Aplicaciones LLM

Desafíos de la técnica RAG

Desafíos de la técnica RAG

- Desafíos en el proceso de recuperación (retrieval).
- Desafíos en el proceso de respuesta (response).

Desafíos en el proceso de recuperación (retrieval)

- Baja precisión: no todos los fragmentos (chunks) en el conjunto recuperado son relevantes.
  - Problemas de alucinación y “lost in the middle”.
  - Hay mucho "relleno" (fluff) en la respuesta devuelta.
- Low recall: no se recuperan todos los fragmentos relevantes.
  - Falta suficiente contexto para que LLM sintetice una respuesta.
- Información desactualizada (outdated): los datos son redundantes o están desactualizados.

Desafíos del proceso de respuesta (response)

- Alucinación: el modelo inventa una respuesta que no está en el contexto.
- Irrelevancia: el modelo inventa una respuesta que no responde a la pregunta.
- Toxicidad/Sesgo (Bias): el modelo inventa una respuesta que es perjudicial/ofensiva.

Formas de superar los desafíos

- Puedes mejorar las cosas en todas las etapas del proceso.
  - Datos.
  - Incrustaciones (embeddings).
  - Recuperación (retrieval).
  - Síntesis (generación de respuesta).

- Antes de introducir mejoras en todas estas áreas, debes tener métricas listas para poder medir el impacto de estos cambios en el rendimiento de la aplicación.

Mejora de los datos

- ¿Puedes almacenar información adicional además de los fragmentos de texto (chunks) sin procesar?
  - Juega con los tamaños de los fragmentos.

Mejora de los embeddings

- ¿Puedes optimizar las representaciones de los embeddings?
  - La configuración predeterminada (default settings) se puede mejorar.

Mejora de la técnica de recuperación (retrieval)

- ¿Puedes encontrar una técnica mejor que la técnica de búsqueda de embeddings top-k?

Mejora de la síntesis (generación de la respuesta)

- ¿Puedes utilizar LLMs para más tareas que la generación de la respuesta? Podrías usar el LLM para tareas de razonamiento (reasoning). Ejemplos:
  - Dada una pregunta, ¿puedes descomponerla en preguntas más simples?
  - Enrutamiento (route) a diferentes fuentes de datos.
  - Tener una forma más sofisticada de consultar (querying) tus datos.