241018 평가 결과

TREX-TS

Exported on 2024-10-31 12:26:46

Table of Contents

1 수행 일자 3

2 데이터셋 4

3 테스트셋 5

3.1 생성에 사용한 Tool 5

3.2 테스트셋 세부 내용 5

4 테스트셋 평가 19

4.1 사용한 Tool 19

4.2 파이프라인 19

4.2.1 아키텍처 19

4.2.2 세부 설명 19

5 Query 수행시 사용하는 시스템 프롬프트 전체 텍스트 21

6 241018 평가 결과 요약 22

7 인사이트 & 다음 수정 내용( 2024. 10. 29 기준) 23

# 수행 일자

2024. 10. 18

# 데이터셋

1. 웹페이지 공통: [자료 링크](https://confluence.samsungds.com/pages/viewpage.action?pageId=117278199)
2. FAQ 리스트: [자료 링크](https://jira.samsungds.com/browse/TREXTS-204)
3. AI Studio 2.0 9/24 버전: [자료 링크](https://confluence.samsungds.com/pages/viewpage.action?pageId=134841292)
4. Remote Farm 문서: [자료 링크](https://confluence.samsungds.com/pages/viewpage.action?pageId=134316824)

# 테스트셋

## 생성에 사용한 Tool

데이터셋 생성 프레임워크: **RAGAS (ver 0.1.10)**

LLM: OpenAI **gpt-4o-mini**,  **gpt-4o(신규 추가 항목)**

## 테스트셋 세부 내용

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 데이터셋번호 | question | ground\_truth | evolution\_type | metadata |
| 3 | What are the different precision options available for weights and activations in the quantizer? | The different precision options available for weights and activations in the quantizer are specified by 'precision\_weight' and 'precision\_activation', which can be values like int8, int16, or fp16. | simple | Header 1 : \*\*How to use\*\* | Header 3 : \*\*Detailed explanation for eht yaml file\*\* | domain : exynos-ai-studio-docs-main/how\_to\_use/[how\_to\_use.md](http://how_to_use.md) |
| 3 | What is the purpose of applying Fixed Precision Quantization to the optimized CNNX model? | The purpose of applying Fixed Precision Quantization to the optimized CNNX model is to perform quantization, which typically involves reducing the precision of the model's weights and activations to improve efficiency, such as reducing model size and increasing inference speed. | simple | Header 1 : Model Optimization Flow | Header 2 : CV | domain : exynos-ai-studio-docs-main/eht/model\_optimization\_flow/[model\_optimization\_flow.md](http://model_optimization_flow.md) |
| 3 | Why is it important for the keys in the h5 dataset to match the input names of the model? | It is important for the keys in the h5 dataset to match the input names of the model to ensure proper mapping occurs. | simple | Header 1 : \*\*Dataset preparation\*\* | Header 3 : \*\*Dataset format\*\* | domain : exynos-ai-studio-docs-main/dataset\_and\_model/[dataset\_and\_model.md](http://dataset_and_model.md) |
| 3 | What are the different precision options available for weights and activations in the quantizer? | The different precision options available for weights and activations in the quantizer are int8, int16, and fp16. | simple | Header 1 : \*\*How to use\*\* | Header 3 : \*\*Detailed explanation for eht yaml file\*\* | domain : exynos-ai-studio-docs-main/how\_to\_use/[how\_to\_use.md](http://how_to_use.md) |
| 3 | What steps are involved in the optimization process of a CNNX model? | The optimization process of a CNNX model involves passing the CNNX model through a Simplifier and 4-Dimensional Conversion, followed by applying an Optimization template. | simple | Header 1 : Model Optimization Flow | Header 2 : LVM | domain : exynos-ai-studio-docs-main/eht/model\_optimization\_flow/[model\_optimization\_flow.md](http://model_optimization_flow.md) |
| 3 | What is the purpose of the debug API in mixed precision quantization? | The purpose of the debug API in mixed precision quantization is to allow users to select different activations and weights and assign them varying levels of precision. | simple | Header 1 : Quantizer | Header 2 : Debug API | Header 3 : Layer-wise mixed precision quantiztion debug API | domain : exynos-ai-studio-docs-main/eht/quantizer/[quantizer.md](http://quantizer.md) |
| 3 | What ONNX opset versions does EHT currently support? | EHT currently supports ONNX opset versions 13 to 17. | simple | Header 1 : Model Requirements and Constraints | Header 3 : opset version | domain : exynos-ai-studio-docs-main/dataset\_and\_model/[dataset\_and\_model.md](http://dataset_and_model.md) |
| 3 | What is the process of static uniform quantization in relation to weights and activations? | Static uniform quantization involves applying a uniform quantization process to both weights and activations, where users can specify the precision (bit-width) for these components. The entire model is quantized to the specified precision. | simple | Header 1 : Quantizer | Header 2 : Basic Quantization Methods | Header 3 : Fixed Precision Quantization | domain : exynos-ai-studio-docs-main/eht/quantizer/[quantizer.md](http://quantizer.md) |
| 3 | How does setting the Add operator to INT4 affect the outputs in a model using mixed precision? | Setting the Add operator to INT4 quantizes all outputs of the Add operators to INT4 in a model using mixed precision. | simple | Header 1 : Quantizer | Header 2 : Basic Quantization Methods | Header 3 : Mixed Precision Quantization | domain : exynos-ai-studio-docs-main/eht/quantizer/[quantizer.md](http://quantizer.md) |
| 3 | How can two CNNX models be compared using their inference outputs and intermediate tensors? | Two CNNX models can be compared using their inference outputs by using the `compare\_model\_by\_inference` feature, which uses the SNR value as the comparison metric. They can also be compared using their intermediate tensors for each layer by using the `compare\_model\_by\_layer` feature. | simple | Header 1 : Simulator | domain : exynos-ai-studio-docs-main/eht/simulator/[simulator.md](http://simulator.md) |
| 3 | How can users apply mixed precision quantization (MPQ) to models using the Quantizer module in EHT? | Users can apply mixed precision quantization to models by specifying activation names or operators. | simple | Header 1 : Quantizer | domain : exynos-ai-studio-docs-main/eht/quantizer/[quantizer.md](http://quantizer.md) |
| 3 | What are the system requirements for running Linux based on Ubuntu 22.04 with NVIDIA support? | The system requirements for running Linux based on Ubuntu 22.04 with NVIDIA support include: NVIDIA driver version 450.80.02 or later, Docker 19.03 or later with NVIDIA Container Toolkit support, and NVIDIA Container Toolkit (nvidia-docker2). | simple | Header 1 : System requirement | Header 2 : Software | domain : exynos-ai-studio-docs-main/system\_requirements/[system\_requirements.md](http://system_requirements.md) |
| 3 | What is the purpose of simulated quantization in the context of CNNX model inference? | Simulated quantization provides the capability to mimic the effects of quantized operations. Floating point values are clipped and divided into several ranges, and values within each range are converted to the same value, simulating the quantization process. | simple | Header 1 : Simulator | domain : exynos-ai-studio-docs-main/eht/simulator/[simulator.md](http://simulator.md) |
| 3 | What functionalities does EHT software offer for neural networks, specifically regarding quantization and model optimization? | EHT software offers functionalities such as quantization and model optimization for neural networks. | simple | Header 1 : Introduction to Exynos AI High-Level Toolchain (EHT) | domain : exynos-ai-studio-docs-main/eht/[eht.md](http://eht.md) |
| 3 | How can different precisions be utilized in a model using mixed precision approaches? | Different precisions can be utilized in a model using mixed precision approaches by specifying precisions for specific activation or weight names, or by defining precisions for different types of operators, such as setting the Add operator to INT4, which quantizes all outputs of the Add operators to INT4. | simple | Header 1 : Quantizer | Header 2 : Basic Quantization Methods | Header 3 : Mixed Precision Quantization | domain : exynos-ai-studio-docs-main/eht/quantizer/[quantizer.md](http://quantizer.md) |
| 3 | What are the system requirements for running Linux based on Ubuntu 22.04 with NVIDIA support? | The system requirements for running Linux based on Ubuntu 22.04 with NVIDIA support include: NVIDIA driver version 450.80.02 or later, Docker 19.03 or later (with NVIDIA Container Toolkit support), and NVIDIA Container Toolkit (nvidia-docker2). | simple | Header 1 : System requirement | Header 2 : Software | domain : exynos-ai-studio-docs-main/system\_requirements/[system\_requirements.md](http://system_requirements.md) |
| 3 | What is the purpose of applying an Optimization template to the CNNX model? | The purpose of applying an Optimization template to the CNNX model is to optimize the model after it has been passed through Simplifier and 4-Dimensional Conversion. | simple | Header 1 : Model Optimization Flow | Header 2 : LLM | domain : exynos-ai-studio-docs-main/eht/model\_optimization\_flow/[model\_optimization\_flow.md](http://model_optimization_flow.md) |
| 3 | How does mixed precision by name allow users to specify precisions for specific activation or weight names? | Mixed precision by name allows users to specify precisions for specific activation or weight names. | simple | Header 1 : Quantizer | Header 2 : Basic Quantization Methods | Header 3 : Mixed Precision Quantization | domain : exynos-ai-studio-docs-main/eht/quantizer/[quantizer.md](http://quantizer.md) |
| 3 | How are optimization scenarios currently predetermined for large language models? | Optimization scenarios are predetermined for large language models as part of the current optimization flow according to model type. | simple | Header 1 : Model Optimization Flow | domain : exynos-ai-studio-docs-main/eht/model\_optimization\_flow/[model\_optimization\_flow.md](http://model_optimization_flow.md) |
| 3 | What are the recommended specifications for a processor when considering an Intel Core i7? | The recommended specifications for a processor when considering an Intel Core i7 are that it should be the latest model with 4 cores or more and have a 64-bit (x86-64) architecture. | simple | Header 1 : System requirement | Header 2 : Hardware | Header 3 : CPU | domain : exynos-ai-studio-docs-main/system\_requirements/[system\_requirements.md](http://system_requirements.md) |
| 3 | How can users apply their own optimization methods to their models using the Optimizer Template? | Users can apply their own optimization methods to their models using the Optimizer Template by following these steps: create custom templates, prepare the model to be optimized, and validate the optimized model. | simple | Header 1 : Optimizer | Header 2 : How to Create Custom Templates | domain : exynos-ai-studio-docs-main/eht/optimizer/[optimizer.md](http://optimizer.md) |
| 3 | What functionalities does Exynos AI Studio provide for neural network models? | Exynos AI Studio provides functionalities such as quantization, conversion, optimization, and compilation to generate NNC models for neural network models. | simple | Header 1 : Introduction to Exynos AI Studio | domain : exynos-ai-studio-docs-main/introduction/[introduction.md](http://introduction.md) |
| 3 | How do users use the EHT module for MPQ with activation IDs and ops to set precision for model parts? | Users can apply mixed precision quantization to models by specifying activation names or operators using the EHT module. | multi\_context | Header 1 : Quantizer | Header 2 : Debug API | Header 3 : Layer-wise mixed precision quantiztion debug API | domain : exynos-ai-studio-docs-main/eht/quantizer/[quantizer.md](http://quantizer.md) |
| 3 | How does the MPQ API aid precision tweaks in a quantized CNNX model with EHT's quantizer? | The MPQ API aids precision tweaks in a quantized CNNX model by allowing users to select different activations and weights and assign them varying levels of precision. | multi\_context | Header 1 : Quantizer | Header 2 : Debug API | Header 3 : Layer-wise mixed precision quantiztion debug API | domain : exynos-ai-studio-docs-main/eht/quantizer/[quantizer.md](http://quantizer.md) |
| 4 | How does real-time synchronization work in the File Directory feature of the Remote Streaming Service? | The file directory is synchronized in real-time with the file browser application of the device and the file updates are immediately reflected. | simple | Header 2 : [File Directory Management] | domain : devicefarm/[Getting\_Started\_20240923.md](http://Getting_Started_20240923.md) |
| 4 | How can a user navigate and download files using the File Directory feature in the Remote Streaming Service? | A user can navigate and download files using the File Directory feature in the Remote Streaming Service by double-clicking any folder to access its location, using the Back button to navigate to the parent folder, and the forward button to navigate to the previous folder. To download files, the user can double-click the uploaded files in the /sdcard/download/ directory or the captured files in the /sdcard/image/ directory. | simple | Header 2 : [File Directory Management] | domain : devicefarm/[Getting\_Started\_20240923.md](http://Getting_Started_20240923.md) |
| 4 | How does Samsung's advanced real-device testing service enhance the quality of models and apps? | Samsung's advanced real-device testing service enhances the quality of models and apps by connecting them across diverse devices, leading to a smarter and more innovative outcome. | simple | domain : devicefarm/[Getting\_Started\_20240923.md](http://Getting_Started_20240923.md) |
| 4 | How can command-line inputs be used to control a device through the Command Shell feature in the Remote Streaming Service? | Command-line inputs can be used to control a device through the Command Shell feature by entering the necessary commands in a modal window and either pressing the Enter key or clicking Send. | simple | Header 2 : [Command Shell Access] | domain : devicefarm/[Getting\_Started\_20240923.md](http://Getting_Started_20240923.md) |
| 4 | How can users select a device on the Device Farm page? | Users can select a device on the Device Farm page by clicking the Device drop-down menu and selecting the required device. The Device area is then refreshed to display the information of the selected device. | simple | Header 2 : [Device Selection] | domain : devicefarm/[Getting\_Started\_20240923.md](http://Getting_Started_20240923.md) |
| 4 | How can users manage and verify the validity and usage of their credits on the Device Farm page? | Users can manage and verify the expiration details, acquisition, and usage of their credits from the My Credit menu of the My Page. | simple | Header 2 : [Credit Management] | domain : devicefarm/[Getting\_Started\_20240923.md](http://Getting_Started_20240923.md) |
| 4 | How can you download logs using the Logcat Window feature in the Remote Streaming Service? | To download logs using the Logcat Window feature in the Remote Streaming Service, navigate to the Logcat Window section and click Download to save the system log file to your PC. | simple | Header 2 : [Logcat Window] | domain : devicefarm/[Getting\_Started\_20240923.md](http://Getting_Started_20240923.md) |
| 4 | How can the Logcat Window feature be used to review and save usage logs of the remote service? | To use the Logcat Window feature to review and save usage logs of the remote service, navigate to the Logcat Window section, review system logs and debugging information as required, and click Download to save the system log file to your PC. | simple | Header 2 : [Logcat Window] | domain : devicefarm/[Getting\_Started\_20240923.md](http://Getting_Started_20240923.md) |
| 4 | What components are integrated into the System on Chip (SoC) developed by Samsung Electronics? | The System on Chip (SoC) developed by Samsung Electronics integrates high-performance CPU, GPU, NPU, and memory management features. | simple | Header 2 : [Processor Specification Check] | domain : devicefarm/[Getting\_Started\_20240923.md](http://Getting_Started_20240923.md) |
| 4 | How does the File Directory feature in the Remote Streaming Service enable file management on a remote device? | The File Directory feature in the Remote Streaming Service enables file management on a remote device by allowing users to access and manage files through double-clicking folders to access them, and double-clicking captured or uploaded files to download them to their PC. The file directory is synchronized in real-time with the device's file browser application, ensuring immediate reflection of file updates. | simple | Header 2 : [File Directory Management] | domain : devicefarm/[Getting\_Started\_20240923.md](http://Getting_Started_20240923.md) |
| 4 | How does real-time synchronization work in the File Directory feature of the Remote Streaming Service? | The file directory is synchronized in real-time with the file browser application of the device, and the file updates are immediately reflected. | simple | Header 2 : [File Directory Management] | domain : devicefarm/[Getting\_Started\_20240923.md](http://Getting_Started_20240923.md) |
| 4 | How can the Command Shell feature be used to control a device in the Remote Streaming Service? | The Command Shell feature can be used to control a device in the Remote Streaming Service by entering commands necessary for device control in a modal window. Users can enter commands and either press the Enter key on the keyboard or click Send to execute them. The modal window is accessed by clicking Command Shell. | simple | Header 2 : [Command Shell Access] | domain : devicefarm/[Getting\_Started\_20240923.md](http://Getting_Started_20240923.md) |
| 4 | How can the Logcat Window feature be used to review and save usage logs of the remote service? | To use the Logcat Window feature to review and save usage logs of the remote service, navigate to the Logcat Window section, review system logs and debugging information as required, and click Download to save the system log file to your PC. | simple | Header 2 : [Logcat Window] | domain : devicefarm/[Getting\_Started\_20240923.md](http://Getting_Started_20240923.md) |
| 4 | How can the Reservation Status List feature be used to manage Remote Service reservations on the Device Farm page? | To use the Reservation Status List feature to manage Remote Service reservations on the Device Farm page, you can click Connect for immediate access to services, click Stop to end a service, and click Cancel to cancel a reservation. | simple | Header 2 : [Reservation Status Check] | domain : devicefarm/[Getting\_Started\_20240923.md](http://Getting_Started_20240923.md) |
| 4 | How can the Logcat Window feature be used to review and save usage logs of the remote service? | To use the Logcat Window feature to review and save usage logs of the remote service, navigate to the Logcat Window section, review system logs and debugging information as required, and click Download to save the system log file to your PC. | simple | Header 2 : [Logcat Window] | domain : devicefarm/[Getting\_Started\_20240923.md](http://Getting_Started_20240923.md) |
| 4 | How does the Command Shell feature facilitate device control in the Remote Streaming Service? | The Command Shell feature facilitates device control in the Remote Streaming Service by allowing users to enter command-line inputs necessary for device control. Users can continuously enter commands similar to a PC command window and send them by pressing the Enter key or clicking Send. The feature is accessed by clicking Command Shell to open a modal window where commands can be entered and executed. | simple | Header 2 : [Command Shell Access] | domain : devicefarm/[Getting\_Started\_20240923.md](http://Getting_Started_20240923.md) |
| 4 | What is the purpose of providing detailed information in the guide for Device Farm users? | The purpose of providing detailed information in the guide for Device Farm users is to enhance their understanding of the Device Farm and ensure a smooth experience with the system. | simple | Header 2 : [Overview] | domain : devicefarm/[Getting\_Started\_20240923.md](http://Getting_Started_20240923.md) |
| 4 | How can you download logs using the Logcat Window feature in the Remote Streaming Service? | To download logs using the Logcat Window feature in the Remote Streaming Service, navigate to the Logcat Window section and click Download to save the system log file to your PC. | simple | Header 2 : [Logcat Window] | domain : devicefarm/[Getting\_Started\_20240923.md](http://Getting_Started_20240923.md) |
| 4 | How can you capture the device screen using the Screenshot feature in the Remote Streaming Service? | To capture the device screen using the Screenshot feature in the Remote Streaming Service, click Screenshot to capture the device screen as an image file. The captured screenshots are downloaded to your PC and saved to the /sdcard/image/ directory. | simple | Header 2 : [Screenshot / Recording] | domain : devicefarm/[Getting\_Started\_20240923.md](http://Getting_Started_20240923.md) |
| 4 | How can the Uploading File feature be used to transfer and verify files in the Remote Streaming Service? | To use the Uploading File feature in the Remote Streaming Service, you can transfer files by either dragging and dropping them from the local directory to the upload area or by clicking the upload area to open the file dialog, selecting a file, and clicking OK. After uploading is complete, you can verify the uploaded files by navigating to the /sdcard/download/ directory within the service page. For APK files, they can be uploaded in the same manner and will be automatically installed and launched on the device. | simple | Header 2 : [Uploading File/APP] | domain : devicefarm/[Getting\_Started\_20240923.md](http://Getting_Started_20240923.md) |
| 4 | What features does the Remote Streaming Service offer for capturing and recording the device screen? | The Remote Streaming Service offers the Screenshot feature to capture images of the device screen and the Record feature to record the device screen in real-time. | simple | Header 2 : [Screenshot / Recording] | domain : devicefarm/[Getting\_Started\_20240923.md](http://Getting_Started_20240923.md) |
| 4 | What components are integrated into the System on Chip (SoC) developed by Samsung Electronics? | The System on Chip (SoC) developed by Samsung Electronics integrates high-performance CPU, GPU, NPU, and memory management features. | simple | Header 2 : [Processor Specification Check] | domain : devicefarm/[Getting\_Started\_20240923.md](http://Getting_Started_20240923.md) |
| 4 | How does the Command Shell feature facilitate device control in the Remote Streaming Service? | The Command Shell feature facilitates device control in the Remote Streaming Service by allowing users to enter command-line inputs necessary for device control. Users can continuously enter commands similar to a PC command window and send them by pressing the Enter key or clicking Send. The feature is accessed by clicking Command Shell to open a modal window where commands can be entered and executed. | simple | Header 2 : [Command Shell Access] | domain : devicefarm/[Getting\_Started\_20240923.md](http://Getting_Started_20240923.md) |
| 4 | What features does the Remote Streaming Service offer for capturing and recording the device screen? | The Remote Streaming Service offers the Screenshot feature to capture images of the device screen and the Record feature to record the device screen in real-time. | simple | Header 2 : [Screenshot / Recording] | domain : devicefarm/[Getting\_Started\_20240923.md](http://Getting_Started_20240923.md) |
| 4 | What is the purpose of providing detailed information in the guide for Device Farm users? | The purpose of providing detailed information in the guide for Device Farm users is to enhance their understanding of the Device Farm and ensure a smooth experience with the system. | simple | Header 2 : [Overview] | domain : devicefarm/[Getting\_Started\_20240923.md](http://Getting_Started_20240923.md) |
| 4 | What steps are involved in launching a remote service? | The steps involved in launching a remote service are outlined in section 4.[Remote Service Launch]. | simple | Header 2 : Table of Contents | domain : devicefarm/[Getting\_Started\_20240923.md](http://Getting_Started_20240923.md) |
| 4 | What guarantees instant file updates in the Remote Streaming Service's directory? | The file directory is synchronized in real-time with the file browser application of the device, and the file updates are immediately reflected. | reasoning | Header 2 : [File Directory Management] | domain : devicefarm/[Getting\_Started\_20240923.md](http://Getting_Started_20240923.md) |
| 1 | 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. | | |
| 1 | What is the purpose of the Contact Us page? | The purpose of the Contact Us page is to obtain personalized assistance by submitting queries. | | |
| 1 | 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. | | |
| 1 | What display capability does the Exynos 2200 chipset support at 144Hz? | The Exynos 2200 chipset supports QHD+ resolution at 144Hz. | | |
| 1 | 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. | | |
| 1 | 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. | | |
| 1 | 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). | | |
| 1 | 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. | | |
| 2 | I would like to know which operators are supported by the ENN SDK when developing Computer Vision for Healthcare. | For the Computer Vision for Healthcare category, the supported operators include CONVOLUTION, MAXPOOL, AVGPOOL, RESHAPE, and CONCATENATION. | | |
| 2 | I would like to know which operators are supported by the ENN SDK when developing Augmented and Virtual Reality. | In the Augmented and Virtual Reality category, the supported operators are CONVOLUTION, DEPTHWISE\_CONVOLUTION, and RESHAPE. | | |
| 2 | I would like to know examples of applications that can be created using Image Classification. | AI Applications that can be created using Image Classification include Face recognition, Plant disease diagnosis, Product identification. | | |
| 2 | I would like to know examples of applications that can be created using Object Detection. | AI Applications that can be created using Object Detection include Traffic monitoring, Security surveillance, Wildlife tracking. | | |
| 2 | I would like to know examples of applications that can be created using Segmentation. | AI Applications that can be created using Segmentation include Medical image analysis, Autonomous vehicle navigation, Agricultural crop monitoring. | | |
| 2 | I would like to know examples of applications that can be created using Pose Estimation. | AI Applications that can be created using Pose Estimation include Sports performance analysis, Physical therapy, Animation and Gaming. | | |
| 2 | I would like to know examples of applications that can be created using Image Enhancement. | AI Applications that can be created using Image Enhancement include Restoration of old photos, Clarity improvement in medical imaging, Enhancement of security camera image. | | |
| 2 | I would like to know examples of applications that can be created using Depth Estimation. | AI Applications that can be created using Depth Estimation include 3D modeling, Augmented reality AI Applications, Robot navigation. | | |
| 2 | I would like to know examples of applications that can be created using Recommendation System. | AI Applications that can be created using Recommendation System include Personalized online shopping, Content recommendation in streaming services, Customized news feed. | | |
| 2 | I would like to know examples of applications that can be created using Anomaly Detection | AI Applications that can be created using Anomaly Detection include Credit card fraud detection, Network security, Predictive maintenance in manufacturing. | | |
| 2 | I would like to know examples of applications that can be created using Computer Vision for Healthcare. | AI Applications that can be created using Computer Vision for Healthcare include Tumor detection in radiology images, Patient monitoring systems, Surgical assistance. | | |
| 2 | I would like to know examples of applications that can be created using Augmented and Virtual Reality. | AI Applications that can be created using augmented reality and virtual reality include Virtual training environments, Immersive gaming, Interactive educational contents. | | |
| 2 | I want to know the planning intention of AI Challenger. | Samsung System LSI's AI Challenger aims to cultivate AI expertise among Korean students and postgraduates through practical engagement and innovation. The program’s objectives include advancing Edge AI technology by equipping participants with essential skills, fostering industry-academia collaboration through networking and mentorship, and transforming innovative ideas into tangible prototypes or solutions. Additionally, it offers practical exposure to real-world tech challenges. Ultimately, the AI Challenger seeks to enhance participants' competitiveness in the AI field and fuel the growth of AI technology and industry by serving as a breeding ground for applicable tech solutions. | | |
| 2 | I would like to know the device specifications provided by 1st AI Challenger. | The device features the Exynos 2200 chipset, with a CPU configuration of Cortex®-X2, Cortex®-A710, and Cortex®-A510, alongside the Samsung Xclipse 920 GPU. It's equipped with an AI Engine incorporating a Dual-core NPU and DSP. The camera capabilities include support for up to 200MP in single camera mode, 108MP at 30fps for a single camera, and a dual-camera setup of 64MP + 32MP at 30fps. Video performance allows up to 8K decoding at 60fps for 10-bit HEVC(H.265) and 30fps for 10-bit VP9, AV1, along with 8K encoding at 30fps for 10-bit HEVC(H.265) and VP9. The display supports 4K/WQUXGA at 120Hz and QHD+ at 144Hz. | | |
| 2 | I would like to know the advantages of Exynos 2200, the base chipset of the reference device provided by First AI Challenger managers. | Console quality graphics now on mobile with the Exynos 2200 mobile processor. The Samsung Xclipse GPU sets to usher in a new era and completely change the way we experience mobile gaming. Playtime is well and truly over. | | |
| 2 | I want to know what types of smartphones were mass-produced based on the Exynos 2200. | Smartphones mass-produced based on the Exynos 2200 chipset include Samsung Galaxy S22 Ultra, Samsung Galaxy S22+, and Samsung Galaxy S22. | | |
| 2 | I want to know the main summary of Samsung Exynos 2200. | Samsung Electronics introduces the Exynos 2200, a premium mobile processor featuring the Samsung Xclipse GPU based on AMD RDNA 2 architecture, crafted using a 4-nanometer EUV process. This processor enhances mobile gaming, social media, and photography, offering the first mobile hardware-accelerated ray tracing for realistic lighting and variable rate shading for smoother gameplay. It boasts an upgraded neural processing unit (NPU), Arm's latest Armv9 CPU cores for superior performance and security, and a fast 5G modem compliant with 3GPP Release 16. With integrated Secure Element (iSE) and robust encryption, it ensures data security. The processor's ISP supports up to 200MP resolution, AI-integrated camera functionalities for professional-quality images, an advanced codec for up to 8K video, and a display solution with HDR10+ and up to 144Hz refresh rates, ensuring a comprehensive and immersive visual experience. | | |
| 2 | How do I post an article in the Best Lab? | Best Lab is a dedicated platform where administrators meticulously select and upload projects that demonstrate excellence and have high potential for development. Currently, up to six projects are featured, receiving special attention. There is no separate process for individuals to submit their projects directly to Best Lab. Instead, administrators personally curate and showcase projects on the platform. | | |
| 2 | How many projects can I create maximum in the SDK Service? | In the SDK Service, you are allowed to create a maximum of 5 tabs. If you attempt to create more than 5 tabs, a warning popup will appear to notify you of the limit. Please adjust your projects accordingly within this constraint. | | |
| 2 | In the SDK Service, I'm unable to download the converted log and NNC file. | You can download the converted log and NNC file for up to 7 days after their creation. After this period, they are deleted, and the download button is disabled. If you need to download them again, you'll have to re-upload the same model file and repeat the conversion under the same conditions. | | |
| 2 | I entered the wrong password multiple times, and now my account is locked. What should I do? | In the Eco System, an account gets locked after 5 consecutive incorrect password attempts. If your account is locked, you can unlock it by going through the verification process sent to your linked email address. Please check your email for the necessary steps to regain access to your account. | | |

# 테스트셋 평가

## 사용한 Tool

데이터셋 평가 프레임워크 **RAGAS (ver 0.1.10)**

LLM: OpenAI **gpt-4o**

## 파이프라인

### 아키텍처

|  |
| --- |
| **AI Studio 2.0 문서 적용 세부 일정과 파이프라인** |
| **참고: 추론(AWS) 및 품질 검증(로컬 PC) 파이프라인** |

### 세부 설명

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Phase | items | description | comments |  |  |
| Document 준비 | Chunking | S3에서 데이터 파일을 읽어와 Markdown Chunker로 Chunking | Markdown Chunker\*로 문서 분리 후 총 Context Chunk갯수 350개  Markdown header + text 단위로 나눈 다음에 token 길이가 100이하면 합치고 300 이상이면 나눈다. |  |  |
| 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 | DocumentDB에서 ***테스트셋 질문***과 코사인 거리가 가까운 top chunk 추출 | DocumentDB에서 max document 20개로 셋팅해서 추출을 하는데 가져온다. HNSW 방식에 따라 관련 있는 일부 chunk만 retrieve 하는 방식. 20개를 가져오라고 시키는데 때에 따라 이보다 적을 수도 있다. |  |  |
| Candidate Selection | 기준에 맞도록 Context Chunk 선정 | 현재 셋팅에서는 retrieval된 20개 중 top-10개만 사용한다. |  |  |
| LLM 집어 넣기 전처리 |  |  |  |  |
| Answer Generation | ***테스트셋 질문***과 선정된 Chunk를 LLM (오픈AI ***gpt-4o***)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 ***gpt-4o***)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 |  |  |

# Query 수행시 사용하는 시스템 프롬프트 전체 텍스트

|  |
| --- |
| You are a helpful assistant designed to help developers create applications using the ENN SDK (Exynos Neural Network Software Development Kit).  With each query from the user, a set of context data will be given.  Your primary role is to answer the query based on the provided context.  Context data may be in the form of a question-and-answer pair or excerpt from a specific section of the documentation.  You will help users with:  - Model Conversion: Providing instructions on converting TensorFlow Lite models to Neural Network Container (NNC) models using the ENN SDK service.  - Model Execution: Explaining how to execute NNC models using the ENN framework on Exynos platforms.  - API Usage: Offering detailed information on using the ENN framework APIs **for** managing tensors, buffers, and model execution processes.  - Support and Troubleshooting: Assisting with support queries by referencing the support matrix, FAQs, and sample codes. Guiding users on reporting bugs and accessing additional resources from the Exynos Developer Society web page.  Answer the query based on the given context information, general AI and/or SDK knowledge, and your role.  If the query is beyond the scope of the given context, respond with : 'We do not have the information you requested. If you wish to contact support about this inquiry, please send an email to seed.ai@samsung.com.'.  Be aware and forgiving of spelling and/or typing errors.  Keep in mind that system messages are invisible to the user. Only user and assistant messages are visible.  Remember to be concise, clear, and refer to specific sections of the documents when providing assistance. |

# 241018 평가 결과 요약

결과 원본 파일 링크(Context 포함)

원본 결과 파일 : [result\_OpenAI\_gpt-4o\_20241018.zip](https://confluence.samsungds.com/download/attachments/145687075/result_OpenAI_gpt-4o_20241018.zip?api=v2&modificationDate=1729559847000&version=2)

원본 결과 파일(CSV) : [result\_OpenAI\_gpt-4o\_20241018.csv](https://confluence.samsungds.com/download/attachments/145687075/result_OpenAI_gpt-4o_20241018.csv?api=v2&modificationDate=1729559865000&version=3)

(\*참고 : OpenAI Credit 부족으로 마지막 테스트 셋 하나만 Score 산출이 안되어서 평균 값 계산할 때 제외했고, 충전 후 마지막 테스트 셋에 대해 업데이트 예정입니다.)

|  |  |  |  |
| --- | --- | --- | --- |
| **context\_relevancy** | **faithfulness** | **answer\_correctness** | **설명** |
| **0.03(0.04)** | **0.88(0.82)** | **0.73(0.61)** | 중요한 변화: **테스트셋이 전체를 커버하도록 업데이트 되었습니다.**  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 나쁜 이유는 top-10개를 추출할 때 20개 chunk를 가져오며 이때 관련성이 먼 chunk들도 꽤 추출된다는 뜻입니다. 품질 보다는 운영 비용에 더 가까운 이슈입니다. * faithfulness(컨텍스트로부터 답변이 잘 추론되었나): 추출된 chunk 개수와 총 데이터 양이 많은 편이라 해당 컨텍스트에 답변이 포함되어 있는 가능성이 높습니다. * answer\_correctness(답변을 ground truth와 비교한 값): 기존 대비 0.73으로 다소 떨어진 이유는 RAGAS로 생성한 항목들이 많이 추가되어(난이도 상승) 하락한 것으로 보입니다.   괄호() 내의 수치는 FAQ 항목(답이 이미 있는)수치을 제외한 통계입니다(이것이 실제에 가깝습니다). |

# 인사이트 & 다음 수정 내용( 2024. 10. 29 기준)

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 번호 | 문제점 | 추가 설명 | 개선 요청 | 수정 가능 여부 | 완료 예상 일정 | 코멘트 |
| 1 | 데이터셋 평가 프레임워크 RAGAS의 버전 업그레이드 필요 | 현재 ver 0.1.10 | 0.2.1 이상 버전으로 업그레이드  테스트셋 생성 및 데이터셋 평가를 공히 0.2.1 버전 이상을 사용하도록 변경 | 가능 | query 비율만 컨펌되면 완료 | ComparativeAbstractQuerySynthesizer, SpecificQuerySynthesizer 두 가지 query 종류 비율을 어떻게 할지 컨펌 필요.  Testset generation 할 때 metadata가 적용되지 않는 것으로 보임(참고). |
| 2 | Metrics 수정 사항 반영 필요 | 버전이 올라가며 주요 metrics의 이름이 바뀌었으며, 추가로 Semantic Similarity도 포함하게 되었음 | Context Recall (context\_relevancy → LLMContextRecall)  Faithfulness (변경 없음)  Factual Correctness (answer\_correctness → Factual Correctness)  Semantic Similarity 추가. 생성된 답변과 레퍼런스 답변의 의미적 유사도. | 가능 | 완료 | 수정 및 샘플 테스트셋으로 테스트 완료. |
| 3 | 테스트셋에 FAQ(QA 데이터) 항목 조정 | FAQ 내용도 집어 넣었는데(잘 나오는지 확인은 필요하니) 그 비중이 50%로(47/110) 과다함 | 대표 사례 몇개만 남기고 나머지 삭제 필요  모든 데이터셋/테스트셋에서 수정 필요 | 가능 |  | FAQ 내용 정리하여 컨펌 필요. |
| 4 | 데이터셋 및 테스트셋 old term 정리 필요 | <Old Term>  "EDS" (Exynos Developer Society)": 포털 사이트 이름  "ENN SDK" | "EDS" ==> "Exynos Developer Society" 사이트로 이름 변경  "ENN SDK" ==> "Exynos AI Studio" 로 이름 변경  모든 데이터셋/테스트셋에서 수정 필요 | 가능 |  | 데이터셋 수정 완료.  테스트셋 수정 필요. |
| 5 | Faithfulness에 0이 있는 항목 정리 필요 | Faithfulness가 0인 문항들 확인 및 조치 예정.  "What is the purpose of applying Fixed Precision Quantization to the optimized CNNX model?" : Faithfulness가 0이며, 이 뜻은 answer가 context의 내용을 전혀 사용하지 않고 만들어졌다는 이야기입니다. 이례적이라 확인 필요.  "How does the data path affect quantization precision":  Faithfulness가 0인데 내용을 보니 애초에 ground\_truth 자체가 잘못 만들어져 있는 것 같습니다: "The answer to given question is not present in context" | "What is the purpose of applying Fixed Precision Quantization to the optimized CNNX model?" 디버깅 및 수정.  "How does the data path affect quantization precision?" 해당 항목 삭제 | 가능 |  | 테스트셋 항목 삭제 예정  업데이트된 버전에 맞는 새로운 테스트셋 컨펌 필요. |
| 6 | 시스템 프롬프트 수정 필요 | ENN SDK 용어 변경 및 기타 수정 필요. | ENN SDK → "Exynos AI Studio"로 변경, 기타 필요한 수정 수행. |  |  |  |
| 7 | 악성/오용 프롬프트 방어 구현 | 키워드 필터링, 문맥 이해 필터링, 응답 제한, 지식 범위 제한, 응답 템플릿 제약, 시스템 프롬프트 추출 방어, 이용약관 검토 | OpenAI moderation API 활용, 프롬프트 엔지니어링, 부적절한 컨텐츠 응답 제한, 지식 범위 제한, 응답 템플릿 제약, 시스템 프롬프트 추출 방어 등 구현. |  |  |  |
| 8 | 프롬프트 선호도 (like/dislike 구현) 및 품질 문제 신고 채널 | 프롬프트 결과에 대해 like/dislike 표시와 dislike를 눌렀을 때 설문이 필요 | like/dislike 아이콘 구현  like 눌렀을 때 설문 페이지 구현  dislike 눌렀을 때 설문 페이지 구현  like와 dislike의 결과를 AWS DB에 저장 구현 |  |  |  |
| 9 | 사용자 사용 패턴 모니터링 방안 고안 | 악성 프롬프트나 시도의 방어를 위한 사용자 사용 패턴 모니터링 방안 고안과 구현 | 지속적인 사용자 usage 모니터링 방안 구현 |  |  |  |
| 10 | Chatbot 이름 지정 | Chatbot에 이름을 짓는다. | 시스템 프롬프트 상에 Chatbot의 이름을 "Exynos AI Studio Assistant"로 명명 |  |  |  |
| 11 | 연속 질문 | 연속 질문에 대한 대답 필요 | 이전 Q/A 쌍의 History 활용 |  |  |  |
| 12 | 이미지 설명 | Document 내 Flow Chart 등의 설명 | 텍스트를 넣고 이미지를 링크로 걸어서 이미지에 대한 정보 제공 |  |  |  |
| 13 | 대화 저장 | 대화 History를 사용자가 PDF/TEXT 등으로 저장 할 수 있는 기능 | 기능 추가 |  |  |  |
| 14 | 세션유지시간 | 사용자 활동이 없는 세션의 유지 시간 | 유지 시간을 정책으로 정하고 세션 종료 추 |  |  |  |
| 15 | 표 출력 | 표 스타일의 설명을 요청할 경우 LLM의 대답을 표 UI 형태로 출력 | Markdown을 표 형태로 보여줄 수 있는 뷰어로 교체 |  |  |  |

참고: 시스템 프롬프트 현재본 ==> 가장 최신 버전을 가져와서 여기에 붙여 주세요!!

|  |
| --- |
| You are a helpful assistant designed to help developers create applications using the ENN SDK (Exynos Neural Network Software Development Kit).  With each query from the user, a set of context data will be given.  Your primary role is to answer the query based on the provided context.  Context data may be in the form of a question-and-answer pair or excerpt from a specific section of the documentation.  You will help users with:  - Model Conversion: Providing instructions on converting TensorFlow Lite models to Neural Network Container (NNC) models using the ENN SDK service.  - Model Execution: Explaining how to execute NNC models using the ENN framework on Exynos platforms.  - API Usage: Offering detailed information on using the ENN framework APIs **for** managing tensors, buffers, and model execution processes.  - Support and Troubleshooting: Assisting with support queries by referencing the support matrix, FAQs, and sample codes. Guiding users on reporting bugs and accessing additional resources from the Exynos Developer Society web page.  Answer the query based on the given context information, general AI and/or SDK knowledge, and your role.  If the query is beyond the scope of the given context, respond with : 'We do not have the information you requested. If you wish to contact support about this inquiry, please send an email to seed.ai@samsung.com.'.  Be aware and forgiving of spelling and/or typing errors.  Keep in mind that system messages are invisible to the user. Only user and assistant messages are visible.  Remember to be concise, clear, and refer to specific sections of the documents when providing assistance. |