Aplicaciones LLM

Optimización de las aplicaciones RAG

Optimización de sistemas RAG

- Una vez que tengas métricas para medir la mejora del rendimiento puedes proceder con la optimización del sistema RAG.

- Dispones de varias técnicas de optimización, que a continuación se listan desde las más simples y económicas hasta las más avanzadas y costosas:
  - Técnicas de optimización inicial.
  - Métodos avanzados de recuperación (Advanced Retrieval Methods).
  - Fine-tuning.
  - Uso de agentes.

Técnicas iniciales de optimización

- Mejores analizadores (parsers).
- Tamaños de fragmentos (chunks).
- Búsqueda híbrida.
- Filtros de metadata.

Notas sobre ajustar el tamaño de los chunks

- Ajustar los tamaños de los fragmentos (chunks) puede tener un impacto en el rendimiento.
- Más tokens recuperados (retrieved tokens) no siempre se traduce en un mejor rendimiento.
- El reordenamiento del orden del contexto (shuffling context order) no siempre es beneficioso.

Notas sobre añadir metadata

- Metadata: contexto que se puede inyectar en cada chunk de texto.
- Es como un diccionario JSON estructurado.
- Ejemplos de datos incluidos en metadatos:
  - Número de página.
  - Título del documento, año.
  - Resumen de fragmentos adyacentes (adjacent chunks).
  - Preguntas que el fragmento puede responder (reverse HyDE).
- Beneficios:
  - Puede ayudar en la recuperación.
  - Puede aumentar la calidad de la respuesta.
  - Se integra con los filtros de metadatos de bases de datos vectoriales.

Advanced Retrieval Methods

- Reordenamiento (reranking).
- Recuperación recursiva (recursive retrieval).
- Tablas incrustadas (embedded tables).
- Recuperación de pequeño a grande (Small-to-Big retrieval).

Notas sobre la Small-to-Big retrieval

- Incrustar fragmentos de texto (chunks) grandes no suele ser óptimo.

- Soluciones:
  - Incrustar texto por frases y luego ampliar esa ventana durante la síntesis de LLM.
  - Incrustar una referencia al fragmento principal. Usar el fragmento principal para la síntesis.

- Esto lleva a una recuperación más precisa y evita problemas de tipo "lost in the middle".

Fine-tuning

- Embedding fine-tuning.
- LLM fine-tuning.

Notas sobre el fine-tuning de los embeddings

- Los embeddings no están optimizadas por defecto.

- Solución: generar un dataset de querys sintéticas a partir de chunks sin procesar utilizando LLMs. Utiliza este dataset sintético para hacer fine-tuning de los embeddings.

Uso de agentes.

- Routing.
- Query planning.
- Multi-document agents.

Notas sobre el uso de multi-document agents

- Hay ciertas preguntas que un sistema RAG "top-k" no puede responder.

- Solución: agentes de múltiples documentos.
  - QA basadas en hechos y resúmenes de subsets de documentos.
  - Chain-of-thought y query planning.