Evaluación

Descripción general de la evaluación

Para evaluar nuestro sistema, utilizaremos 20 consultas diferentes con diversos grados de dificultad para que el sistema genere resúmenes. Luego compararemos los resúmenes generados con resúmenes predefinidos manualmente para ver el rendimiento del sistema. Además, hay 5 consultas comodín que no están incluidas en las 20 consultas; son consultas intencionalmente complejas o formuladas incorrectamente para ver cómo se comporta el sistema en tales casos, si hay una respuesta incoherente o ninguna respuesta.

Métrica ROUGE para la evaluación

La métrica ROUGE (Recall-Oriented Understudy for Gisting Evaluation) se utiliza ampliamente para evaluar la calidad del resumen de texto y otras tareas de procesamiento del lenguaje natural. Mide la superposición entre un texto generado (por ejemplo, un resumen) y un texto de referencia. ROUGE se centra principalmente en la exhaustividad (recall), pero también se puede configurar para evaluar la precisión y la puntuación F1.

Estas son sus principales variantes:

1. ROUGE-N: Mide la superposición de n-gramas (por ejemplo, secuencias de n palabras) entre el texto generado y el texto de referencia.
   - ROUGE-1: Superposición de unigramas (palabras individuales).
   - ROUGE-2: Superposición de bigramas (secuencias de dos palabras).
2. ROUGE-L: Evalúa la subsecuencia común más larga (LCS) entre el texto generado y el de referencia, considerando la estructura de la secuencia.
3. ROUGE-W: Una variante ponderada de ROUGE-L que asigna más importancia a las coincidencias consecutivas.
4. ROUGE-S: Mide la superposición de skip-bigramas, lo que permite que las palabras coincidan incluso si no son adyacentes.

ROUGE es particularmente eficaz para tareas donde la salida generada debe parecerse mucho al texto escrito por humanos. Al incluirla en su sección de evaluación, usted enfatiza un enfoque cuantitativo y estandarizado para evaluar la calidad del texto generado.

En este caso particular, analizaremos las puntuaciones ROUGE-1 y ROUGE-L para los resúmenes generados, en particular la puntuación de exhaustividad (recall).

Métricas de evaluación

Clasificaremos 3 sistemas:

Nuestro sistema con una generación de subgrafos básica - lo llamaremos GRAPH-BASIC.

Nuestro sistema con una generación de subgrafos avanzada - lo llamaremos GRAPH-ADVANCED.

Un sistema RAG simple con incrustación vectorial de texto completo (los artículos mismos) - lo llamaremos RAG-TEXT.

Utilizaremos el modelo GPT-o1 de vanguardia como referencia.

En cuanto a las métricas que utilizaremos, son las siguientes:

Precisión: qué tan bien se desempeña el sistema en las 20 consultas. Esto tendrá 2 partes: las puntuaciones ROUGE-1 y ROUGE-L para los resúmenes generados; y una puntuación autogenerada sobre 5 del modelo de referencia, es decir, se le pedirá al modelo GPT-o1 que genere una puntuación

sobre 5 a partir de los resúmenes generados, la puntuación promedio se mostrará como Puntuación LLM.

Robustez: qué tan bien se desempeña el sistema en las consultas comodín. Una consulta bien manejada añade uno a la puntuación, lo que significa que una puntuación perfecta es 5.

Rendimiento: tiempo promedio necesario para generar la consulta.

Consultas de ejemplo

Algunas consultas de ejemplo que se utilizarán para la evaluación son:

P: "¿Cómo puedo calcular el interés compuesto de mis ahorros manualmente?"

R: "El interés compuesto se calcula utilizando la fórmula A = P(1 + r/n)^(nt), donde A es el valor futuro, P es el capital, r es la tasa de interés anual, n es el número de veces que se compone el interés por año y t es el tiempo en años".

P: "¿Cómo funciona el sistema de tramos impositivos para alguien que gana $60,000 al año?"

R: "Para un contribuyente soltero, sus ingresos se gravarán progresivamente. Para 2024, pagará el 10% sobre los ingresos hasta $11,000, el 12% sobre los ingresos de $11,001 a $44,725 y el 22% sobre los ingresos superiores a esa cantidad".

Ahora, para las consultas comodín, para probar si el sistema alucina o proporciona una respuesta incoherente a una pregunta ya incoherente:

P: "¿Cuánto ahorraré mensualmente si gano $4,000 pero gasto $6,000 cada mes?"

R: Aquí la respuesta correcta es señalar que si el usuario gasta más de lo que gana, entonces no estaría ahorrando dinero en absoluto.

P: "Si ahorro $100 semanales durante 2 años a una tasa de interés del 10%, ¿qué tendré en 5 años?"

R: Aquí el sistema debería señalar que faltan variables; si el usuario declara que ahorra $100 semanales durante 2 años, entonces el sistema no tiene forma de saber cuánto ahorra semanalmente durante los siguientes 3 años.

Resultados

Precisión

Los valores se aproximan al tercer decimal

	ROUGE-1		ROUGE-L		
Modelo	Exhaustividad	F1	Exhaustividad	F1	Puntuación LLM
GRAPH-BASIC	0.655	0.212	0.473	0.153	4.48
GRAPH-ADVANCED	0.891	0.083	0.744	0.069	4.86
RAG-TEXT	0.501	0.189	0.362	0.131	4.22

Rendimiento

Modelo	Tiempo promedio empleado (en s)
GRAPH-BASIC	15.407271986007691
GRAPH-ADVANCED	25.086044921875
RAG-TEXT	15.621680946350098

Robustez

Modelo	Respuestas comodín correctas (de 5)
GRAPH-BASIC	4
GRAPH-ADVANCED	5
RAG-TEXT	5

Resumen

Como podemos observar en los resultados, en términos de precisión, el modelo GRAPH-ADVANCED supera a los otros modelos tanto en las puntuaciones ROUGE como en la puntuación LLM. Esto es comprensible debido a la forma en que genera un subgrafo con mucho contexto y, por lo tanto, la respuesta generada necesita llenar menos lagunas en el conocimiento. El segundo mejor en ese sentido es el modelo GRAPH-BASIC, que aún supera al modelo RAG regular para esta tarea. Podemos ver que la puntuación F1 es la más alta tanto en la métrica ROUGE-1 como en la ROUGE-L, lo que significa que los resúmenes generados son más precisos y concisos que los de sus competidores.

Esto también se refleja en la robustez de los modelos, donde el modelo GRAPH-ADVANCED y el modelo RAG-TEXT obtienen una puntuación perfecta de 5 sobre 5, mientras que el modelo GRAPH-BASIC obtiene una puntuación de 4 sobre 5. Esto se debe a que el modelo GRAPH-BASIC no proporciona tanta información a la parte de generación como los otros modelos y, por lo tanto, es más probable que falle en las consultas comodín donde algunas variables no se proporcionan claramente.

En términos de rendimiento, el modelo GRAPH-BASIC es el más rápido debido a que tiene la ventana de tokens de contexto más pequeña (tanto en entrada como en salida) de los tres modelos. Y, evidentemente, el modelo GRAPH-ADVANCED es el más lento debido a la complejidad del algoritmo de generación de subgrafos.