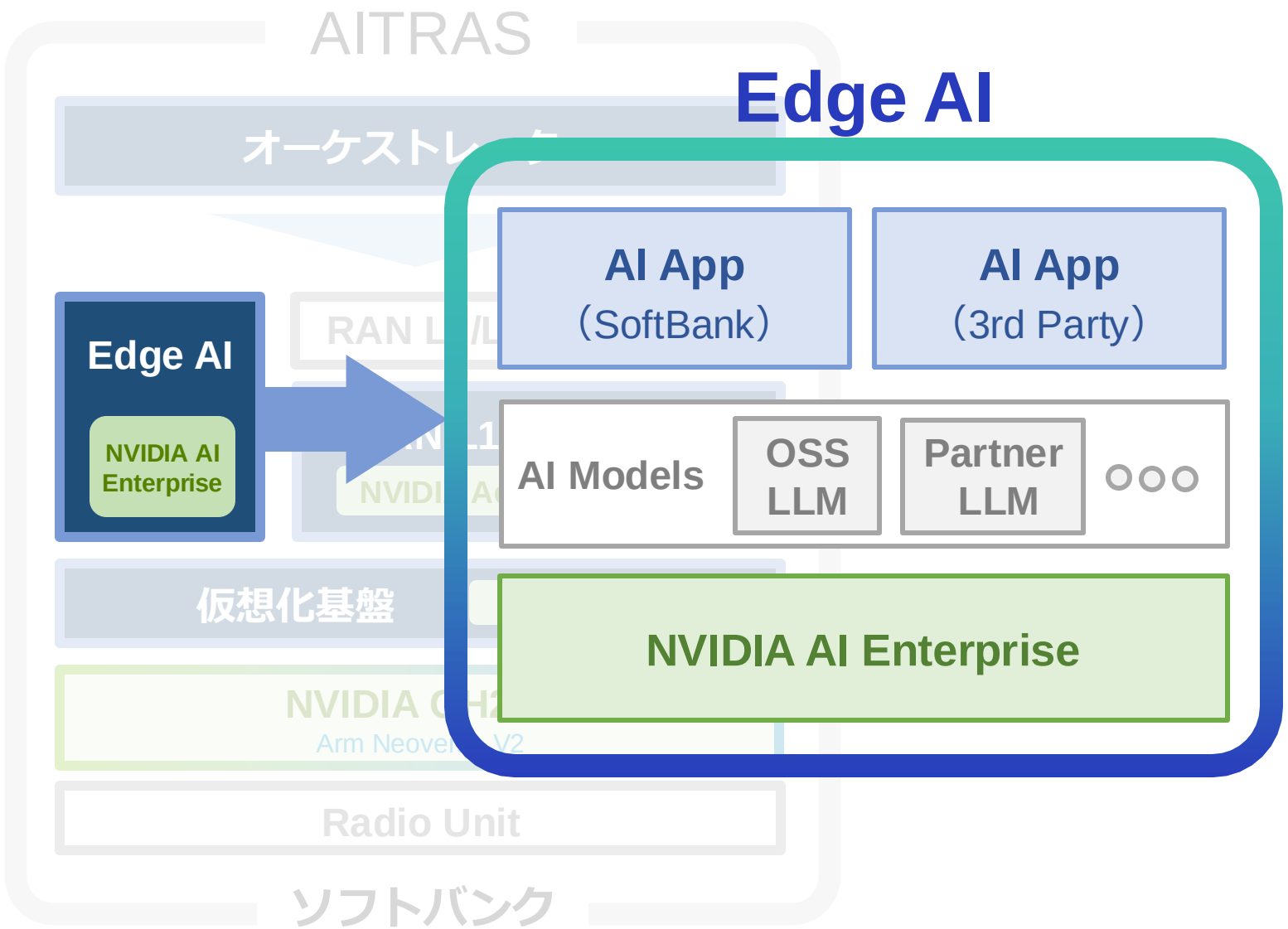
SoftBank

RAG Menu @Edge

Key Technology of AITRAS：AI Edge



Original AI Service

- Autonomous Driving Monitoring App
- Cloud Robot
- **RAG Menu @edge**

NVIDIA Serverless API

Allocates unused resources to meet customer demands on demand

NVIDIA AI Enterprise

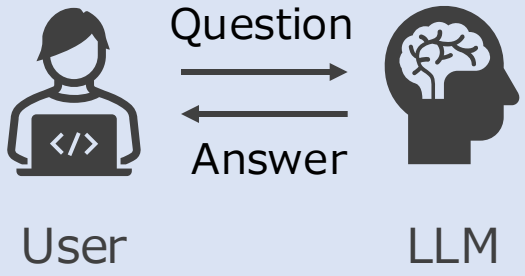
Industry-standard AI framework

What is RAG?
(Retrieval-Augmented Generation)

SoftBank

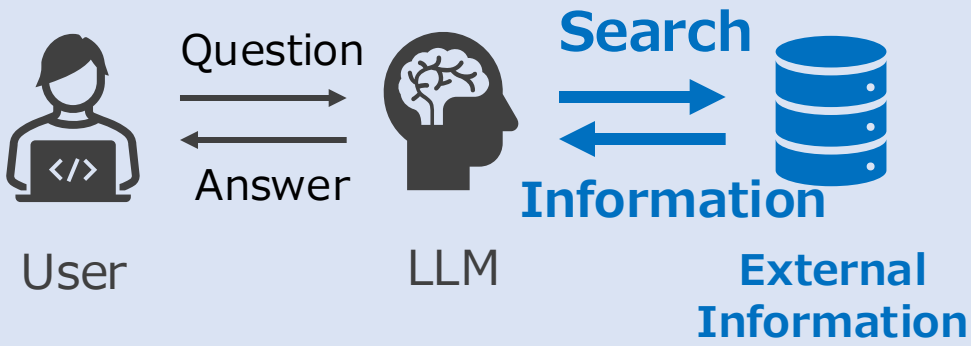
Allows answering information that LLMs do not know, such as internal confidential information

✓ General LLM



Answers based only
on pre-trained information

✓ RAG



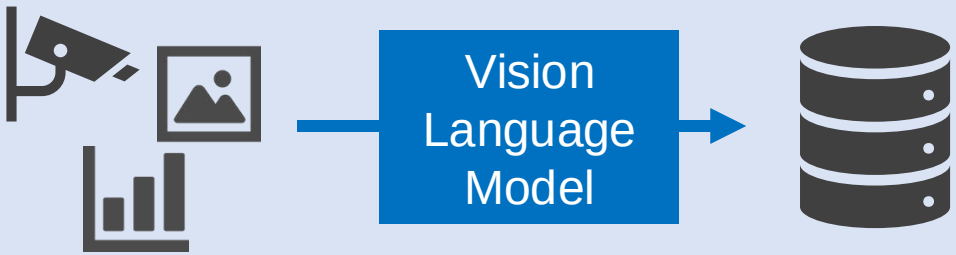
Answers based on
pre-trained information
+ External Information

Multimodal RAG

SoftBank

Handling various data such as images and audio with RAG

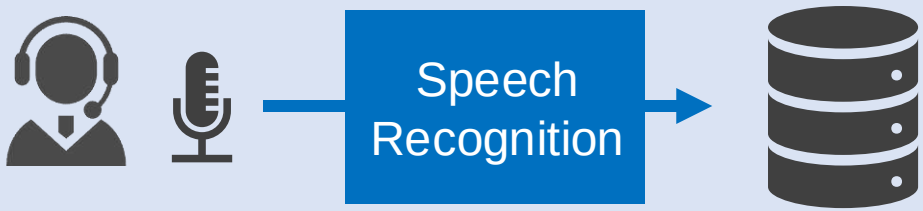
Visual Information



Text Conversion

Example: Diagrams and graphs

Audio Information



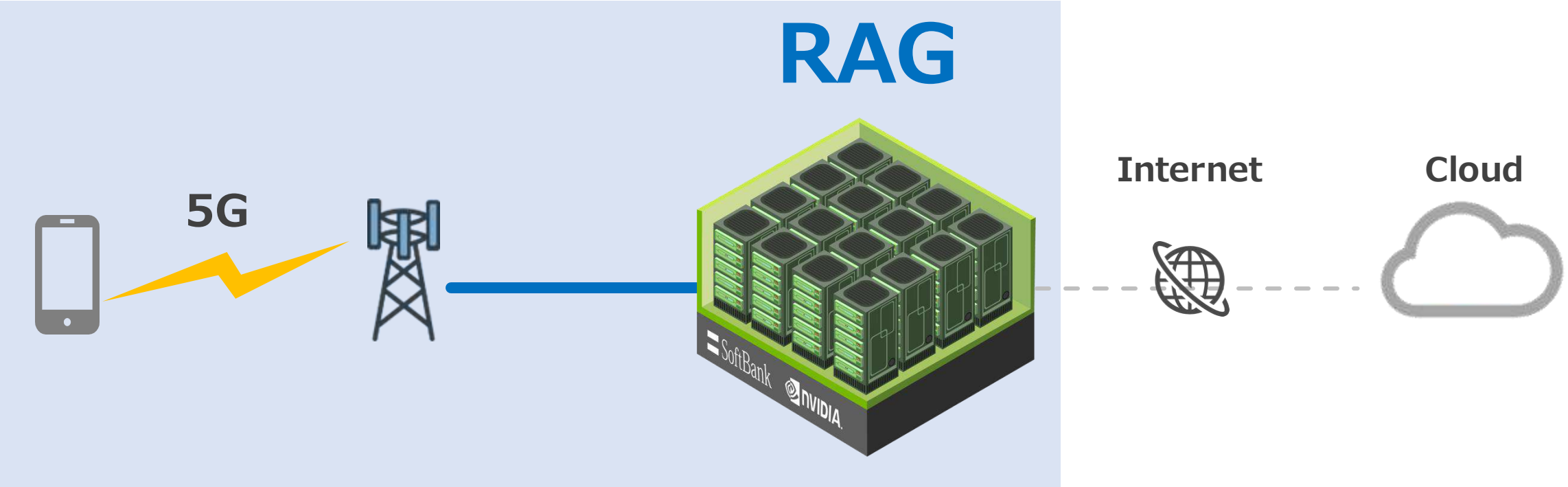
Text Conversion

Example: Meetings, Call Centers

RAG Deployment at the Edge

SoftBank

RAN is a closed network
→Effective for ensuring the security of RAG using confidential information



RAG Menu @Edge

SoftBank

Additional Options to Meet Customer Needs

Base Application

RAG

+

Optional Features

Visual Data

Expanded input for images and video

Voice Stream

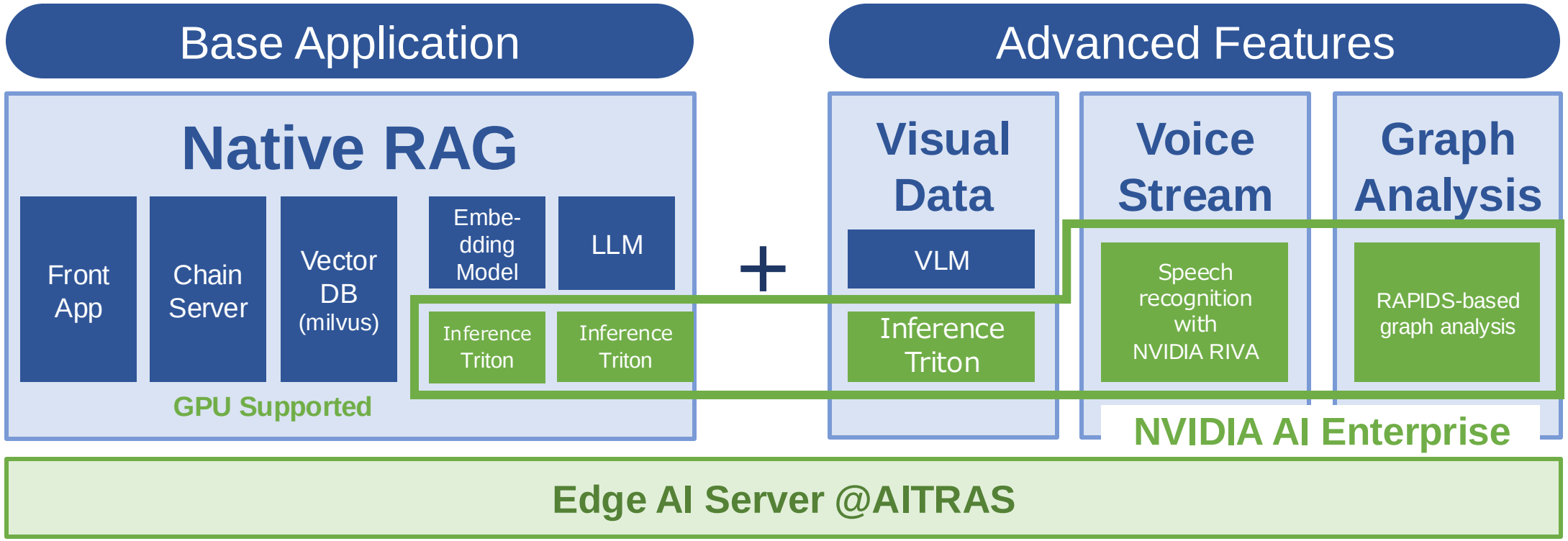
Expanded input for audio

Graph Analysis

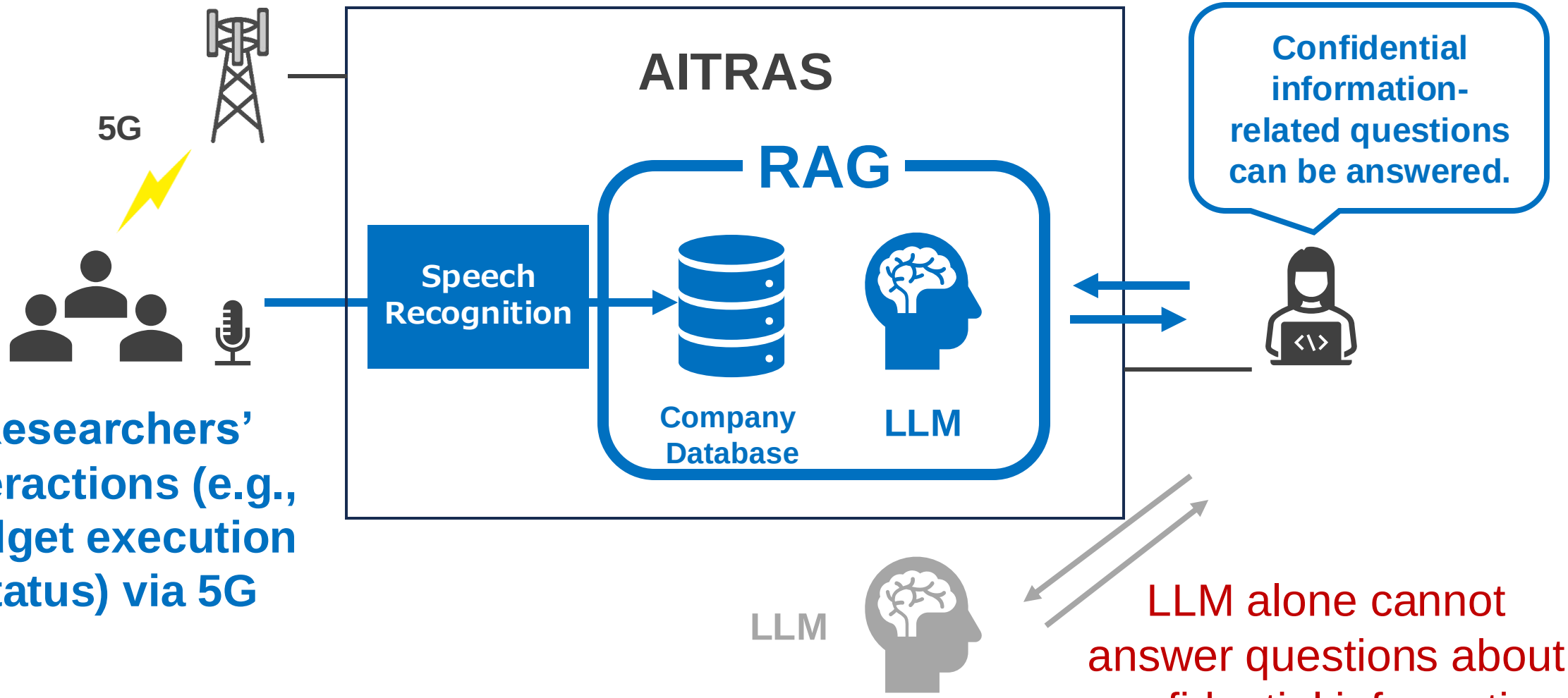
Improved response accuracy

Provided Features

Delivering optimal GPU-enables feature sets implemented with NVIDIA AI Enterprise



Demo : Research Institute of Advanced Technology RAG



Example：Enterprise Scenario
Improving Factory Operations Efficiency with RAG