AI Agent 과정

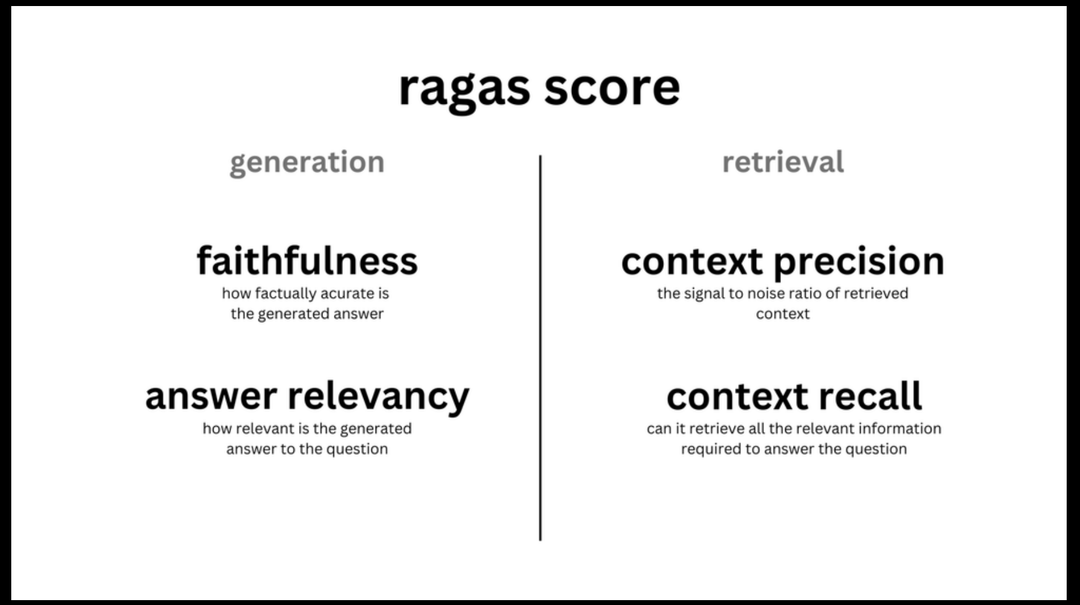
평가 및 모니터링

목차 평가 및 모니터링

2. 평가 자동화 프레임워크

RAGAS

- RAGAS(Retrieval-Augmented Generation Answer Scoring)
  - RAG 시스템이 생성한 답변의 품질을 평가하기 위한 평가 체계
    - 외부 문서나 데이터베이스와 같은 근거 자료를 활용하여 답변을 생성하기 때문에
    - 단순히 문법이나 유창성만 평가하는 기존 지표로는 부족
  - 답변이 근거 자료와 얼마나 일치하며 신뢰할 수 있는지를 평가
- 평가 대상
  - Retrieval(Context)
  - LLM output(Answer)



RAGAS

- Retrieval(Context) 평가
  - **Context Recall**
    - 회수된 Chunk 중 LLM이 생성한 답변과 일치하는 비율
    - 질문에 답하기 위해 필요한 모든 관련 정보를 검색할 수 있는지를 평가
    - Ground truth Answer를 sentence로 쪼갬 → sentence 가 context에 속하는지 판단
    - 정답 답변과 검색된 Context 간의 관련성을 비교하여 계산

$$\text{context recall} = \frac{|\text{GT sentences that can be attributed to context}|}{|\text{Number of sentences in GT}|}$$

https://datainsider.tistory.com/entry/LLMRAGRAGASRAG-evaluationRAGAS

RAGAS

- Retrieval(Context) 평가
  - **Context Recall 예시**
    - Query : 대한민국은 어디에 위치하고 있고, 수도는 어디인가요?
    - Ground truth Answer : 대한민국은 동아시아에 위치하며, 수도는 서울입니다.
    - Context Recall이 높은 경우
      - Retrivered Context : 대한민국은 한반도에 자리한 동아시아의 국가이며, 수도는 서울입니다.
    - Context Recall이 낮은 경우
      - Retrivered Context : 대한민국은 삼면이 바다로 둘러싸인 반도 국가입니다.

https://datainsider.tistory.com/entry/LLMRAGRAGASRAG-evaluationRAGAS

RAGAS

- Retrieval(Context) 평가
  - **Context Precision**
    - Context가 얼마나 잘 회수되었는지를 평가
    - Context 중 Query에 부합하는 중요 내용들이 Top-K에 얼마나 포진되었는지를 평가
    - context → chunk → chunk 마다 유사한지 여부 판단 → 최종 precision 계산

$$\text{Context Precision@K} = \frac{\sum_{k=1}^{K} (\text{Precision@k} \times v_k)}{\text{Total number of relevant items in the top } K \text{ results}}$$

$$\text{Precision@k} = \frac{\text{true positives@k}}{(\text{true positives@k} + \text{false positives@k})}$$

Where K is the total number of chunks in `contexts` and $v_k \in \{0, 1\}$ is the relevance indicator at rank k.

https://datainsider.tistory.com/entry/LLMRAGRAGASRAG-evaluationRAGAS

RAGAS

- Retrieval(Context) 평가
  - **Context Precision 예시**
    - Query : 대한민국은 어디에 위치하고 있고, 수도는 어디인가요?
    - 높은 정밀도 Context :
      - 대한민국은 동아시아에 위치하며, 수도는 서울입니다.
      - 서울은 대한민국의 수도입니다.
    - 낮은 정밀도 Context :
      - 한국의 고유한 전통 발표식품에는 김치가 있습니다.
      - 서울에는 한강이 가로지르고 있습니다.
    - 점수 계산 : 높은 정밀도 Context가 상위에 위치 → 높은 점수

https://datainsider.tistory.com/entry/LLMRAGRAGASRAG-evaluationRAGAS

RAGAS

- LLM output(Answer) 평가
  - **Relevancy(Supportiveness)**
    - 생성된 답변이 제공된 근거 자료의 내용을 얼마나 잘 반영하는지를 평가
    - 답변이 주어진 질문에 얼마나 적합한지를 평가
    - 답변으로 부터 여러 인공 질문 생성 → 생성된 질문들과 원래 질문 간의 평균 코사인 유사도 계산

$$\text{answer relevancy} = \frac{1}{N}\sum_{i=1}^{N} cos(E_{g_i}, E_o)$$

$$\text{answer relevancy} = \frac{1}{N}\sum_{i=1}^{N} \frac{E_{g_i} \cdot E_o}{\|E_{g_i}\|\|E_o\|}$$

Where:

- E_{g_i} is the embedding of the generated question i.
- E_o is the embedding of the original question.
- N is the number of generated questions, which is 3 default.

https://datainsider.tistory.com/entry/LLMRAGRAGASRAG-evaluationRAGAS

RAGAS

- LLM output(Answer) 평가
  - **Relevancy(Supportiveness) 예시**
    - Query : 대한민국은 어디에 위치하고 있고, 수도는 어디인가요?
    - Answer : 대한민국은 동아시아에 있습니다.
    - Generated Query :
      - 대한민국은 아시아의 어느 지역에 있나요?
      - 대한민국은 동아시아에 위치해 있나요?
      - 대한민국은 아시아의 동쪽에 위치하나요?
    - 점수 계산 : 각 인공 질문과 원래 질문 간의 코사인 유사도 계산 후 평균

https://datainsider.tistory.com/entry/LLMRAGRAGASRAG-evaluationRAGAS

RAGAS

- LLM output(Answer) 평가
  - **Faithfulness**
    - 답변에 포함된 정보가 실제 자료에 기반하고 있는지를 확인 (Hallucination 평가)
    - 근거 자료의 정보를 왜곡하거나 잘못 해석하지 않고, 사실 그대로 전달하는지를 측정

$$\text{Faithfulness score} = \frac{|\text{Number of claims in the generated answer that can be inferred from given context}|}{|\text{Total number of claims in the generated answer}|}$$

    - 생성된 답변으로 부터 Claim 추출 후 Claim들이 context로 추론이 가능한지를 평가
      - Claim : 생성된 답변에서 주장하는 핵심 point

https://datainsider.tistory.com/entry/LLMRAGRAGASRAG-evaluationRAGAS

RAGAS

- LLM output(Answer) 평가
  - **Faithfulness 예시**
    - Query : 대한민국은 어디에 있고, 수도는 어디이고, 피파 랭킹은 몇 위야?
    - Context : 대한민국은 동아시아 국가 중 그 다음으로 … , 대한민국은 수도인 서울을 중심으로 …
    - Hight Faithful Answer : 대한민국은 동아시아에 위치하고 있으며 수도는 서울입니다.
    - Low Faithful Answer : 대한민국은 동아시아에 위치하고 있으며 수도는 서울이고 피파 랭킹은 23위입니다.

      → 피파 랭킹은 Context에 없는 내용

https://datainsider.tistory.com/entry/LLMRAGRAGASRAG-evaluationRAGAS