Estrategias de RAG y evaluación de preguntas en modelos de lenguaje

Felipe Clariá, Santiago Fada

FAMaFyC, UNC, Argentina

Resumen

El uso de Retrieval Augmented Generation (RAG) ha ganado popularidad como una técnica para mejorar la generación de respuestas en modelos de lenguaje natural. Esta metodología permite integrar información externa al modelo con el objetivo de proporcionar contexto a las respuestas generadas. En este estudio, exploramos cómo RAG puede optimizar la respuesta a preguntas complejas y especializadas, mejorando la calidad de las interacciones con los modelos de lenguaje.

Introducción

Los Large Language Model (LLM) han alcanzado un éxito notable en diversas tareas de procesamiento de texto, pero todavía enfrentan limitaciones en tareas específicas de dominio o que requieren un alto nivel de conocimiento [1]. Estos modelos son propensos a producir "alucinaciones" cuando se les presentan consultas que están más allá de sus datos de entrenamiento o que exigen información actualizada.

En este contexto, el enfoque de Retrieval Augmented Generation (RAG) se presenta como una solución prometedora. Al integrar la recuperación de información con la generación de texto, RAG tiene el potencial de proporcionar respuestas más precisas y contextualizadas, superando algunas de las limitaciones inherentes a los modelos de lenguaje tradicionales [2].

Este proyecto se centra en investigar las ventajas de RAG, clasificar los tipos de preguntas que son más adecuadas para este enfoque y analizar la efectividad de estas estrategias en comparación con modelos que no las implementan.

Objetivos

El proyecto busca explorar y analizar los contextos en los que la utilización de RAG ofrece ventajas significativas sobre los modelos de lenguaje tradicionales (LLM). El objetivo principal es identificar los tipos de preguntas que se responden de manera más efectiva con este enfoque. Evaluaremos cómo el uso de RAG puede mejorar la precisión, relevancia y calidad de las respuestas del modelo, en comparación con LLMs que no implementen estas estrategias.

Hipótesis

Proponemos que el uso de RAG proporcionará respuestas más precisas y contextualmente relevantes que un LLM tradicional en la mayoría de los casos, y creemos que esto será especialmente evidente con preguntas que requieren información actualizada o específica de un dominio acotado.

Metodología

Los experimentos se realizarán dentro de un dominio de textos a definir, que abarcarán conocimientos específicos sobre ciencias, historia y cultura. Estos dominios deberán ser lo suficientemente amplios para permitir la recolección de datos significativos y facilitar el entrenamiento y la posterior evaluación de modelos especializados.

Se desarrollará un conjunto de preguntas con la intención de evaluar la generación de respuestas. Las preguntas se clasificarán en las siguientes categorías:

Preguntas de contexto cerrado Resueltas dentro del conocimiento del modelo.

Preguntas contextuales Requieren información adicional no contenida en el modelo.

Preguntas ambiguas o complejas No pueden ser resueltas adecuadamente con o sin información adicional debido a la falta de datos precisos.

Preguntas especializadas Implican la necesidad de información filtrada o criterios específicos.

Planificación futura

- **Semana 1:** Revisión de literatura sobre RAG y definición de contextos específicos. Especificación y formulación de preguntas.
- **Semana 2:** Extracción y preparación de datos de fuentes relevantes. Implementación de técnicas de preprocesamiento y limpieza de datos.
- **Semana 3:** Desarrollo y entrenamiento de modelos utilizando los datos extraídos. Evaluación inicial de la calidad de las respuestas generadas mediante preguntas de contexto cerrado y contextuales.
- **Semana 4:** Experimentación con preguntas ambiguas y especializadas, analizando la efectividad de RAG en estos casos.
- **Semana 5:** Elaboración del informe final con las comparaciones realizadas y las conclusiones.

Referencias

[1] Nikhil Kandpal et al. *Large Language Models Struggle to Learn Long-Tail Knowledge*. 2023. arXiv: 2211.08411 [cs.CL]. URL: https://arxiv.org/abs/2211.08411.

[2] Yunfan Gao et al. *Retrieval-Augmented Generation for Large Language Models: A Survey*. 2024. arXiv: 2312.10997 [cs.CL]. URL: https://arxiv.org/abs/2312.10997.