테스트셋 v0.1

TREX-TS

Exported on 2024-09-13 09:48:53

Table of Contents

1 수행 일자 3

2 테스트셋 데이터 소스 4

3 테스트셋 생성 Tool 5

4 생성된 테스트셋 원본 6

5 테스트셋 필터링 후 내용 7

6 테스트셋 평가 Tool 10

7 테스트셋 평가 RAG 파이프라인 11

8 테스트셋 v0.1 평가 결과 12

9 인사이트와 향후 진행 내용 13

10 참고: AI Studio 1.0용의 테스트셋 14

10.1 ChatGPT-3.5-Turbo 14

10.2 ChatGPT-4o-mini 16

# 수행 일자

2024. 7. 29

# 테스트셋 데이터 소스

부연 설명: 특정 AI Studio 문서 버전과는 상관이 먼 일반 부분만 먼저 정리한 내용입니다(양이 많지 않습니다). 매뉴얼 부분 관련 AI Studio 1.0 문서는 쓰지 않기로 했고 AI Studio 2.0은 8월말에 나오므로 그때까지 기다려야 합니다.

|  |  |
| --- | --- |
| 데이터 소스(클릭하면 열립니다) | 설명 |
| [challenger.txt](https://confluence.samsungds.com/download/attachments/117278199/challenger.txt?api=v2&modificationDate=1722485246000&version=1) | AI challenger 이벤트 관련 내용 |
| [chatbot.txt](https://confluence.samsungds.com/download/attachments/117278199/chatbot.txt?api=v2&modificationDate=1722485263000&version=1) | Samsung Chatbot 일반 설명 (향후 수정 확충 예정) |
| [webpage.txt](https://confluence.samsungds.com/download/attachments/117278199/webpage.txt?api=v2&modificationDate=1722485289000&version=1) | AI Studio 웹페이지 관련 일반 설명 (향후 수정 확충 예정) |
| [exynos.txt](https://confluence.samsungds.com/download/attachments/117278199/exynos.txt?api=v2&modificationDate=1722485279000&version=1) | Exynos 2200 등 Exynos 칩셋 관련 일반 설명 |

# 테스트셋 생성 Tool

데이터셋 생성 프레임워크: **RAGAS**

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

# 생성된 테스트셋 원본

json 파일: [testset\_general.json](https://confluence.samsungds.com/download/attachments/117278199/testset_general.json?api=v2&modificationDate=1722485178000&version=1)

# 테스트셋 필터링 후 내용

엑셀 파일: <https://jira.samsungds.com/secure/attachment/52228/testset_general_0.1.csv>

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **question** | **ground\_truth** | **evolution\_type** | **metadata** | **TS Comment** |
| What capabilities do satellite connections provide in relation to the Exynos 2400 chipset? | The Exynos 2400 chipset supports non-terrestrial network (NTN) satellite connections, which enhance connectivity even in cellular dead zones. | simple | exynos | 문제 없음 |
| What is the purpose of the Contact Us page? | The purpose of the Contact Us page is to obtain personalized assistance by submitting queries. | simple | webpage | 문제 없음 |
| What is the future goal of the EDS Chatbot regarding the Samsung S.LSI Exynos platform? | The future goal of the EDS Chatbot is to provide information about technical issues, product details, events, and announcements pertinent to the Samsung S.LSI Exynos platform. | simple | chatbot | 문제 없음 |
| What are the video performance capabilities of the Exynos 2400 chipset? | The Exynos 2400 chipset's video performance capabilities include up to 8K decoding at 60fps and encoding at 30fps for 10-bit HEVC (H.265) and VP9. | simple | exynos | 문제 없음 |
| What display capability does the Exynos 2200 chipset support at 144Hz? | The Exynos 2200 chipset supports QHD+ resolution at 144Hz. | simple | exynos | 문제 없음 |
| What advancements does the Exynos 2400 bring to mobile gaming? | The Exynos 2400 brings advancements to mobile gaming through console-quality graphics and advanced GPU technology. | simple | exynos | 문제 없음 |
| What are the key features of the Exynos 2200 chipset? | The key features of the Exynos 2200 chipset include being built using a 4-nanometer EUV process, incorporating Arm's latest Armv9 CPU cores, and including a fast 5G modem compliant with 3GPP Release 16. | simple | exynos | 문제 없음 |
| ~~What is the significance of industry-academia collaboration in the advancement of Edge AI technology?~~ | ~~The answer to given question is not present in context~~ | simple | challenger | 삭제 |
| ~~What role does Samsung S. LSI play in the management of the Exynos Developer Society?~~ | ~~Samsung S. LSI manages the Exynos Developer Society.~~ | simple | webpage | 추가 정보 필요 |
| What video performance capabilities does the Exynos 2400 chipset offer for 10-bit HEVC? | The Exynos 2400 chipset offers video performance capabilities of up to 8K decoding at 60fps and encoding at 30fps for 10-bit HEVC (H.265). | simple | exynos | 문제 없음 |
| ~~What 5G features come from the Exynos 2200's CPU?~~ | ~~The answer to given question is not present in context~~ | multi\_context | exynos | 삭제 |
| ~~Which initiative enhances AI skills for Korean students and addresses industry issues?~~ | ~~The AI Challenger enhances AI skills for Korean students and addresses industry issues by targeting students and postgraduates to improve their AI expertise and competitiveness in the industry.~~ | multi\_context | challenger | 추가 정보 필요 |
| ~~What 8K 60fps video features set the Exynos 2400 apart from the 2200?~~ | ~~The Exynos 2400 chipset supports 8K decoding at 60fps and encoding at 30fps for 10-bit HEVC(H.265) and VP9, while the Exynos 2200 also supports 8K decoding at 60fps but has 8K encoding at 30fps. The distinction lies in the encoding capabilities, as both chipsets support the same decoding frame rate.~~ | multi\_context | exynos | 틀린 대답 |
| What CPU cores set the Exynos 2400 apart from the 2200, and which GPU does it have? | The Exynos 2400 chipset is set apart from the Exynos 2200 by its CPU configuration, which includes Cortex짰-X4, Cortex짰-A720, and Cortex짰-A520. It is paired with the Samsung Xclipse 940 GPU. | multi\_context | exynos | 문제 없음 |
| How do the display resolutions and refresh rates of the Exynos 2200 compare to the Exynos 2400? | The display capabilities of the Exynos 2200 and Exynos 2400 chipsets are identical, as both support 4K/WQUXGA resolutions at 120Hz and QHD+ at 144Hz. | multi\_context | exynos | 문제 없음 |
| ~~What advanced features does the Exynos 2200's 5G modem get from its CPU/GPU?~~ | ~~The answer to given question is not present in context~~ | multi\_context | exynos | 삭제 |
| ~~What criteria do Samsung S. LSI admins use for top projects on the Exynos Best Lab?~~ | ~~The answer to given question is not present in context~~ | multi\_context | webpage | 추가 정보 필요 |
| ~~What sets the Exynos 2400 apart from the 2200 for 8K at 60fps?~~ | ~~The Exynos 2400 chipset supports 8K decoding at 60fps and encoding at 30fps, while the Exynos 2200 also supports 8K decoding at 60fps but has 8K encoding at 30fps. The key difference is in the encoding capabilities, where both support 10-bit HEVC(H.265) and VP9.~~ | multi\_context | exynos | 틀린 대답 |
| ~~How does the 14.7x AI boost of Exynos 2400 improve efficiency & security?~~ | ~~The 14.7x improvement in AI performance of the Exynos 2400 chipset enhances efficiency by reducing latency and boosting device security.~~ | reasoning | exynos | 질문 수정필요 |
| ~~What are the Exynos 2200's 5G modem specs in relation to its CPU/GPU?~~ | ~~The answer to given question is not present in context~~ | reasoning | exynos | 삭제 |

# 테스트셋 평가 Tool

데이터셋 평가 프레임워크 **RAGAS**

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

# 테스트셋 평가 RAG 파이프라인

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Phase | items | description | comments |  |  |
| Document 준비 | Chunking | 수동 임시 newline chunking | 빠른 작업을 위한 임시 방편입니다. LangChain 등 기존의 프레임워크에서 사용하는 Chunking 라이브러리로 교체 예정입니다. |  |  |
| Embedding | 오픈AI  **text-*embedding*-3-*small*** 모델 사용 |  |  |  |
| Vector Storage | CSV file에 저장 | 차주 AWS DocumentDB로 이관 예정입니다. |  |  |
| Q&A 수행 | Context Retrieval | ***테스트셋 질문***과 코사인 거리가 가까운 top 10 chunk 추출 | 차주 AWS DocumentDB로 향후 이관 예정입니다. |  |  |
| Candidate Selection | Top-10 chunk 선정 |  |  |  |
| Answer Generation | ***테스트셋 질문***과 Top-10 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 v0.1 평가 결과

엑셀 파일: <https://jira.samsungds.com/secure/attachment/52398/ragas_general_0.1.csv>

|  |  |  |  |
| --- | --- | --- | --- |
| **context\_relevancy** | **faithfulness** | **answer\_correctness** | **설명** |
| **0.22** | **0.90** | **0.85** | 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): 수치가 좋지 않은데 top-10개를 추출할 때 관련성이 먼 chunk들이 추출된다는 뜻입니다. 이 부분은 AWS DocumentDB 셋업 및 AI Studio 2.0 매뉴얼 준비 후 더 튜닝 예정입니다. * faithfulness(컨텍스트로부터 답변이 잘 추론되었나): 추출된 chunk 양이 많은 편이라 답변과의 관련성이 높습니다. * answer\_correctness: 현재는 데이터 소스 크기가 (일반 내용으로만 구성) 크지 않아 점수가 높게 나오는 편입니다. 향후 AI Studio 2.0 매뉴얼이 준비되는 8월 이후에 돌려야 제대로 된 수치가 기대됩니다. |

# 인사이트와 향후 진행 내용

선행 연구와 워크샵 등을 통해 신규로  데이터셋을 작성한 초기 결과입니다 (기존 FAQ 내용은 취지가 달라 평가 데이터셋보다는 RAG 데이터 소스로 사용하기로 했습니다)

RAGAS 프레임워크를 이용한 데이터셋 생성 방법론과 데이터셋 평가 방법론을 모두 적용해 본 결과입니다.

이제 RAGops 자동화를 위한 초석을 다진 것이고 이후는 AWS 인프라의 셋업과 AI Studio 2.0 매뉴얼 (8월말)을 거치며 지속적으로 RAG 파이프라인을 고도화/튜닝하면 될 것 같습니다.

다음주에는 AWS DocumentDB가 새로이 셋업되니 파이프라인이 변경됩니다. 그에 대해 다시 테스트 및 평가 진행하겠습니다.

문서 반영/평가 로드맵은 다음주까지 작성 예정입니다. 썬더소프트 내부에서 시스템 운영적인 측면도 같이 고려해야 해서 시간이 걸리고 있습니다.

# 참고: AI Studio 1.0용의 테스트셋

부연설명: 사용하지 않을 것이므로 현재 큰 의미는 없습니다

AI Studio V1

## ChatGPT-3.5-Turbo

|  |  |  |
| --- | --- | --- |
| question | ground\_truth | evolution\_type |
| What are the hardware preferences available for model conversion in the ENN SDK service? | The hardware preferences available for model conversion in the ENN SDK service are Default and Accelerate. | simple |
| Where can I find the list of samples for ENN SDK? | The list of samples for ENN SDK is available in [ENN SDK Samples](enn-sdk-samples). | simple |
| How can you report a bug or issue with the ENN SDK? | To report a bug or issue with the ENN SDK, you need to follow the instructions provided in the 'Reporting ENN SDK Issues' section. | simple |
| What is the purpose of merging successive concatenate layers that share the same Output Feature Maps (OFM)? | The purpose of merging successive concatenate layers that share the same Output Feature Maps (OFM) is to combine the information from multiple layers and enhance the representation of the features. | simple |
| How can users provide feedback on the ENN SDK? | Users can provide feedback on the ENN SDK by sharing general questions, feedback, or suspected bugs on the forums for public discussion. Alternatively, they can submit their concerns through the Contact Us page for a more direct approach or personalized assistance. | simple |
| How can I receive personalized support for ENN SDK inquiries from the Exynos Developer Society? | If you prefer a more direct approach or need personalized assistance, submit your concerns to our Contact Us page. | multi\_context |
| What are the steps to convert TFLite models into NNC models and integrate them into an application using the ENN SDK service? | To convert TFLite models into NNC models and integrate them into an application using the ENN SDK service, follow these steps:    1. Access the Exynos Developer Society by signing up to create an account if you are a new user or logging in if you are an existing user.  2. Navigate to the ENN SDK service page.  3. Provide project information by entering a descriptive title for your project and uploading your TFLite model.  4. Choose hardware preferences, either Default (CPU and GPU) or Accelerate (CPU, GPU, and NPU).  5. Initiate the conversion by confirming your selections and clicking Convert.  6. If the conversion is successful, the NNC Download button will be enabled.  7. Click NNC Download to download the NNC model.  8. Integrate the downloaded NNC model into your application as required. | multi\_context |
| Why is it recommended to merge successive concatenate layers with the same IFM and OFM, excluding the dropout layer, to improve performance? | Merging successive concatenate layers with the same IFM and OFM, excluding the dropout layer, is recommended to improve performance because it reduces the number of parameters and computations in the network. By merging these layers, the network becomes more efficient and less prone to overfitting. Additionally, it helps to preserve important features and gradients throughout the network, leading to better overall performance. | multi\_context |
| Where can I find the ENN SDK samples and information on using the ENN SDK service? | The ENN SDK samples can be found in the ENN SDK Samples section. Detailed information on using the ENN SDK service can be found in the ENN SDK service section. | multi\_context |
| What tool can users use to optimize TensorFlow Lite neural network models for Samsung Exynos hardware, and why? | Users can use the ENN SDK to optimize TensorFlow Lite neural network models for Samsung Exynos hardware. The ENN SDK allows users to convert trained TensorFlow Lite models to a format that can run efficiently on Samsung Exynos hardware, ensuring optimal performance and resource utilization. | reasoning |

(원본)



## ChatGPT-4o-mini

|  |  |  |
| --- | --- | --- |
| question | ground\_truth | evolution\_type |
| What are the steps to execute the program using ADB? | The steps to execute the program using ADB are described in the 'Using ADB to Execute Native Program' guide. | simple |
| What are the steps to report an issue with the ENN SDK? | The answer to given question is not present in context | simple |
| What is the procedure for connecting the ERD board to the computer when utilizing the sample application? | The procedure for connecting the ERD board to the computer when utilizing the sample application is to connect the ERD board to the computer. | simple |
| What is the significance of NNC models in the context of using the ENN SDK? | NNC models are significant in the context of using the ENN SDK because they are a format that can run efficiently in Samsung Exynos hardware, allowing for effective execution of neural network models. | simple |
| What functionalities does the Android application provide in relation to image classification? | The Android application provides functionalities for executing the NNC model using the ENN framework and running the corresponding TFLite model. It also displays the results and inference time of both the ENN framework and TFLite at the bottom of the application interface. | simple |
| What are the steps to connect the ERD board to a PC for image data inference? | To connect the ERD board to a computer for image data inference, you need to follow these steps: 1. Connect the ERD board to the computer. 2. Run the application (using Shift + F10). 3. Provide the image data for inference. | multi\_context |
| What's the process to run an image classification app on the ERD board with ADB? | The process to run an image classification app on the ERD board with ADB involves the following steps: First, download the samples by cloning the Github repository using the command 'git clone <https://github.com/exynos-eco/enn-sdk-samples-9925.git>'. Then, open the downloaded 'image-classification' project in Android Studio. After connecting the ERD board, click 'run 'app'' to launch the application on the ERD board after the build. | multi\_context |
| What are the constraints for TensorFlow Lite models in ENN SDK, like dims, size, and quantization for NPU? | The constraints for TensorFlow Lite models in ENN SDK include: tensors with up to four dimensions, a maximum model size of 1 GB, and for NPU, the models must be quantized. | multi\_context |
| Which two dirs in enn-sdk-samples-9925 GitHub repo have the sample? | The two directories in the enn-sdk-samples-9925 GitHub repository that have the sample are 'perf-compare' and 'image-enhance'. | multi\_context |
| How does input binary data affect NNC execution and inference time? | The answer to given question is not present in context | reasoning |

(원본)

