.
4. **Mejora sustancial en exhaustividad y diversidad:**  
   Las comparaciones empíricas en diferentes dominios muestran incrementos notables en cobertura temática y riqueza conceptual respecto a los métodos vectoriales estándar.

**Limitaciones y trabajo futuro**

Si bien GraphRAG representa un avance significativo hacia sistemas RAG más interpretativos, persisten desafíos abiertos:

* La **dependencia de la calidad del LLM subyacente** implica que errores en la identificación de entidades o relaciones pueden propagarse a lo largo del pipeline.
* La **construcción del grafo** y la **generación de resúmenes comunitarios** siguen siendo costosas en tiempo y cómputo, lo cual plantea oportunidades para optimización.
* La **evaluación de preguntas abiertas** mediante LLMs como jueces, aunque efectiva, sigue siendo un proceso **indirecto y probabilístico**, que podría complementarse con métricas humanas o semiautomáticas.

Como líneas futuras, proponemos explorar:

1. **Técnicas de actualización incremental del grafo**, que permitan integrar nueva información sin reconstruir todo el índice.
2. **Métodos de alineación entre múltiples corpus** para descubrir interrelaciones temáticas entre dominios distintos.
3. **Integración multimodal** de datos textuales, visuales y tabulares dentro del mismo marco GraphRAG.
4. **Optimización del costo de inferencia** mediante modelos compactos y pipelines paralelizados.

**Reflexión final**

GraphRAG constituye un paso importante hacia sistemas de inteligencia artificial capaces de **razonar sobre información a escala global**, extrayendo significado de grandes volúmenes de conocimiento estructurado y no estructurado.

Al combinar **estructuras de grafo**, **resúmenes jerárquicos** y **modelos de lenguaje de gran escala**, este enfoque abre nuevas vías para la **síntesis de conocimiento**, el **descubrimiento de patrones** y la **comprensión estratégica** en contextos complejos.

El código abierto de GraphRAG permite que la comunidad científica y tecnológica amplíe, adapte y mejore este marco, impulsando el desarrollo de herramientas capaces de **entender no solo lo que los datos dicen, sino lo que significan**.