Estrategias de RAG y evaluación de preguntas en modelos de lenguaje

Felipe Clariá, Santiago Fada
FAMaFyC, UNC, Argentina

Abstract

Este informe de avances presenta los resultados iniciales y los desafíos encontrados en el análisis de la efectividad de Retrieval Augmented Generation (RAG) en comparación con modelos de lenguaje tradicionales (Large Language Models, LLMs). Manteniendo los objetivos originales, el estudio se enfocó en encontrar contextos en los que el uso de RAG proporciona ventajas en la precisión y relevancia de las respuestas. La metodología incluyó la definición de preguntas y el análisis de las respuestas generadas por un modelo al que se le proporciona un contexto de dominio específico. Los resultados sugieren que una correcta configuración del modelo puede mitigar alucinaciones y mejorar la relevancia de las respuestas en función del contexto proporcionado. Se proponen pasos futuros para optimizar la evaluación y extender el alcance del modelo.

Introducción

El uso de Retrieval Augmented Generation (RAG) en modelos de lenguaje ha emergido como una solución prometedora para mejorar la precisión y relevancia de las respuestas generadas, especialmente en consultas que requieren información específica o actualizada [1]. Aunque los LLMs han logrado un notable éxito en diversas tareas de procesamiento de texto, todavía enfrentan limitaciones en tareas de dominio específico y son propensos a generar "alucinaciones" cuando las consultas sobrepasan sus datos de entrenamiento o exigen información actualizada [2].

En este contexto, el enfoque RAG combina la recuperación de información con la generación de texto, ofreciendo el potencial de proporcionar respuestas más exactas y contextualizadas. Este proyecto busca evaluar los contextos en los cuales RAG ofrece ventajas sobre los modelos tradicionales de lenguaje, explorando su capacidad para generar respuestas precisas y adecuadas. La metodología sigue una estructura de generación y evaluación de respuestas para distintos tipos de preguntas: algunas resolubles solo con el conocimiento del modelo y otras que requieren un contexto adicional.

Objetivos

El proyecto conserva sus objetivos iniciales: encontrar y analizar los contextos en los cuales el uso de RAG ofrece ventajas sobre los modelos de lenguaje tradicionales (LLMs). En particular, se busca determinar los tipos de preguntas en los que el uso de RAG proporciona respuestas más precisas y relevantes.

Metodología

La experimentación se llevó a cabo empleando **Ollama** [3], una herramienta de uso libre que permite ejecutar LLMs de manera local, reduciendo la dependencia de conexiones a servidores externos y ofreciendo un mejor control y mayor privacidad de los datos. Usamos el modelo y los embeddings de **Llama2** [4], que al ser de peso ligero proporcionan una buena *performance* para equipos locales.

Para el desarrollo técnico, se utilizó **Jupyter Notebooks** y el framework **LangChain** [5] para estructurar el flujo de trabajo con RAGs, unificando la recuperación de información y la comunicación con el modelo para realizar consultas y generar respuestas. Se utilizó además **Pinecone** [6] para gestionar los embeddings, proporcionando una recuperación rápida y relevante.

Procedimiento:

1. **Selección del Contexto**: Se eligieron los diez primeros capítulos del documento *Git Notes for Professionals* [7] desarrollado por GoalKicker, para proporcionar un contexto adecuado en el dominio de comandos de Git.
2. **Clasificación de Preguntas**: Las preguntas se dividieron en:
    - Preguntas de contexto cerrado.
    - Preguntas contextuales.
    - Preguntas ambiguas o complejas.
    - Preguntas especializadas.
3. **Evaluación de Respuestas**: Las consultas se realizaron con y sin el contexto definido, unificando las preguntas con una plantilla general.

Resultados Obtenidos

En el archivo de resultados adjunto se incluyen las preguntas definidas, la plantilla de consultas utilizada y las respuestas obtenidas en ambas condiciones (con y sin contexto).

1. **Precisión y "Alucinaciones"**:
    - Observamos respuestas precisas incluso cuando la información no estaba en el documento.

    - Hubo presencia de "alucinaciones" en algunas respuestas, añadiendo detalles irrelevantes o incorrectos.
2. **Impacto del Contexto en la Precisión**:
    - Algunas respuestas fueron menos precisas con contexto, priorizando fragmentos textuales cercanos.
3. **Mejora con Nueva Plantilla**:
    - Se definió una nueva plantilla indicando que toda la información debía provenir del contexto.
    - Los resultados fueron más precisos, eliminando en gran medida las alucinaciones.

Problemas Encontrados

- **Dificultades Técnicas**: Complicaciones con el versionado de librerías y ejecución del entorno.
- **Evaluación Subjetiva**: Dependencia de evaluaciones anecdóticas para valorar las respuestas, lo que dificulta establecer criterios objetivos.

Siguientes Pasos

1. **Métricas de Evaluación**: Incorporar indicadores objetivos para una valoración cuantitativa.
2. **Ampliación del Alcance**: Probar otros contextos y preguntas para una evaluación más robusta.
3. **Incorporación de Data-Scraping**: Permitir acceso a información más diversa y actualizada para respuestas en tiempo real.

Conclusiones

Los avances realizados hasta ahora indican que el uso de RAG, con la configuración adecuada, tiene el potencial de mejorar significativamente la calidad y relevancia de las respuestas generadas por los modelos de lenguaje. Sin embargo, aún se requieren ajustes en la plantilla de consultas para optimizar el uso del contexto proporcionado y reducir de manera efectiva la aparición de alucinaciones o información incorrecta.

Los próximos pasos se centrarán en refinar las métricas de evaluación para lograr una valoración más objetiva del rendimiento, así como en explorar nuevos contextos y dominios para extender el alcance del análisis y evaluar la efectividad del enfoque en situaciones variadas.

Referencias

1. Yunfan Gao et al. *Retrieval-Augmented Generation for Large Language Models: A Survey*. 2024. arXiv: 2312.10997 [cs.CL]. Disponible en: https://arxiv.org/abs/2312.10997.
2. Nikhil Kandpal et al. *Large Language Models Struggle to Learn Long-Tail Knowledge*. 2023. arXiv: 2211.08411 [cs.CL]. Disponible en: https://arxiv.org/abs/2211.08411.
3. Ollama
4. Llama2
5. LangChain
6. Pinecone
7. Git Notes for Professionals