RN00459

Ara2 Software Packages Release Notes

Rev. 2.1 — 28 July 2026 **Release notes**

Document information

Information	Content
Keywords	RN00459, Ara240, Ara2 Runtime SDK, Ara2 Vision Example, LLM, VLM, eIQ AAF Connector, CNN
Abstract	This document contains important information about the contents of different packages and demos available for Ara240 16GB M.2 and USB Modules with FRDM i.MX 8M Plus, FRDM i.MX 95, and FRDM i.MX 95 Pro boards.

1 Overview

This document contains important information about the contents of different packages and demos available for Ara240 16GB M.2 and USB Modules with FRDM i.MX 8M Plus, FRDM i.MX 95, and FRDM i.MX 95 Pro boards. It covers supported features, known issues, and limitations in this release.

2 Features

Table 1. Ara2 software packages

Package	Features
Ara2 Runtime Package	Provides a runtime environment for Ara240. Provides scripts to download and run pre-compiled CNN and GenAI models on Ara240 modules.
Ara2 Vision Example	Demonstrate a Vision pipeline capable of handling up to 8 simultaneous video streams for object detection on the Ara240 discrete NPU.
LLM Edge Studio	Showcase unimodal capabilities on i.MX for supported Large Language Models (LLMs) accelerated by the Ara240 discrete NPU at the edge.
VLM Edge Studio	Showcase Ara240 multimodal capabilities on i.MX using VLMs for video and image analysis.
eIQ AAF Connector	Service interface that exposes an OpenAI-compatible REST API to client applications while brokering requests to locally hosted LLM/VLM backends on Ara240 discrete NPU.
Smart Device Gateway	Service that allows devices to become voice-enabled endpoints that communicate with a central gateway running advanced AI models.

2.1 Models

CNNs

- https://huggingface.co/nxp/YOLOv8
  - YOLOv8n (nano)
  - YOLOv8s (small)
  - YOLOv8m (medium)
  - YOLOv8l (large)
  - YOLOv8x (extra large)
- https://huggingface.co/nxp/YOLOX-Detection-Ara240
  - YOLOXs (small)
  - YOLOXm (medium)
  - YOLOXl (large)
  - YOLOXx (extra large)

LLMs

- https://huggingface.co/nxp/Qwen2.5-Coder-1.5B-Ara240
- https://huggingface.co/nxp/Qwen2.5-7B-Instruct-Ara240

VLMs

All information provided in this document is subject to legal disclaimers.

© 2026 NXP B.V. All rights reserved.

- https://huggingface.co/nxp/Qwen2.5-VL-7B-Instruct-Ara240

3 Licenses

Table 2. Package license information

Package name	Version	License
Ara2 Runtime SDK	2.1.1	LA_OPT_NXP_Software_License
Ara2 Vision Examples	1.0.1	LA_OPT_Online Code Hosting NXP_Software_License
LLM Edge Studio	2.0.1	LA_OPT_Online Code Hosting NXP_Software_License
Smart Device Gateway	1.0.0	LA_OPT_OnlineCode Hosting NXP_Software_License
VLM Edge Studio	1.0.1	LA_OPT_Online Code Hosting NXP_Software_License
eIQ Connector	2.1	LA_OPT_Online Code Hosting NXP_Software_License

4 Supported devices

Target Hosts:

- FRDM i.MX 8M Plus
- FRDM i.MX 95
- FRDM i.MX 95 PRO

5 System configuration

Supported OS:

- i.MX Embedded Linux Versions:
  - LF6.18.20_2.0.0

Supported/Required hardware:

- Supported FRDM i.MX platform
- Ara240 DNPU
- Ara240 16GB M.2 Module
- microSD >= 64 GB storage
- USB-C debug cable
- Internet (Ethernet)
- HDMI monitor
- 100 W power supply (Anker Charger is recommended)
- USB mouse
- USB keyword
- 1080p USB camera

6 Release contents

This section provides an overview of the software packages, components, and features included in this release.

All information provided in this document is subject to legal disclaimers.

© 2026 NXP B.V. All rights reserved.

6.1 Ara2 runtime SDK package contents

Package: `rt-sdk-ara2`

Version: 2.1.1

The `rt-sdk-ara2` Debian package provides a complete runtime environment for AI/ML acceleration using the Ara240 Discrete NPU on FRDM i.MX 8M Plus, FRDM i.MX 95, and FRDM i.MX 95 Pro.

This release includes significant improvements to the `RT-SDK-ARA2`. It focuses on automatic device management, installation improvements, performance optimizations, enhanced model support, and bug fixes.

6.1.1 New features

- The package is now pre-installed into the Board Support Package (BSP) image (i.MX Embedded Linux image starting 6.18.20_2.0.0), therefore no need to manually install the Debian package for runtime.
- The debian package is still supported but only for custom builds.
- Smart driver installation: Driver only installed if not present and `vermagic` matches kernel version.
- Reboot message: Add a gentle message after installation to reboot the system.

Yocto

- `uiodma` is removed from the package and it is built as an out-off tree driver. It is located in `/usr/lib/modules/$(uname -r)/extra/`.

Service

- Ara2 Management utility: `aam` utility for controlling automatic device detection and bring-up.
- Target-based architecture: with `ara2-device.target` and `rt-sdk-ara2.service`.
- Service management: `rt-sdk-ara2.service` now triggered by `udev` (not auto-enabled).
- Hot-plug Support: Added udev-based hot-plug support for PCIe and USB Ara240 devices.

Multi-device support

- Hardware utilities: Enhancement for multi-endpoint scenarios.
- System service: Enable capabilities to run the bring-up process on all devices.
- Multi-modules support: Support for USB and PCIe Ara240 Modules.

GStreamer

- Merged preprocess, inference, and postprocess plugins to optimize CPU usage.
- Full YOLOx support: Added comprehensive post-processing support for YOLOx models.
- Model variants: Support for YOLOx variants (l, m, s, x)
- Multi-resolution: Enabled support for both 512×512 and 640×640 input resolutions.
- Performance optimizations: Add multiple optimizations on pre-processing and post-processing stages.

Breaking changes

- Package rename: Changed package name from `rt-sdk-ara2` to `imx-nxp-ara2`.
- Optimum-ara version: Reset version to v1.0.0.

Bug fixes

- Service dependencies: Fixed service dependency relationships using Wants/PartOf/Upholds directives, which was affecting `eiq-aaf-connector.service` when running from startup.
- RGB mem-set bug: Fixed bug requiring 6 mem-set (2 for each plane from RGB).
- `fetch_model` script: Fixed script not returning error code.
- Deprecated scripts: Removed deprecated scripts from `rt-sdk-ara2.service`.

All information provided in this document is subject to legal disclaimers.

© 2026 NXP B.V. All rights reserved.

6.2 Ara2 vision example release contents

Package: `ara2-vision-examples`

Version: 1.1.0

Comprehensive vision example demonstrating real-time multi-stream object detection using YOLOv8 models accelerated by Ara240 DNPU on FRDM i.MX 8M Plus development board, FRDM i.MX 95, add FRDM i.MX 95 Pro development boards.

Key features:

- Multi-stream processing: Supports 1-8 concurrent video streams with real-time object detection.
- Hardware acceleration: Leverages Ara240 DNPU for efficient YOLOv8/YOLOX detection inference.
- GStreamer integration: Robust pipeline management with hardware-accelerated video decoding (H.265 support).
- Visual output: Mosaic display with bounding boxes for detected objects.
- GTK3 GUI launcher: Full-featured graphical interface for easy configuration and management.
- Multiple input sources: Support for cameras (V4L2), RTSP streams (low latency ~100 ms), and test patterns.
- Runtime model selection: Automatic model download, validation, and GUI-based management.
- Flexible display options: Configurable bounding boxes, OSD statistics, and performance metrics.
- GoPoint launcher — Dedicated launcher integration for GoPoint.

Supported platforms:

- FRDM i.MX 8M Plus
- FRDM i.MX 95
- FRDM i.MX 95 Pro

Models supported:

- nxp/YOLOv8
  - YOLOv8n (nano)
  - YOLOv8s (small)
  - YOLOv8m (medium)
  - YOLOv8l (large)
  - YOLOv8x (extralarge)
- nxp/YOLOX
  - YOLOXs (small)
  - YOLOXm (medium)
  - YOLOXl (large)
  - YOLOXx (extralarge)

Target applications:

- Multi-camera surveillance systems
- Industrial monitoring and quality control
- Smart city infrastructure
- Retail analytics
- Autonomous vehicle testing
- RTSP-based remote monitoring

License: NXP Proprietary (LA_OPT_Online Code Hosting)

All information provided in this document is subject to legal disclaimers.

© 2026 NXP B.V. All rights reserved.

6.2.1 What is New in v1.1.0 (new features and enhancements)

GTK3 GUI launcher

- Added a full-featured GTK3 GUI launcher.
- Supports configuration for:
  - Streams
  - Models
  - Display settings
  - Synchronization
  - Statistics

Multi-source support

- Added support for multiple input sources:
  - Cameras (V4L2)
  - RTSP streams (low latency ~100ms)
  - Test patterns
  - Mixed inputs
- JSON-based configuration support

Runtime model selection

- Added support for YOLOX

Video handling improvements

- Added video auto-discovery
- On-demand video downloads
- Support for H.265 codec
- Removed bundled videos to reduce package size

New performance and display flags

- `--no-bbox`
- `--no-osd-stats`
- `--only-bbox`
- Configurable statistics interval

Pipeline performance improvements

- Optimized buffering
- Improved FPS performance
- Reduced CPU usage

Resource management enhancements

- Per-process shared memory handling
- Unique shared memory (SHM) paths for improved stability

User experience improvements

- Automatic resource downloads
- Connectivity checks
- Progress feedback during operations
- Improved help and documentation

Bug fixes

All information provided in this document is subject to legal disclaimers.

© 2026 NXP B.V. All rights reserved.

- Fixed stream configuration issues
- Resolved endpoint selection problems
- Improved memory handling stability
- Fixed installer formatting issues

Performance improvements

- Better FPS scaling across workloads
- Improved synchronization stability on supported hardware

Known issues

- YOLOX models are not available to select the GTK GUI yet.
- RTSP streams do not show bounding boxes.
- Gopoint GUI does not close the `ara2-vision-exmpales` when is running inferences.

6.3 LLM Edge Studio release contents

Package: `llm-edge-studio`

Version: 2.0.1

Large Language Model (LLM) Edge Studio is a Qt/QML-based GUI application. It supports FRDM i.MX 95, FRDM i.MX 95 Pro, and FRDM i.MX 8M Plus. LLM Edge Studio also enables rapid interaction with locally running LLMs. It uses the eIQ AAF connector to communicate with the Ara240 Runtime SDK and provides a user-friendly GUI for model selection and prompt input.

Key features:

- User-friendly GUI for model selection and prompt submission
- Real-time streaming token responses via Server-Sent Events (SSE)
- Integration with eIQ AAF connector for HTTP-based communication
- Support for Qwen2.5 models (7B-Instruct and Coder-1.5B)
- Model loading/unloading with progress tracking
- Endpoint support — Configurable backend endpoints for flexible deployment
- Automatic model download — Models fetched directly from NXP hugging face repository
- GoPoint launcher — Dedicated launcher integration for GoPoint
- Model loading progress — Animated progress bar with time estimation per model

Architecture:

- Frontend: Qt 6.5+ with QML for responsive UI
- Backend: C++ with AAF Connector for REST API communication
- Communication: HTTP POST/GET with SSE streaming
- Target: Embedded Linux (>= LF6.18.2_1.0.0)

Supported platforms:

- FRDM i.MX 8M Plus
- FRDM i.MX 95
- FRDM i.MX 95 Pro

Models supported:

- nxp/Qwen2.5-Coder-1.5B-Ara240
- nxp/Qwen2.5-7B-Instruct-Ara240

License: NXP Proprietary (LA_OPT_Online Code Hosting)

All information provided in this document is subject to legal disclaimers.

© 2026 NXP B.V. All rights reserved.

The known issues were addressed in the following link: DNPU Hub.

6.3.1 New features and enhancements

- Enable endpoint support in backend and update layout UI — Backend now supports configurable endpoints, accompanied by layout UI improvements.
- Implement automatic model download from NXP Hugging Face — Models can now be automatically fetched and downloaded directly from the NXP Hugging Face repository.
- Added launcher for GoPoint — Introduced a dedicated launcher integration for GoPoint.

Known issues

- Display inference metrics failure — Inference metrics may fail to display correctly in the UI after model inference completes.
- Load model payload endpoint key change — The load model request payload endpoint key was renamed from "endpoint" to "ara_endpoint" to match the Ara runtime API specification, causing model loading to fail with older versions of the Ara SDK.

6.4 VLM Edge Studio release contents

Package: `vlm-edge-studio`

Version: 1.0.0

Vision Language Model (VLM) Edge Studio is a Qt/QML-based GUI application designed for NXP's FRDM i.MX (8M Plus and i.MX 95). It interacts with vision language models accelerated by the Ara240 discrete NPU at the edge. It enables rapid interaction with locally running VLMs. It uses the eIQ AAF connector to communicate with the Ara240 Runtime SDK and provides a user-friendly GUI for model selection and prompt input.

Key features:

- User-friendly GUI for model selection and prompt submission
- Real-time streaming token responses via Server-Sent Events (SSE)
- Integration with eIQ AAF connector for HTTP-based communication
- Model loading/unloading with progress tracking
- Performance metrics display (TTFT, TPS)
- Support for Qwen2.5 VL 7B model
- Video or image mode analysis support
- USB-C type camera support for video recording/image capture for vision analysis
- Endpoint support — Configurable backend endpoints for flexible deployment
- Automatic model download — Models fetched directly from NXP hugging face repository
- Video analysis — Support for video analysis using VLM inference model
- Camera integration — Live camera stream with video recording
- GoPoint launcher — Dedicated launcher integration for GoPoint
- Model loading progress — Animated progress bar with time estimation per model

Architecture:

- Frontend: Qt 6.5+ with QML for responsive UI
- Backend: C++ with AAF connector for REST API communication
- Communication: HTTP POST/GET with SSE streaming
- Target: Embedded Linux (>= LF6.18.2_1.0.0)

Supported platforms:

- FRDM i.MX 8M Plus

All information provided in this document is subject to legal disclaimers.

© 2026 NXP B.V. All rights reserved.

- FRDM i.MX 95
- FRDM i.MX 95 Pro

Models supported:

- nxp/Qwen2.5-VL-7B-Instruct-Ara240

License: NXP Proprietary (LA_OPT_Online Code Hosting)

The known issues were addressed in the following link: DNPU Hub.

6.4.1 New features and enhancements

- Enable endpoint support in backend and update layout UI — Backend now supports configurable endpoints, accompanied by layout UI improvements.
- Refactor `main.qml` to support column and row layouts for content — Enhanced layout flexibility in the main QML file to support both column and row orientations.
- Implement automatic model download from NXP Hugging Face — Models can now be automatically fetched and downloaded directly from the NXP Hugging Face repository.
- Added launcher for GoPoint — Introduced a dedicated launcher integration for GoPoint.
- Added missing methods — Implemented previously missing methods to ensure complete functionality.
- Updated `README.md` — General documentation updates and improvements.

Bug fixes

- Headers and updated SBOM files — Corrected header inconsistencies and refreshed Software Bill of Materials (SBOM) files.
- Image and video mode switching content — Resolved an issue where content was not properly switching between image and video modes.
- Update load model progress bar — Fixed the progress bar behavior when loading a model.
- EIQ AAF connector `server_config.json` path — Corrected the file path resolution for the EIQ AAF-Connector server configuration.
- Loading bar animation — Small but impactful fix to ensure the loading bar animation runs smoothly.

Known issues

- Qwen-VL initialization failure — Qwen-VL model may fail to initialize correctly on startup when a demo runs without Internet connectivity.
- Display inference metrics failure — Inference metrics may fail to display correctly in the UI after model inference completes.
- Text field cursor not visible — The text input cursor is not visible when the prompt input box is selected or activated.
- Load model payload endpoint key change — The load model request payload endpoint key was renamed from "endpoint" to "ara_endpoint" to match the Ara runtime API specification, causing model loading to fail with older versions of the Ara SDK.

6.5 eIQ AAF connection release contents

Package: `eiq-aaf-connector`

Version: 2.1

eIQ AAF connector — OpenAI compatible server for LLMs. The eIQ AAF connector provides a REST-based server for LLM inference on Ara240 discrete NPU. It offers OpenAI-compatible APIs for language model interactions on i.MX devices.

Supported platforms:

All information provided in this document is subject to legal disclaimers.

© 2026 NXP B.V. All rights reserved.

- FRDM i.MX 8M Plus
- FRDM i.MX 95
- FRDM i.MX 95 Pro

License: NXP Proprietary (LA_OPT_Online Code Hosting)

6.5.1 New features

ARA endpoint selection and device information API

- Added `ara_endpoint` parameter to `POST /v1/models` to enable selecting which ARA device to load a model on
- Implemented new `GET /devices` endpoint that provides:
  - Information about the current state of ARA endpoints
  - Models loaded on each endpoint
  - Memory usage statistics

Structure for metrics response

- Added structured Pydantic model for metrics response

Integration with Yocto images

- Installation logic moved from `postinst` to `install.sh` for unified Yocto and Debian package support.
- The `aaf-connector` script now validates the virtual environment before running and creates it if not found.

Bug fixes

- Chat template fix: Prevents passing `None` values in the `apply_chat_template()` to avoid template errors on certain models.

Known limitations

- When running `uv` commands over serial port, use `--no-progress` flag to avoid hangs.
- `GET /devices` endpoint cannot distinguish between modules of 16 GB or 8 GB.

6.6 Smart device gateway release contents

Version: 1.0.0

This demo application transforms the way we interact with everyday devices. Instead of embedding expensive AI chips into every device, our architecture centralizes intelligence in one powerful server. Your devices become voice-enabled endpoints that communicate with a central gateway running advanced AI models.

This demo can be deployed on FRDM i.MX 95, FRDM i.MX 95 Pro, or FRDM i.MX 8M Plus development boards.

Key features:

- You only need one server with AI: Centralized intelligence reduces hardware costs across your device ecosystem
- Hardware acceleration: Uses the Qwen 2.5-7B model accelerated by the Ara240 for the inference on user prompts
- Context-rich responses: Each device can tap into its own Retrieval-Augmented Generation (RAG) database—recipe collections, user manuals, personal notes, or domain-specific knowledge.
- Privacy-first design: Your voice never leaves your premises. Everything runs locally, keeping your conversations private and secure.
- Universal compatibility: Any device with a microphone, a speaker, and the ability to send HTTP requests can join the ecosystem.

All information provided in this document is subject to legal disclaimers.

© 2026 NXP B.V. All rights reserved.

- Easy-to-use design: You can talk naturally to any connected device, and get a natural voice response back, with personalized knowledge of your device.

Supported platforms:

- FRDM i.MX 8M Plus
- FRDM i.MX 95
- FRDM i.MX 95 Pro

Model supported:

- `nxp/Qwen2.5-7B-Instruct-Ara240`

Target applications:

- Smart home applications
- Voice-enabled IoT devices
- Information retrieval systems
- Smart appliance control
- Device management and monitoring
- Easy-to-access device manuals and documentation

License: NXP Proprietary (LA_OPT_Online Code Hosting)

6.6.1 Features and enhancements

- Compatible with host PC client: You can connect your own PC (Mac OS, Linux, or Windows) as a client to the smart device gateway server using the provided `Push_to_talk` client application.
- RAG database integration: Seamlessly integrate custom knowledge bases and documents into each device's local RAG database for contextual and personalized responses.
- Multi-device support: Connect and manage multiple IoT devices simultaneously through a single gateway server with independent context management per device.
- Voice processing pipeline: Integrated speech-to-text (STT) and text-to-speech (TTS) conversion enabling natural voice interactions with AI models.

Important considerations

To ensure the best experience with the smart device gateway, keep the following points in mind:

- Microphone quality matters
  The quality of your microphone directly impacts transcription accuracy. For optimal results:
  – Use a high-quality microphone with good noise cancellation
  – Minimize background noise during recording
  – Speak clearly and at a moderate pace
- Speech recognition sensitivity
  The STT model used in this demo is highly sensitive to accents and pronunciation variations. If you experience transcription issues:
  – Try speaking more slowly and enunciating clearly
  – Use standard pronunciation when possible
  – Be aware that strong regional accents may affect accuracy
- Knowledge base boundaries
  For the best responses, formulate queries that align with the information available in the device manuals. The system works best when:
  – Questions relate directly to features, operations, or specifications covered in the user manuals
  – Queries are specific to the selected device (oven or barista)

All information provided in this document is subject to legal disclaimers.

© 2026 NXP B.V. All rights reserved.

– Questions focus on practical usage, troubleshooting, or device capabilities

Known issues

- The first time running the demo can take a long time. Be patient while the model is being loaded and optimized on the Ara240.
- If a connector was previously running, ensure to properly disconnect it before starting a new session, this can cause the demo to fail.

7 Known issues

This section provides a summary of known issues, limitations, and available workarounds for the current release.

7.1 Ara240 with FRDM i.MX 8M Plus boot failures

Description:

FRDM i.MX 8M Plus Ara240 initialization may fail or freeze the boot-up process if ASPM PCIe is not disabled.

Workaround:

1. Install Q2'2026 BSP 6.18.20_2.0.0.
2. Stop at U-Boot, reboot and press any key.
3. Append `pcie_aspm=off` to the `mmcargs`, save setting using `saveenv`, and then continue boot.
4. Power cycle the FRDM i.MX 8M Plus board and continue firmware upgrade using steps in README.

```
u-boot=> editenv mmcargs
edit: setenv bootargs ${jh_clk} ${mcore_clk} console=${console} root=
${mmcroot} pcie_aspm=off u-boot => saveenv
u-boot => boot
```

7.2 Qwen2.5-Coder-1.5B overflow error in eIQ AAF connector

Description:

eIQ AAF Connector gets into the following error: "OverflowError: out of range integral type conversion attempted" for Qwen2.5-Coder-1.5B model.

Expected behavior

The model should load and work correctly when prompted, without the error preventing its use.

Workaround:

1. Edit the following config file: `/usr/share/eiq/aaf-connector/server_config.json`.
2. Set temperature = 0.1 (should be float > 0.0).
3. Set `top_p` = 0.0 (between 0.0 and 1.0).
4. Set `top_k = 1` (between 1 and vocab_size).
5. Relaunch the eIQ AAF connector to ensure these changes take effect. The model should work as expected.

7.3 FRDM i.MX 8MP, FRDM i.MX 95, and FRDM i.MX 95 Pro freeze after firmware upgrade

Description

After updating ARA240 firmware and rebooting, the board freezes during boot.

All information provided in this document is subject to legal disclaimers.

© 2026 NXP B.V. All rights reserved.

Mitigation

Doing a power cycle after this condition happens, the board should boot correctly this time.

7.4 Model load stalls at ~95 % in VLM-Edge-Studio and LLM-Edge-Studio

Description

When loading a model in VLM Edge Studio or LLM Edge Studio, the progress bar may halt around 95 % and never complete.

This can occur due to two independent root causes:

1. **SSL certificate verification failure due to no Internet connectivity**
   As part of the model load sequence, eiq-connector attempts to download/update Qwen2.5 VL resources by making HTTPS requests to the model's source repository. If the host system has no Internet connectivity, the eiq-connector cannot validate or retrieve metadata, causing the load to stall silently.
   Workaround:
   Patch is available on community.nxp.com.
2. **Model not fully cleaned up when switching between applications**
   If you run vlm-edge-studio and then open llm-edge-studio (or vice versa), closing the first application does not always cause eiq-connector to fully unload the previously loaded model. When the second application attempts to load a new model, the connector may partially reject the load, again causing a 95 % stall.
   Workaround:
   Before switching between VLM and LLM Edge Studio applications, reboot the system to ensure that the connector starts clean.

Recommended combined action

To avoid the 95 % stall entirely, apply both workarounds together:

1. Reboot the system before switching between Edge Studio applications.
2. Ensure that the machine has Internet access and the correct system time before loading a model.

This gives the highest reliability until the underlying issues are fixed. As an extra precaution, monitor `/tmp/connector.log` for any error message.

7.5 ara2_metrics script failure on FRDM i.MX 8MP, FRDM i.MX 95, and FRDM i.MX 95 Pro

Description

Running `ara2_metrics.sh` and selecting option 6 (Utilization, Temperature, DRAM continuously) causes an immediate segmentation fault in the underlying `binary hw_metrics.out`. The issue appears after the first temperature printout.

Expected behavior

The script should start printing periodically the information.

Workaround

Wait one second after the menu is fully displayed.

7.6 Silent hugging face model download failures under poor network conditions

Description

All information provided in this document is subject to legal disclaimers.

© 2026 NXP B.V. All rights reserved.

When installing either llm-edge-studio or vlm-edge-studio, the installation triggers the `fetch_models` script from `rt-sdk-ara2`. The download process appears to complete normally, but the model files, `model.dvm`, are downloaded incompletely. As a result, the model fails to load or initialize during runtime.

Observed behavior

- Download process shows no errors.
- Runtime components fail when trying to load the corrupted/incomplete model.

Expected behavior

- All model files should download completely and load successfully after installation.

Workaround

Do a fresh installation of the package by removing `/usr/share/llm/<AFECTED_MODEL>`, purging the package with `dpkg -P`, and trying again. Lastly, as a double check you can manually compute the md5sum and compare it with the one under `/usr/share/llm/<MODEL>/model.json`.

8 Revision history

Table 3 summarizes revisions to this document.

Table 3. Revision history

Document ID	Release date	Description
RN00459 v.2.1	28 July 2026	• Updated to include information related to the **FRDM-IMX95-PRO** board across the document. • Added Section 7.2 "Qwen2.5-Coder-1.5B overflow error in eIQ AAF connector".
RN00459 v.2.0	3 July 2026	• Added **smart device gateway** package and related documentation. • Introduced **new features, enhancements, and bug fixes** across Ara2 software packages. • Updated **package versions** (RT-SDK, Vision Examples, LLM/VLM Edge Studio, eIQ Connector). • Added **YOLOX model support** and expanded model coverage. • Enhanced **multi-device support, installation flow, and service architecture** in RT-SDK. • Improved **vision example capabilities** with GUI launcher, multi-source input, and performance optimizations. • Added **endpoint support, automatic model download, and GoPoint integration** in LLM and VLM Edge Studio. • Expanded **eIQ AAF connector APIs** (device info and endpoint selection). • Updated **supported platforms and system configuration** (including FRDM i.MX95 PRO and new BSP version). • Revised **known issues and workarounds**.
RN00459 v.1.0	13 April 2026	Initial public release

All information provided in this document is subject to legal disclaimers.

© 2026 NXP B.V. All rights reserved.

Legal information

Definitions

Draft — A draft status on a document indicates that the content is still under internal review and subject to formal approval, which may result in modifications or additions. NXP Semiconductors does not give any representations or warranties as to the accuracy or completeness of information included in a draft version of a document and shall have no liability for the consequences of use of such information.

Disclaimers

Limited warranty and liability — Information in this document is believed to be accurate and reliable. However, NXP Semiconductors does not give any representations or warranties, expressed or implied, as to the accuracy or completeness of such information and shall have no liability for the consequences of use of such information. NXP Semiconductors takes no responsibility for the content in this document if provided by an information source outside of NXP Semiconductors.

In no event shall NXP Semiconductors be liable for any indirect, incidental, punitive, special or consequential damages (including - without limitation - lost profits, lost savings, business interruption, costs related to the removal or replacement of any products or rework charges) whether or not such damages are based on tort (including negligence), warranty, breach of contract or any other legal theory.

Notwithstanding any damages that customer might incur for any reason whatsoever, NXP Semiconductors' aggregate and cumulative liability towards customer for the products described herein shall be limited in accordance with the Terms and conditions of commercial sale of NXP Semiconductors.

Right to make changes — NXP Semiconductors reserves the right to make changes to information published in this document, including without limitation specifications and product descriptions, at any time and without notice. This document supersedes and replaces all information supplied prior to the publication hereof.

Suitability for use — NXP Semiconductors products are not designed, authorized or warranted to be suitable for use in life support, life-critical or safety-critical systems or equipment, nor in applications where failure or malfunction of an NXP Semiconductors product can reasonably be expected to result in personal injury, death or severe property or environmental damage. NXP Semiconductors and its suppliers accept no liability for inclusion and/or use of NXP Semiconductors products in such equipment or applications and therefore such inclusion and/or use is at the customer's own risk.

Applications — Applications that are described herein for any of these products are for illustrative purposes only. NXP Semiconductors makes no representation or warranty that such applications will be suitable for the specified use without further testing or modification.

Customers are responsible for the design and operation of their applications and products using NXP Semiconductors products, and NXP Semiconductors accepts no liability for any assistance with applications or customer product design. It is customer's sole responsibility to determine whether the NXP Semiconductors product is suitable and fit for the customer's applications and products planned, as well as for the planned application and use of customer's third party customer(s). Customers should provide appropriate design and operating safeguards to minimize the risks associated with their applications and products.

NXP Semiconductors does not accept any liability related to any default, damage, costs or problem which is based on any weakness or default in the customer's applications or products, or the application or use by customer's third party customer(s). Customer is responsible for doing all necessary testing for the customer's applications and products using NXP Semiconductors products in order to avoid a default of the applications and the products or of the application or use by customer's third party customer(s). NXP does not accept any liability in this respect.

Limiting values — Stress above one or more limiting values (as defined in the Absolute Maximum Ratings System of IEC 60134) will cause permanent damage to the device. Limiting values are stress ratings only and (proper) operation of the device at these or any other conditions above those given in the Recommended operating conditions section (if present) or the Characteristics sections of this document is not warranted. Constant or repeated exposure to limiting values will permanently and irreversibly affect the quality and reliability of the device.

Terms and conditions of commercial sale — NXP Semiconductors products are sold subject to the general terms and conditions of commercial sale, as published at https://www.nxp.com/profile/terms, unless otherwise agreed in a valid written individual agreement. In case an individual agreement is concluded only the terms and conditions of the respective agreement shall apply. NXP Semiconductors hereby expressly objects to applying the customer's general terms and conditions with regard to the purchase of NXP Semiconductors products by customer.

No offer to sell or license — Nothing in this document may be interpreted or construed as an offer to sell products that is open for acceptance or the grant, conveyance or implication of any license under any copyrights, patents or other industrial or intellectual property rights.

Quick reference data — The Quick reference data is an extract of the product data given in the Limiting values and Characteristics sections of this document, and as such is not complete, exhaustive or legally binding.

Export control — This document as well as the item(s) described herein may be subject to export control regulations. Export might require a prior authorization from competent authorities.

Suitability for use in non-automotive qualified products — Unless this document expressly states that this specific NXP Semiconductors product is automotive qualified, the product is not suitable for automotive use. It is neither qualified nor tested in accordance with automotive testing or application requirements. NXP Semiconductors accepts no liability for inclusion and/or use of non-automotive qualified products in automotive equipment or applications.

In the event that customer uses the product for design-in and use in automotive applications to automotive specifications and standards, customer (a) shall use the product without NXP Semiconductors' warranty of the product for such automotive applications, use and specifications, and (b) whenever customer uses the product for automotive applications beyond NXP Semiconductors' specifications such use shall be solely at customer's own risk, and (c) customer fully indemnifies NXP Semiconductors for any liability, damages or failed product claims resulting from customer design and use of the product for automotive applications beyond NXP Semiconductors' standard warranty and NXP Semiconductors' product specifications.

HTML publications — An HTML version, if available, of this document is provided as a courtesy. Definitive information is contained in the applicable document in PDF format. If there is a discrepancy between the HTML document and the PDF document, the PDF document has priority.

Translations — A non-English (translated) version of a document, including the legal information in that document, is for reference only. The English version shall prevail in case of any discrepancy between the translated and English versions.

All information provided in this document is subject to legal disclaimers.

© 2026 NXP B.V. All rights reserved.

Security — Customer understands that all NXP products may be subject to unidentified vulnerabilities or may support established security standards or specifications with known limitations. Customer is responsible for the design and operation of its applications and products throughout their lifecycles to reduce the effect of these vulnerabilities on customer's applications and products. Customer's responsibility also extends to other open and/or proprietary technologies supported by NXP products for use in customer's applications. NXP accepts no liability for any vulnerability. Customer should regularly check security updates from NXP and follow up appropriately.

Customer shall select products with security features that best meet rules, regulations, and standards of the intended application and make the ultimate design decisions regarding its products and is solely responsible for compliance with all legal, regulatory, and security related requirements concerning its products, regardless of any information or support that may be provided by NXP.

NXP has a Product Security Incident Response Team (PSIRT) (reachable at PSIRT@nxp.com) that manages the investigation, reporting, and solution release to security vulnerabilities of NXP products.

NXP B.V. — NXP B.V. is not an operating company and it does not distribute or sell products.

Trademarks

Notice: All referenced brands, product names, service names, and trademarks are the property of their respective owners.

NXP — wordmark and logo are trademarks of NXP B.V.

eIQ — is a trademark of NXP B.V.

All information provided in this document is subject to legal disclaimers.

© 2026 NXP B.V. All rights reserved.

Contents

1 Overview 2
2 Features 2
2.1 Models 2
3 Licenses 3
4 Supported devices 3
5 System configuration 3
6 Release contents 3
6.1 Ara2 runtime SDK package contents 4
6.1.1 New features 4
6.2 Ara2 vision example release contents 5
6.2.1 What is New in v1.1.0 (new features and enhancements) 6
6.3 LLM Edge Studio release contents 7
6.3.1 New features and enhancements 8
6.4 VLM Edge Studio release contents 8
6.4.1 New features and enhancements 9
6.5 eIQ AAF connection release contents 9
6.5.1 New features 10
6.6 Smart device gateway release contents 10
6.6.1 Features and enhancements 11
7 Known issues 12
7.1 Ara240 with FRDM i.MX 8M Plus boot failures 12
7.2 Qwen2.5-Coder-1.5B overflow error in eIQ AAF connector 12
7.3 FRDM i.MX 8MP, FRDM i.MX 95, and FRDM i.MX 95 Pro freeze after firmware upgrade 12
7.4 Model load stalls at ~95 % in VLM-Edge-Studio and LLM-Edge-Studio 13
7.5 ara2_metrics script failure on FRDM i.MX 8MP, FRDM i.MX 95, and FRDM i.MX 95 Pro 13
7.6 Silent hugging face model download failures under poor network conditions 13
8 Revision history 14
Legal information 15

Please be aware that important notices concerning this document and the product(s) described herein, have been included in section 'Legal information'.

© 2026 NXP B.V. **All rights reserved.**