테스트셋 v0.2 ⬝ 테스트셋 v0.2

TREX-TS

Exported on 2024-09-13 13:47:39

Table of Contents

1 수행 일자 3

2 사이트 링크 4

3 테스트셋 데이터 소스 5

4 테스트셋 생성 Tool 6

5 생성된 테스트셋 원본 7

6 테스트셋 필터링 후 내용 8

7 테스트셋 평가 Tool 9

8 테스트셋 평가 RAG 파이프라인 10

9 테스트셋 v0.2 평가 결과 12

10 테스트셋 v0.2 평가 결과 13

11 인사이트와 향후 진행 내용 14

# 수행 일자

2024. 9. 13

# 사이트 링크

<https://d1pkmbiqrxemem.cloudfront.net/>

# 테스트셋 데이터 소스

(수정사항: Format을 Markdown syntax로 변경)

|  |  |
| --- | --- |
| 데이터 소스(클릭하면 열립니다) | 설명 |
| [challenger.txt](https://confluence.samsungds.com/download/attachments/125927675/challenger.txt?api=v2&modificationDate=1723010424000&version=2) | AI challenger 이벤트 관련 내용 |
| [chatbot.txt](https://confluence.samsungds.com/download/attachments/125927675/chatbot.txt?api=v2&modificationDate=1723010434000&version=2) | Samsung Chatbot 일반 설명 (향후 수정 확충 예정) |
| [webpage.txt](https://confluence.samsungds.com/download/attachments/125927675/webpage.txt?api=v2&modificationDate=1723010453000&version=2) | AI Studio 웹페이지 관련 일반 설명 (향후 수정 확충 예정) |
| [exynos.txt](https://confluence.samsungds.com/download/attachments/125927675/exynos.txt?api=v2&modificationDate=1723010466000&version=2) | Exynos 2200 등 Exynos 칩셋 관련 일반 설명 |

# 테스트셋 생성 Tool

데이터셋 생성 프레임워크: **RAGAS**

LLM: OpenAI **chatgpt-4o-mini**

# 생성된 테스트셋 원본

(v0.1와 동일)

# 테스트셋 필터링 후 내용

(v0.1와 동일)

# 테스트셋 평가 Tool

데이터셋 평가 프레임워크 **RAGAS**

LLM: OpenAI **chatgpt-4o-mini**

# 테스트셋 평가 RAG 파이프라인

|  |
| --- |
| **DocumentDB** |
| **AWS with DocumentDB, Lambda** |

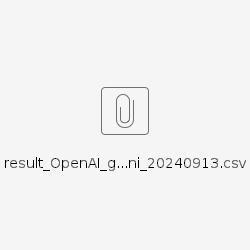
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Phase | items | description | comments |  |  |
| Document 준비 | Chunking | S3에서 데이터 파일을 읽어와 Markdown Chunker로 Chunking | Markdown Chunker\*로 변경 후 총 Context Chunk갯수 10개  Markdown header + text 단위로 나눈 다음에 token 길이가 400이 안되면 합친다. 청킹한다. 람다 함수. |  |  |
| Embedding | 오픈AI  **text-*embedding*-3-*small*** 모델 사용 |  |  |  |
| Vector Storage | AWS DocumentDB에 저장 | vector, text, meta data (page section link, section title) |  |  |
| Q&A 수행 | Context Retrieval | DpcumentDB에서 ***테스트셋 질문***과 코사인 거리가 가까운 top chunk 추출 | DocumentDB에서 max document 50개로 셋팅해서 추출을 하는데 가져온다. HNSW 방식에 따라 관련 있는 일부 chunk만 retrieve 하는 방식. 3~5개 정도 가져온다. |  |  |
| Candidate Selection | 기준에 맞도록 Context Chunk 선정 | chunk size 400 token |  |  |
| LLM 집어 넣기 전 ㅓ리 |  | chunk size 400 token |  |  |
| Answer Generation | ***테스트셋 질문***과 선정된 Chunk를 LLM (오픈AI ***chatgpt-4o-mini***)에 넣고 응답 수집 |  |  |  |
| 평가 | RAGAS evaluation | ***테스트셋(질문, Ground Truth)***의 내용과 LLM이 응답한 ***컨텍스트, 답변*** 내용 총 4가지 필드를 모두 고려하여 RAGAS와 LLM (오픈AI ***chatgpt-4o-mini***)을 통해 지표 평가 수행 | 다음 3가지(상세 설명 링크: [RAGAS 평가 지표](https://confluence.samsungds.com/pages/viewpage.action?pageId=110560495))를 계산합니다:  context\_relevancy, faithfulness, answer\_correctness |  |  |

# 테스트셋 v0.2 평가 결과

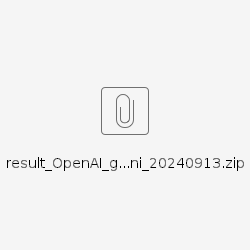
아래의 두 파일 모두 동일한 결과이며 포맷만 csv와 json으로 구분 된 결과 파일 입니다.

# 테스트셋 v0.2 평가 결과

원본파일(csv):



[원본파일(Context 포함):](https://confluence.samsungds.com/download/attachments/125927675/result_OpenAI_gpt-4o-mini_20240807.zip?api=v2&modificationDate=1723011853000&version=1)



|  |  |  |  |
| --- | --- | --- | --- |
| **context\_relevancy** | **faithfulness** | **answer\_correctness** | **설명** |
| **0.019364** | **0.76662** | **0.832666** | 다음 3가지(상세 설명 링크: [RAGAS 평가 지표](https://confluence.samsungds.com/pages/viewpage.action?pageId=110560495))를 계산합니다.   * context\_relevancy(질문과 컨텍스트 간의 관계) * faithfulness(컨텍스트로부터 답변이 잘 추론되었나) * answer\_correctness |

# 인사이트와 향후 진행 내용