INFORME TÉCNICO DE INVESTIGACIÓN

ESTADO DEL ARTE RAG 2024-2025

1. DEFINICIÓN Y PROPÓSITO

La Generación Aumentada por Recuperación (RAG, por sus siglas en inglés: *Retrieval-Augmented Generation*) es una técnica de IA que permite a los Grandes Modelos de Lenguaje (LLM) consultar datos externos y actualizados en tiempo real antes de generar una respuesta.

Su propósito principal es resolver los dos grandes problemas de los LLMs actuales:

- Alucinaciones: Cuando el modelo inventa información con total seguridad.
- Desactualización: Los modelos solo saben lo que aprendieron durante su entrenamiento (su "fecha de corte").

2. ACTUALIDAD DEL PLANTEAMIENTO (2024)

Hoy en día, el RAG se ha convertido en el estándar para aplicaciones empresariales. Ya no basta con un "chat" genérico; las empresas necesitan que la IA conozca sus documentos internos, manuales técnicos y bases de datos privadas.

Fases Críticas del Proceso:

1. Recuperación (Retrieval): El sistema busca en una base de datos vectorial los fragmentos de información más relevantes para la consulta del usuario.

2. Generación (Generation): El LLM recibe la consulta original + la información recuperada (contexto) y redacta una respuesta coherente y verídica.

3. TENDENCIAS Y EVOLUCIÓN PARA 2025

El planteamiento RAG está evolucionando hacia sistemas mucho más complejos y capaces:

A. Agentic RAG (RAG Agéntico)

No es solo buscar y pegar. Ahora, agentes inteligentes planifican la búsqueda, evalúan si el resultado es útil y, si no es suficiente, reformulan la consulta o buscan en otras fuentes de forma autónoma.

B. GraphRAG (RAG basado en Grafos)

En lugar de solo buscar texto por "similitud", se utilizan grafos de conocimiento que conectan conceptos. Esto permite a la IA entender relaciones complejas que un sistema vectorial simple podría pasar por alto.

C. RAG Multimodal

La capacidad de recuperar y generar respuestas no solo basadas en texto, sino integrando imágenes, esquemas técnicos, audio y datos estructurados de forma simultánea.

4. POSIBILIDADES Y APLICACIONES PRÁCTICAS

Las posibilidades son prácticamente infinitas en el entorno profesional actual:

- Atención al Cliente: Asistentes que conocen cada detalle de los productos de una marca sin fallar.
- Investigación Legal y Médica: Análisis instantáneo de miles de normativas o historiales clínicos buscando patrones específicos.
- Marketing y SEO: Generación de contenido basada en tendencias recogidas en tiempo real de internet.
- Consultoría Interna: Sistemas que actúan como el "cerebro" de una empresa, permitiendo a los empleados consultar cualquier dato corporativo en segundos.

5. CONCLUSIÓN

El RAG no es una moda pasajera, sino el puente necesario para que la IA pase de ser un "juguete curioso" a una herramienta de producción masiva, fiable y segura para el sector empresarial.