Propuesta de Trabajo Fin de Grado

Título: Mitigación de alucinaciones en LLMs mediante el desarrollo de un sistema RAG para la verificación de noticias.

1. Definición del Problema y Motivación

Los Modelos de Lenguaje (LLMs) han demostrado una capacidad excepcional para generar texto coherente. Sin embargo, sufren de dos limitaciones críticas cuando se aplican al dominio periodístico:

1. **Alucinaciones:** La generación de hechos plausibles pero falsos.
2. **Obsolescencia del Conocimiento:** Los modelos tienen una fecha de corte de entrenamiento y desconectan de la actualidad en tiempo real.

Este TFG propone abordar estos problemas mediante la arquitectura **RAG (Retrieval-Augmented Generation)**, que desacopla el razonamiento (LLM) del conocimiento (Base de Datos Vectorial/Buscador), permitiendo verificar noticias recientes (2024-2025) sin necesidad de reentrenar el modelo.

2. Objetivos del Proyecto

2.1. Objetivo General

Diseñar y evaluar un sistema RAG capaz de reducir la tasa de alucinaciones en modelos de lenguaje (tanto SLMs como LLMs) al responder preguntas sobre actualidad, utilizando un corpus de noticias reales como fuente de verdad (Ground Truth).

2.2. Objetivos Específicos

1. **Implementación de Infraestructura de Recuperación:** Construir una base de conocimiento indexada y optimizada (Elasticsearch) a partir de un dataset de noticias no estructuradas.
2. **Desarrollo de Metodología de Evaluación (Prompting):** Diseñar una estrategia para generar automáticamente una "batería de preguntas de control" basada en el contenido real de las noticias, permitiendo distinguir cuándo el modelo responde con datos y cuándo inventa.
3. **Estudio Comparativo Multimodelo:** Analizar el impacto del tamaño del modelo en la mitigación de alucinaciones, comparando el rendimiento de **Small Language Models (SLMs)** frente a **Large Language Models (LLMs)**.
4. **Análisis de Estrategias de Recuperación:** Evaluar cómo influyen diferentes métodos de búsqueda (búsqueda léxica BM25, búsqueda semántica, variación de Top-K) en la calidad de la respuesta generada.

3. Metodología Experimental (El núcleo del TFG)

El proyecto se estructura en torno a una comparativa experimental que responderá a las siguientes preguntas de investigación:

A. Generación de la Batería de Preguntas (Mitigación de Alucinaciones)

Para medir la alucinación, se necesita un estándar de verdad. Se desarrollará un sistema automatizado que:

- Extraiga hechos clave de las noticias (fechas, entidades, eventos).
- Genere pares de **Pregunta - Respuesta Esperada** utilizando técnicas de *Prompt Engineering*.
- Clasifique las preguntas en categorías (Factuales, Temporales, Adversarias) para "estresar" al modelo y detectar fallos de razonamiento.

B. Matriz de Comparación (Escenarios)

Se evaluará la precisión y fidelidad (*faithfulness*) de las respuestas en tres escenarios distintos:

Escenario	Descripción	Hipótesis a Validar
1. Baseline (Sin RAG)	El modelo (SLM o LLM) responde usando solo su memoria pre-entrenada.	Se espera una alta tasa de alucinación y desconocimiento de eventos recientes (2024-2025).
2. RAG con SLMs	Modelos pequeños (ej. Phi-2, TinyLlama) aumentados con contexto recuperado.	¿Puede un modelo pequeño y eficiente igualar a un gigante si tiene acceso a buenos datos externos?
3. RAG con LLMs	Modelos grandes (ej. Llama-3, Mistral) aumentados con contexto recuperado.	Se espera el mejor rendimiento, combinando capacidad de razonamiento con datos actualizados.

4. Fases de Desarrollo (Pasos a Seguir)

1. **Ingeniería de Datos (ETL):**
    - Recopilación y limpieza del corpus de noticias.
    - Estructuración de metadatos (Fuente, Fecha, Cuerpo) para optimizar la búsqueda.
    - *Estado: Avanzado (Indexación en Elasticsearch completada).*
2. **Diseño del Sistema de Recuperación (Retrieval):**
    - Configuración del motor de búsqueda.
    - Ajuste de algoritmos de relevancia (BM25, Boosting de títulos).
    - *Estado: Avanzado (Validación de búsquedas realizada).*
3. **Integración del Modelo Generativo (Generation):**
    - Despliegue de modelos en entorno GPU.
    - Desarrollo del pipeline RAG: Conexión *Query -> Buscador -> Contexto -> Prompt -> LLM*.
4. **Creación del Dataset de Evaluación:**

    - Implementación de scripts para generar preguntas de prueba a partir de las noticias indexadas.
5. **Ejecución de Experimentos y Análisis:**
    - Lanzar la batería de preguntas sobre los distintos modelos.
    - Medir métricas de calidad (Precisión factual, ausencia de alucinaciones).
    - Comparar resultados entre SLMs y LLMs.

5. Herramientas y Tecnologías

- **Lenguaje:** Python.
- **Base de Datos Vectorial/Buscador:** Elasticsearch.
- **Modelos de Lenguaje:** Hugging Face Transformers (Llama, Mistral, Gemma, Phi).
- **Entorno de Ejecución:** Servidores Linux con aceleración GPU (CUDA).
-