Smarter SOS: Generative AI at the Edge for Emergency Communication over Satellite Links

FISHING FOR EXPERIENCE

TUHH

BACKGROUND

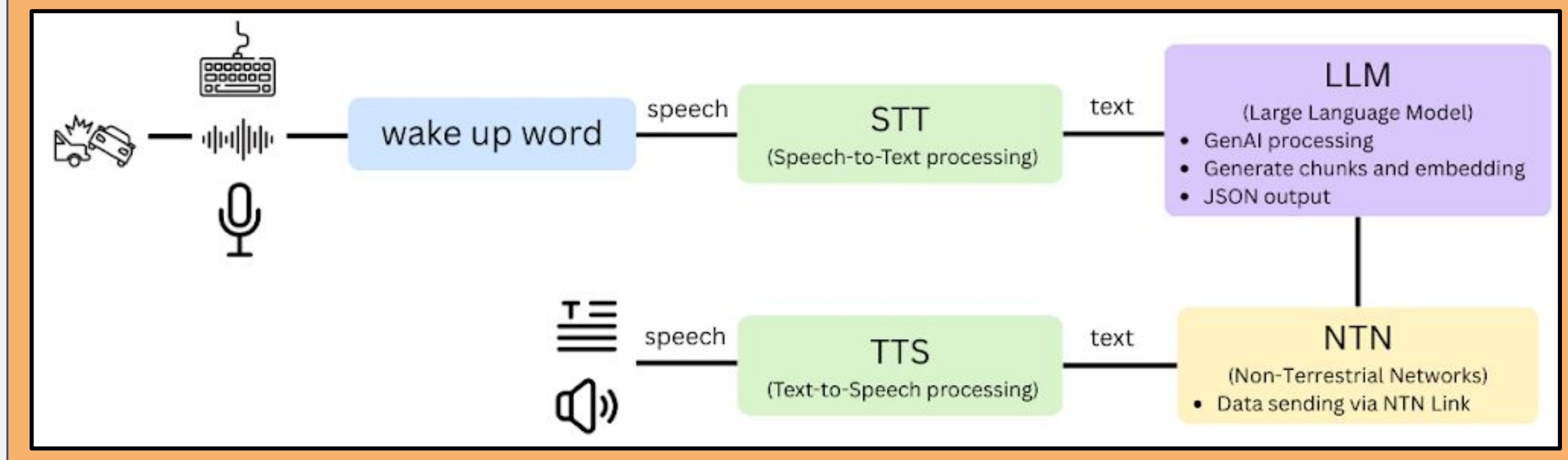
Modern emergency systems, like SOS service on smartphones or BMW eCall demonstrator over satellite, transmit emergency calls via satellite links. However, due to **severe bandwidth limitations of Non-Terrestrial Networks (NTN)**, only minimal data such as GPS coordinates or standardized SOS activation message can be sent. **Rich semantic communication** (speech, contextual details, medical information) is currently **not possible**.

So far, GenAI models have mainly been deployed in cloud, which is impractical in emergencies without stable connectivity. With modern edge hardware platforms (NXP i.MX95), **GenAI at the Edge** becomes possible. **Speech-to-Text (STT)** and **Text-to-Speech (TTS)** convert spoken language to text and back, while **Generative AI models** compress, condense, and reconstruct semantic content, enabling **meaningful communication** even with **minimal bandwidth**.

AIM

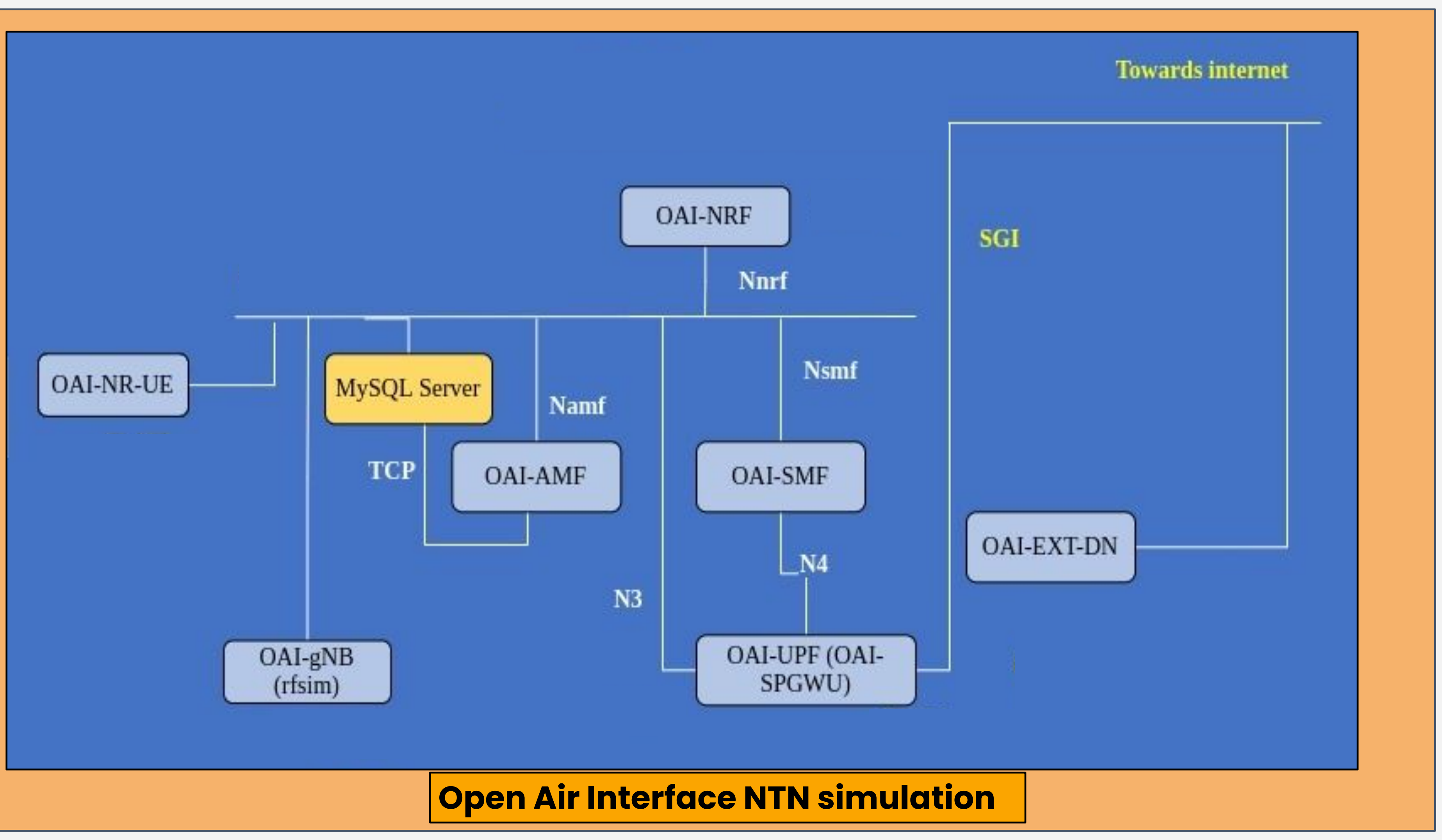
- **Integration of Speech-to-Text (STT) and Text-to-Speech (TTS)** pipelines for transforming spoken language into text and vice versa.
- **Combining GenAI at the Edge** with **Retrieval-Augmented Generation** (RAG).
- **Transmitting data with an NTN simulation** using an open-source 5G stack.
- **Development of a GUI for ground station to display car accident details** – incident alert (beep sound), vehicle information (number plate), location data, and LLM-generated responses on the interface.
- **Demonstration of an end-to-end emergency communication prototype**.

ARCHITECTURES



wake up word
speech
STT
(Speech-to-Text processing)
text
LLM
(Large Language Model)
• GenAI processing
• Generate chunks and embedding
• JSON output
speech
TTS
(Text-to-Speech processing)
text
NTN
(Non-Terrestrial Networks)
• Data sending via NTN Link

eIQ GenAI with RAG



Open Air Interface NTN simulation

HARDWARE & SOFTWARE

Hardware

- NXP i.MX95 board running Danube-500M LLM

Software

- eIQ GenAI Flow framework with STT and TTS pipelines
- Linux (Ubuntu), Python, Docker
- OpenAirInterface (OAI) 5G stack with NTN LEO RF satellite link simulator using Docker Compose

CHALLENGES

Custom Database Generation

Due to limited hardware resources on PC, difficulty in generating RAG database, chunks and embeddings for custom use case.

Integrating OAI 5G NTN stack on i.MX 95

Configuring the OpenAirInterface 5G stack with the NTN LEO RF simulator on the i.MX 95 platform, which is based on the ARM architecture, requires modifications to the Docker containers, as they were originally designed for x86 architectures.

RESULTS

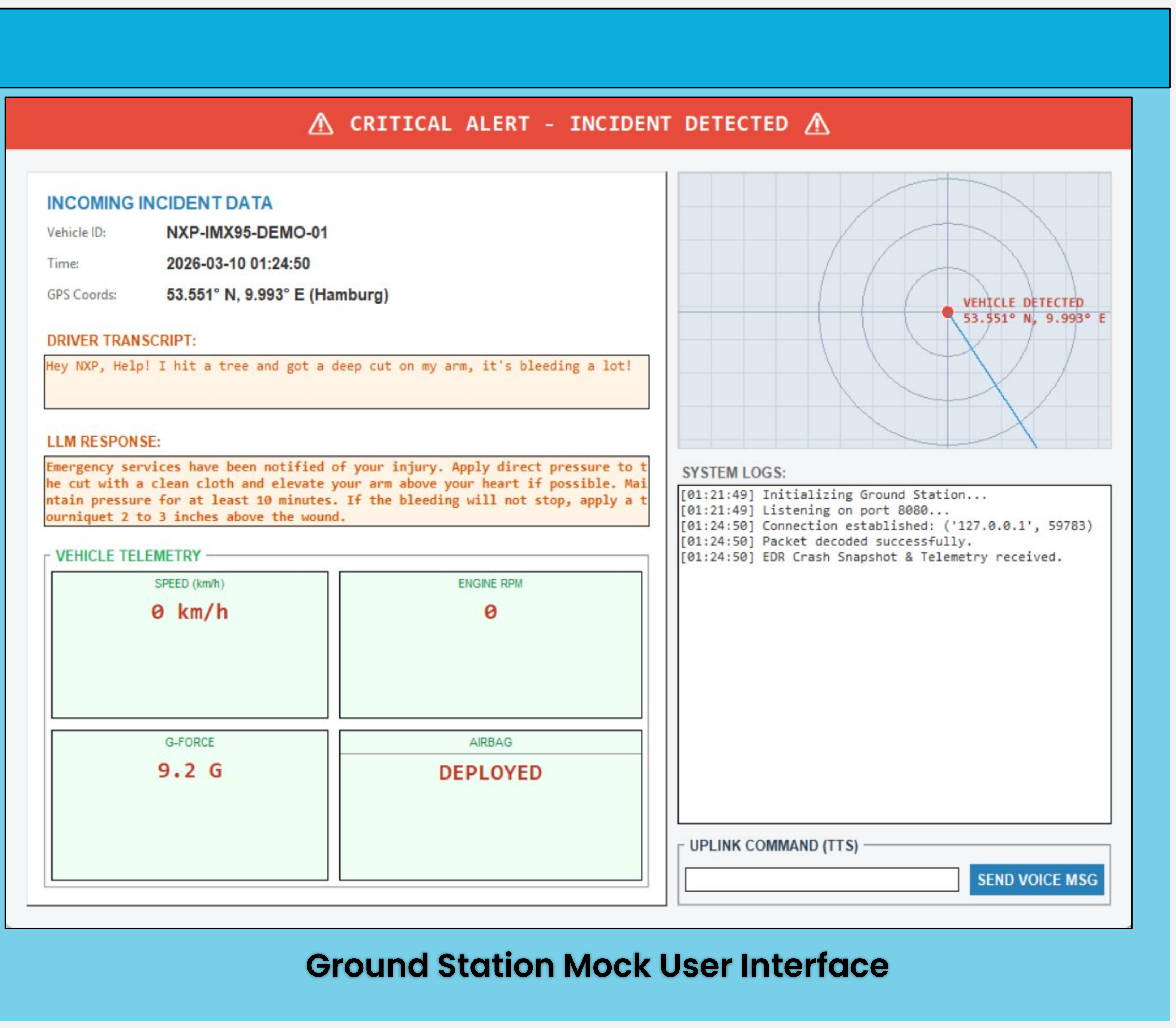
Enhanced LLM's knowledge base by creating new **RAG dataset for car emergency scenarios,** including new information PDF and JSON file with handmade chunks and Q&A pairs.

Generated embeddings using existing **all-MiniLM-L6-v2** and integrated them with **Danube-500M (q4/q8) LLM.** Successfully tested on **i.MX**, and LLM now produces accurate responses using RAG-based retrieval process.

NTN (Non-Terrestrial Network) Simulation (host PC, Docker Compose):

Structured JSON incident payload successfully transmitted from **Edge Device (UE1 — User Equipment)** through **OAI (OpenAirInterface) gNB (5G Base Station) + 5G Core + NTN LEO (Low Earth Orbit) RF Simulator** to **Ground Station (UE2)**.

Ground Station (host PC): Python, Tkinter Command Centre **displays** received LLM response, mocked vehicle telemetry, and Uplink Command module for sending instructions back to the edge device.



⚠ CRITICAL ALERT - INCIDENT DETECTED ⚠

INCOMING INCIDENT DATA

Vehicle ID: NXP-IMX95-DEMO-01

Time: 2026-03-10 01:24:50

GPS Coords: 53.551° N, 9.993° E (Hamburg)

DRIVER TRANSCRIPT:

Hey NXP, Help! I hit a tree and got a deep cut on my arm, it's bleeding a lot!

LLM RESPONSE:

Emergency services have been notified of your injury. Apply direct pressure to the cut with a clean cloth and elevate your arm above your heart if possible. Maintain pressure for at least 10 minutes. If the bleeding will not stop, apply a tourniquet 2 to 3 inches above the wound.

VEHICLE TELEMETRY

SPEED (km/h): 0 km/h

ENGINE RPM: 0

G-FORCE: 9.2 G

AIRBAG: DEPLOYED

SYSTEM LOGS:

```
[01:21:49] Initializing Ground Station...
[01:21:49] Listening on port 8080...
[01:24:50] Connection established: ('127.0.0.1', 59783)
[01:24:50] Packet decoded successfully.
[01:24:50] EDR Crash Snapshot & Telemetry received.
```

Ground Station Mock User Interface

FUTURE SCOPE

- **Deploy OAI UE directly on NXP i.MX95:** Enable true end-to-end edge communication - removing host PC dependency, validating direct NTN transmission between i.MX95 and Ground Station.
- **Expand RAG Knowledge Base:** Broaden document set and refine retrieval to improve LLM accuracy across diverse emergency scenarios.
- **Full LLM response without truncation:** Increase token output limit to prevent incomplete answers (currently truncated beyond word limit).
- **Multilingual support:** Language detection and multi-language response generation.

INDIVIDUAL WORK OVERVIEW

Rohit Varma Jampana	Product Owner	90 h
Gauri Gajanan Amin	Scrum Master, LLM & RAG	92 h
Parvesh Ramesh	NTN	92 h
Yi-Hsuan Wu	LLM & RAG	90 h
Sreelakshmi Ramachandran	Integration & GUI	90 h
Berkin Sahin	Integration & GUI	90 h