GenAI Large Telecom Model: The Future of Mobile Network Operations

May 12, 2025 11:00 am ET / May 13, 2025 0:00 am JT

FIERCE Network

SoftBank

AvidThink

Today’s Speakers

Rajeev Koodli
Principal Fellow,
Research Institute of
Advanced Technology,
SoftBank Corp.

Shun Tamura
AI Engineer,
Research Institute of
Advanced Technology,
SoftBank Corp.

Speaker

Rajeev Koodli

Principal Fellow

Research Institute of Advanced Technology

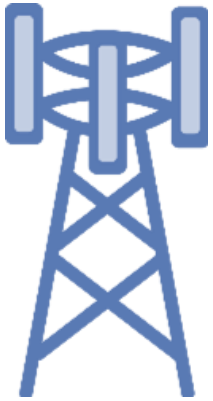
SoftBank Corp.

Technical Overview of LTM

Introduction

AI-RAN and LTM

AI-RAN



Machine AI

AI for RAN

Better gNB performance

CAPEX Reduction

LTM

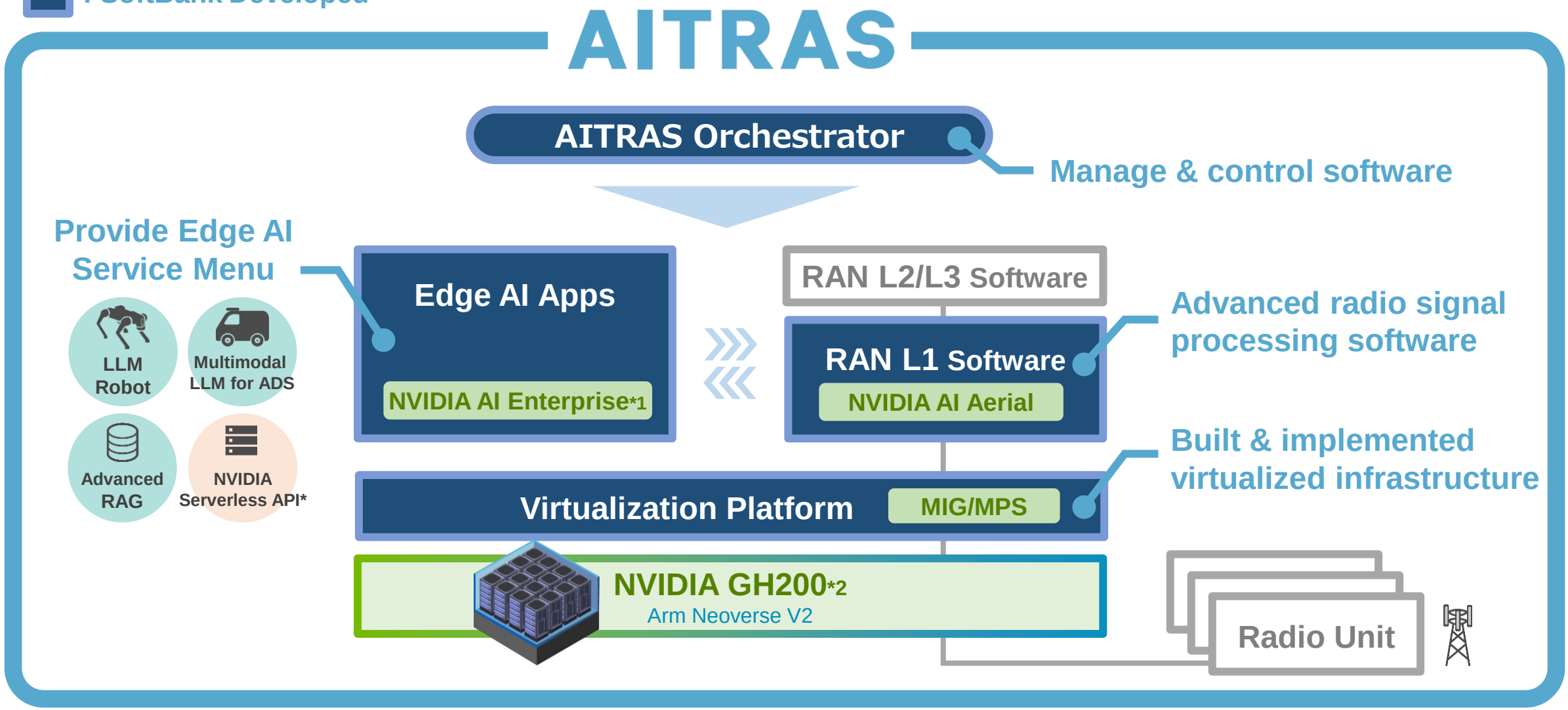
Human AI

Better Operation by AI

OPEX Reduction

SoftBank AI-RAN: AITRAS

: SoftBank Developed

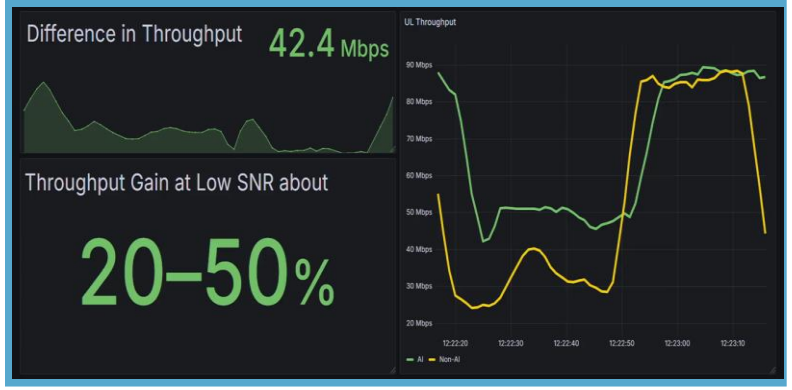


*1:Serverless API powered by NVIDIA AI Enterprise *2: NVIDIA GH200 Grace Hopper Superchip

SoftBank's AI-for-RAN

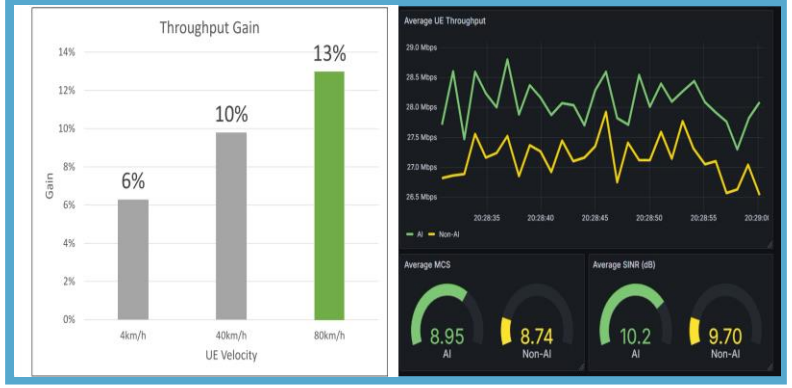
3 Use Cases Introduced at MWC25

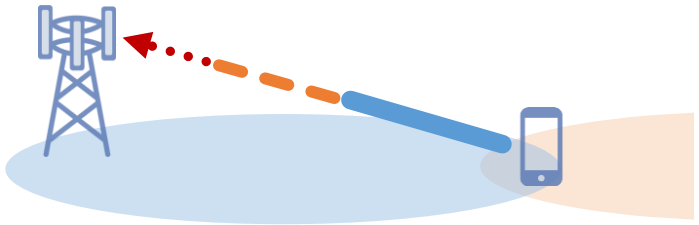
1 UL Channel Interpolation

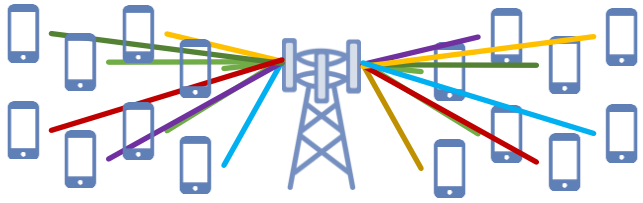


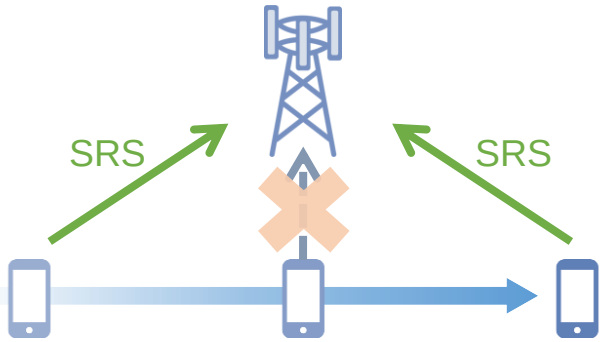
2 MU-MIMO Scheduler

3 SRS Prediction

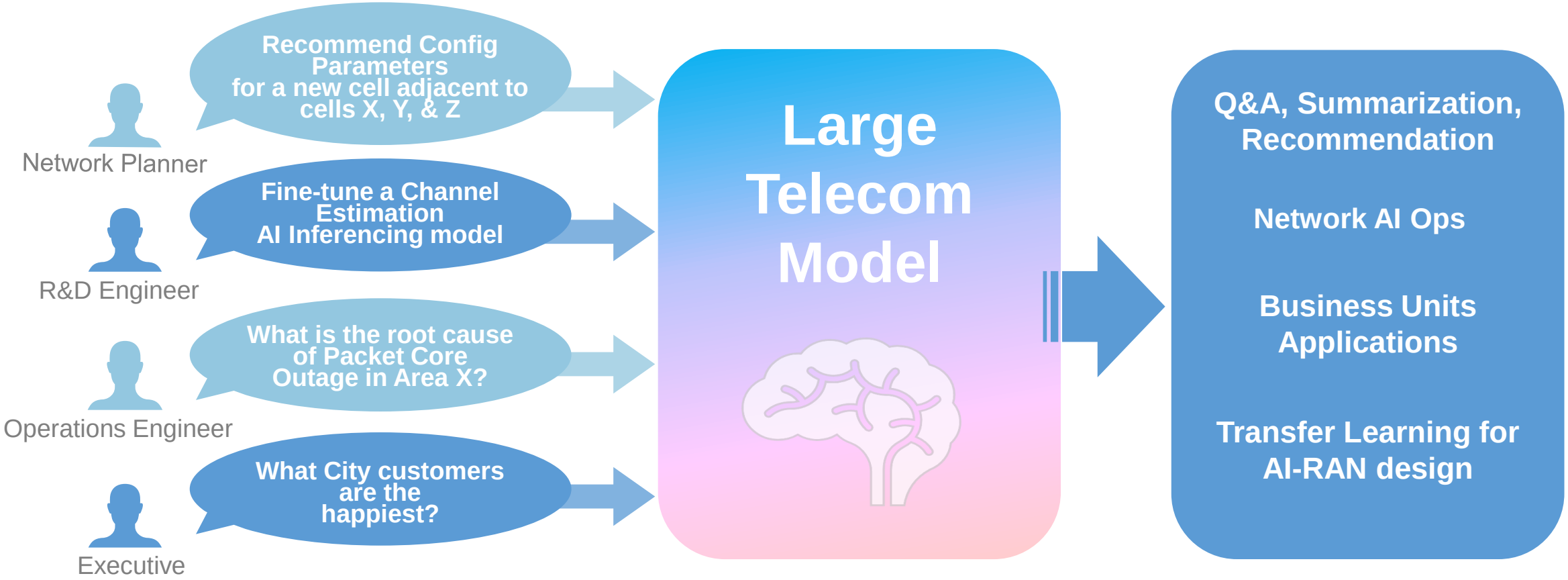








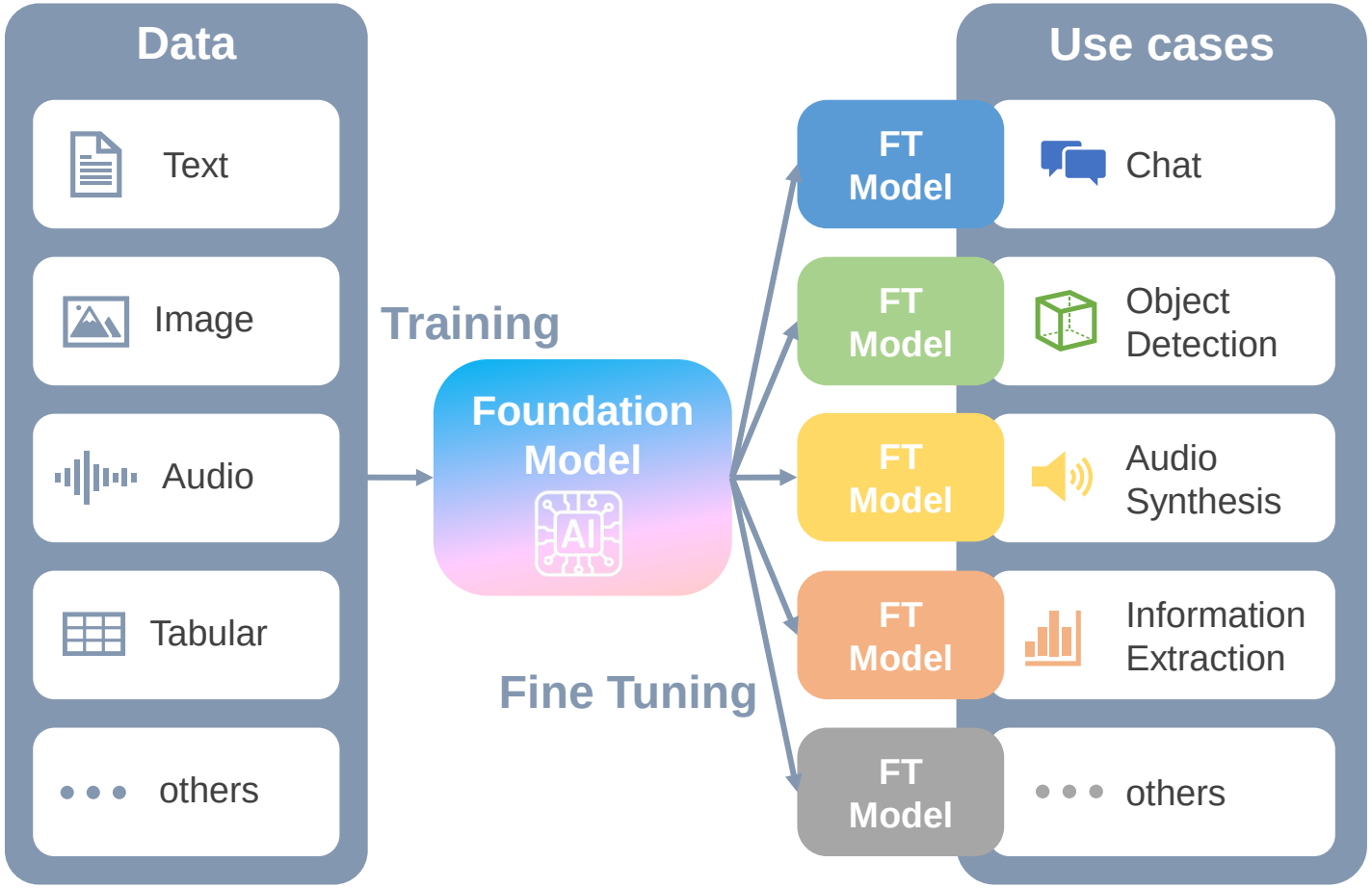
Human AI: Objectives



Democratize Operational Knowledge across the Entire Company using Natural Language Interaction

What is LTM ?

Training a single foundation model on diverse data
and fine-tuning it for each specific use case



Advantages

General Expertise

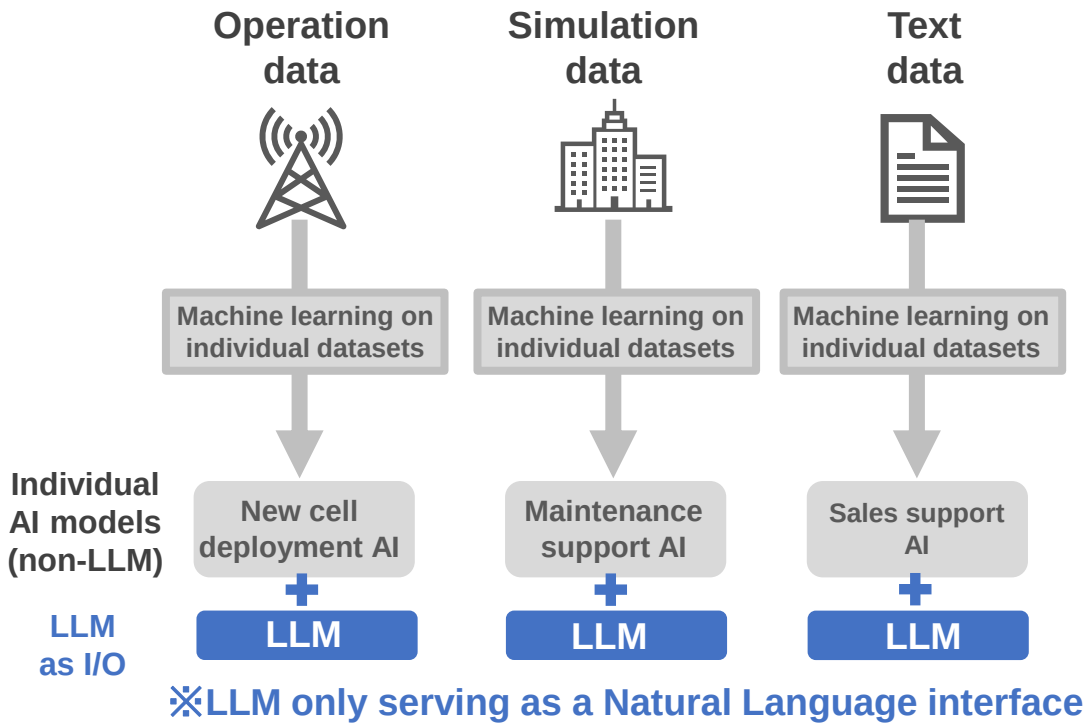
Transfer Learning

Interdisciplinary capability

LTM is not an NLT interface to existing Network OPs

Differences With Conventional Methods

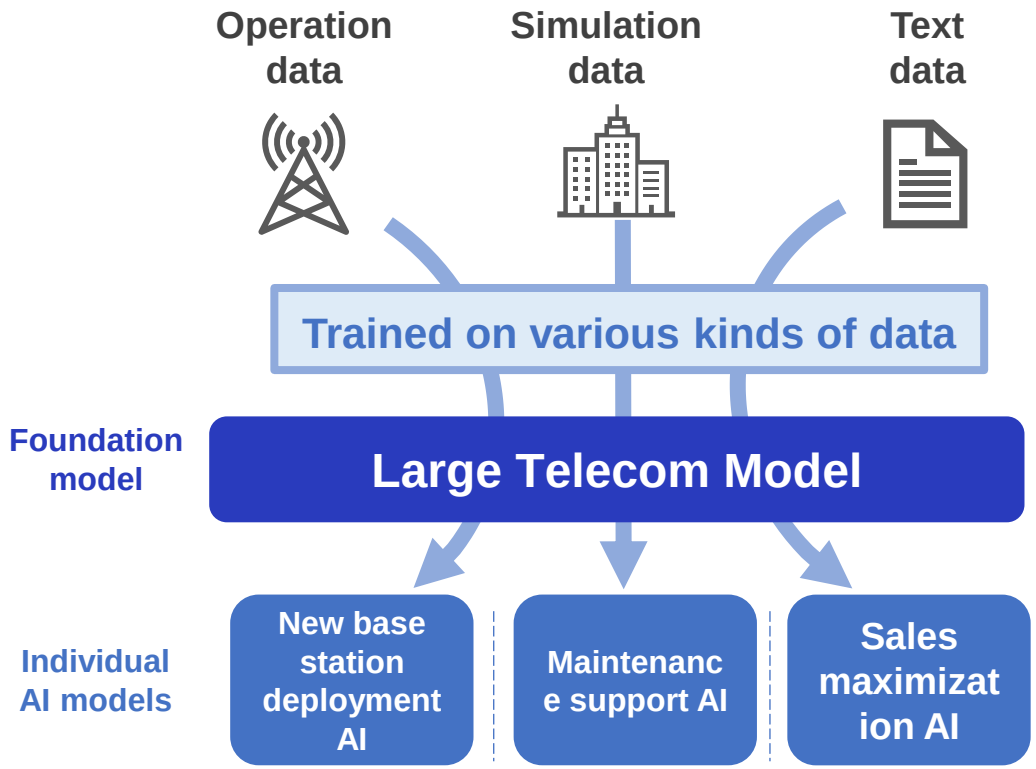
Conventional Methods



※LLM only serving as a Natural Language interface

Comparable to a smart, but inexperienced worker

Large Telecom Model



Comparable to an expert who has learned and experienced everything

Journey

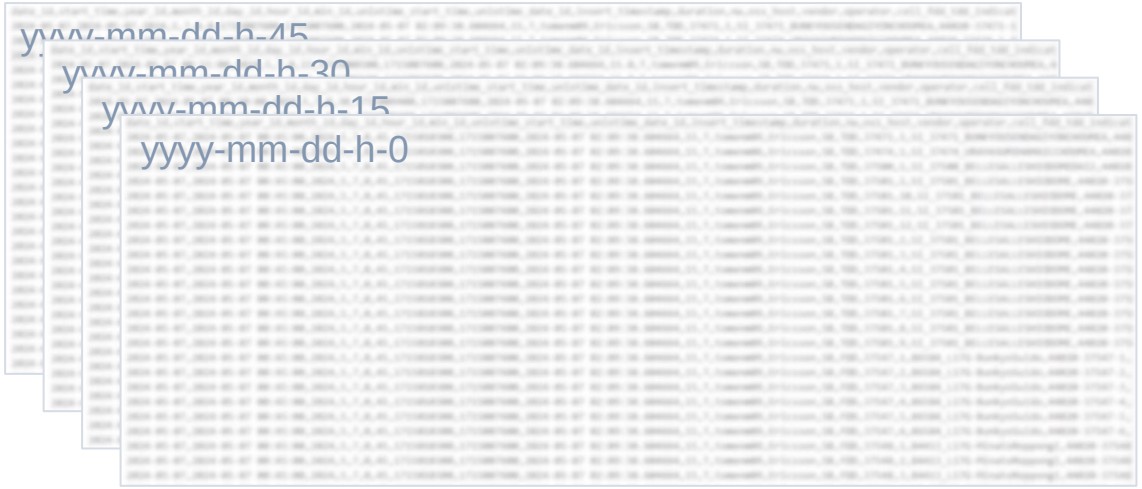
Context

- Started in summer 2024
- Single Modal models
- Vast amount of heterogeneous network operations data (disparate sources, private to SoftBank)
- Need for large enough infra for data processing and training loops

Data: Total of 4TB of SoftBank-Specific Data Corpus

	15min	hourly	daily
RAN KPI data	2024/05/07-2024/08/12 (1840GB ≈ 20GB a day)	2023/08/01-2024/08/12 (1748GB)	-
Core KPI data	2024/05/05-2024/08/12 (27GB)	2023/08/01-2024/08/12 (28GB)	-
Config data	-	-	2024/07/16-2024/08/12 (573GB *zipped ≈ 500GB a day)

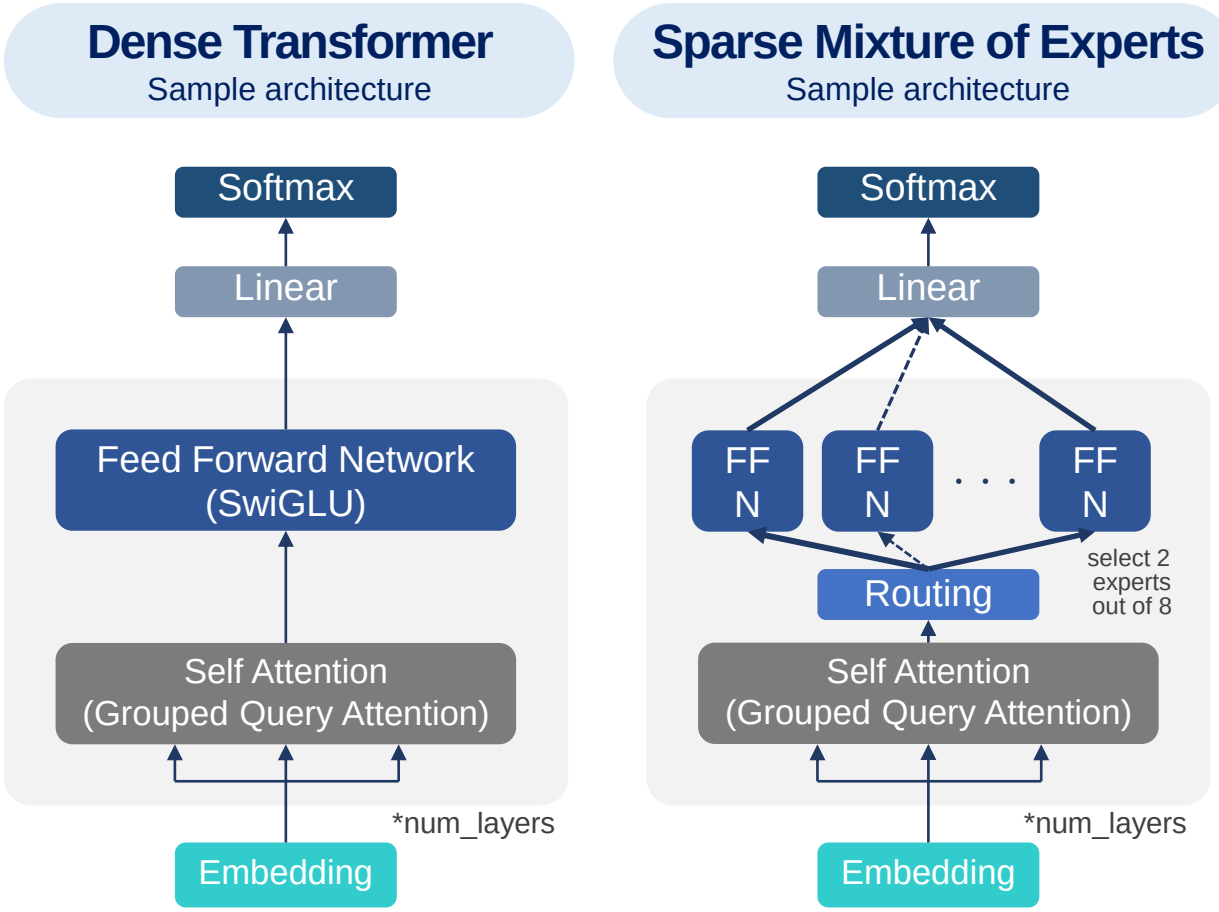
KPI data sample (every 15min)



Config data sample (every 15min)

Model

- Dense Transformer and Sparse Mixture of Experts (SMoE) are popular architectures
- SMoE allows for faster inference because of fewer active parameters
- In version 1, we were data processing & training-bound. SMoE is also harder to debug
- Therefore, we first focus on a pre-trained model based on the dense transformer



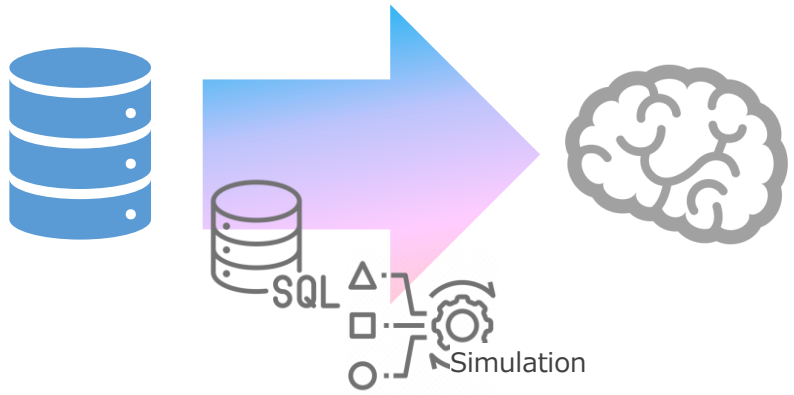
*Positional encoding, Residual network, Normalization, etc. omitted

Challenges with Infrastructure

Data Processing

Huge raw data

LLM training ready data



Significant compute capacity needed to process the raw data

Training

Fine tuning

Full training

Computation resource required
(Estimated, Unit: GPU Hour = 1 Hour spent on 1 DGX A100)
Based on our approach with LoRA fine tuning

Data size	1k cells	200k cells (All 5G cells)
7 days of data	42 GPU Hours	350x24 GPU Hours
365 days	91x24 GPU Hours	50x365x24 GPU Hours

GPU infrastructure is assumed for any kind of AI training

In LLM full training, Data Center level GPU infrastructure is needed

Options for Human AI (and Machine AI)

SoftBank

In-context based | **Training based**

Prompting

Cost: Lowest

✅ Flexibility

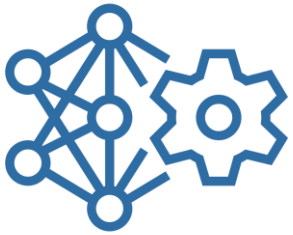
❌ Limited knowledge added

RAG

✅ Scalability

❌ Dependency on retrieval quality

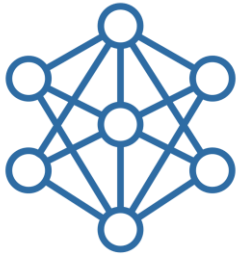
Fine tuning



✅ Task/Domain adaptability

❌ Limited customizability

Full training



Cost: Highest

✅ Full Customizability

❌ Difficulty in updating

Human AI

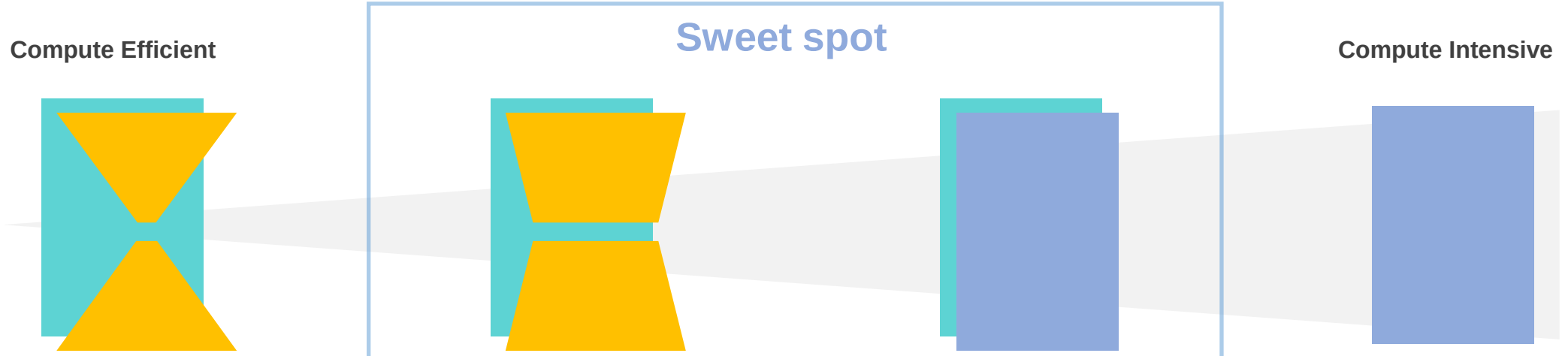
Machine AI

Depending on the model size, data volume and available infra, the best option can be determined for each use case

In-context approaches can boost the base or fine-tuned models' performance (without updating model weights)

Training paradigm can first start with fine tuning and gradually move onto Full training

Supervised Fine Tuning

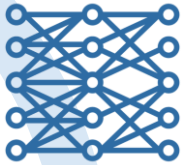


End-End Pipeline

Data collection → **Data processing** → **Training & Evaluation** → **Containerization** → **Serving as an API**

- Collect necessary data without loss
- Design for different data granularity
- Confidentiality, integrity, availability
- Smooth export to machine learning cluster

- Import into databases
- SQL queries, ETL
- Data-grounded synthetic data generation with/without LLM
- Distributed processing on a data lake



- Supervised learning/reinforcement learning
- Distributed learning
- Job scheduling
- Hardware/job monitoring
- Evaluation using public/private datasets

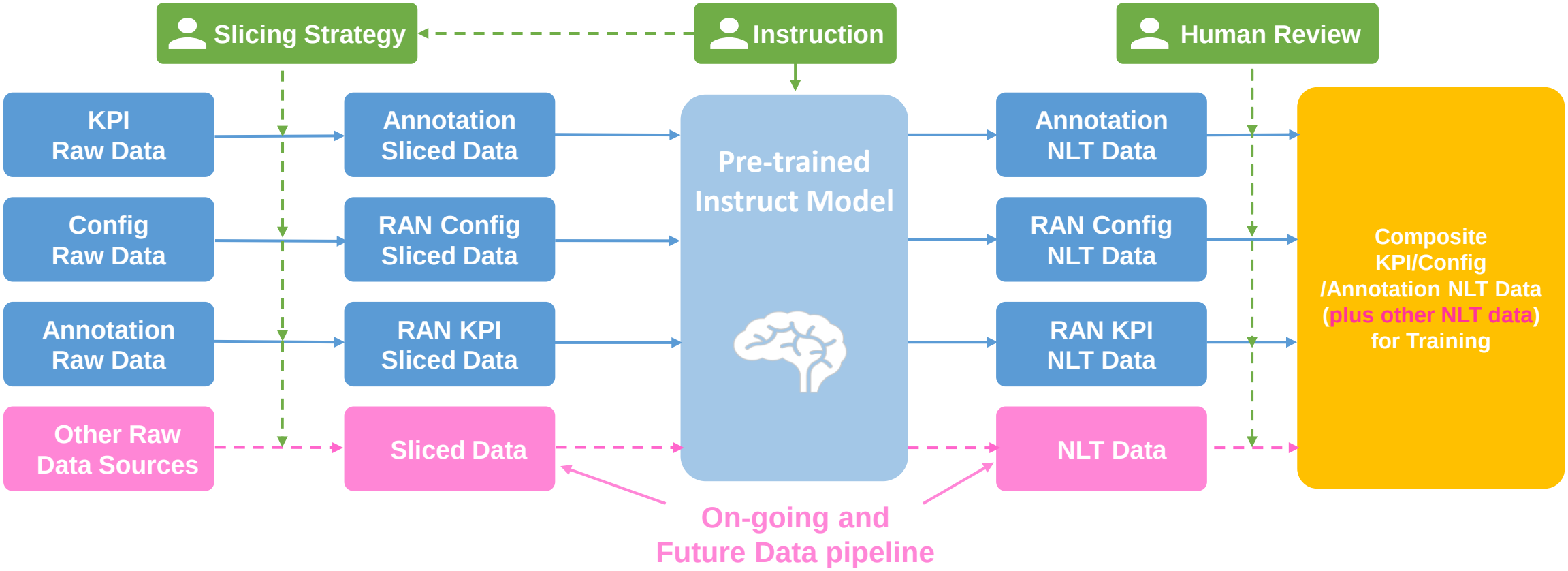
- Model format conversion (ONNX, etc.)
- Integration with inference server
- Continuous updates of environment
- Security scanning

- Deployment on cloud infrastructure
- Auto-scaling
- Observability
- Performance optimization

Loop

- Continuous updates with latest data
- Continuous performance improvement

Data Pipeline



Expert Annotation for Learning

cellReselection Priority

Vanilla Description

Absolute priority of carrier frequency used by cell reselection procedure. Corresponds to parameter priority in 3GPP TS 36.304, Evolved Universal Terrestrial Radio Access (E-UTRA); User Equipment (UE) procedures in idle mode.

Value 0 means lowest priority. UE behavior when no value is entered is specified in 3GPP TS 36.304. Value -1 means frequency is excluded from SIB5 and IdleModeMobilityControlInfo.

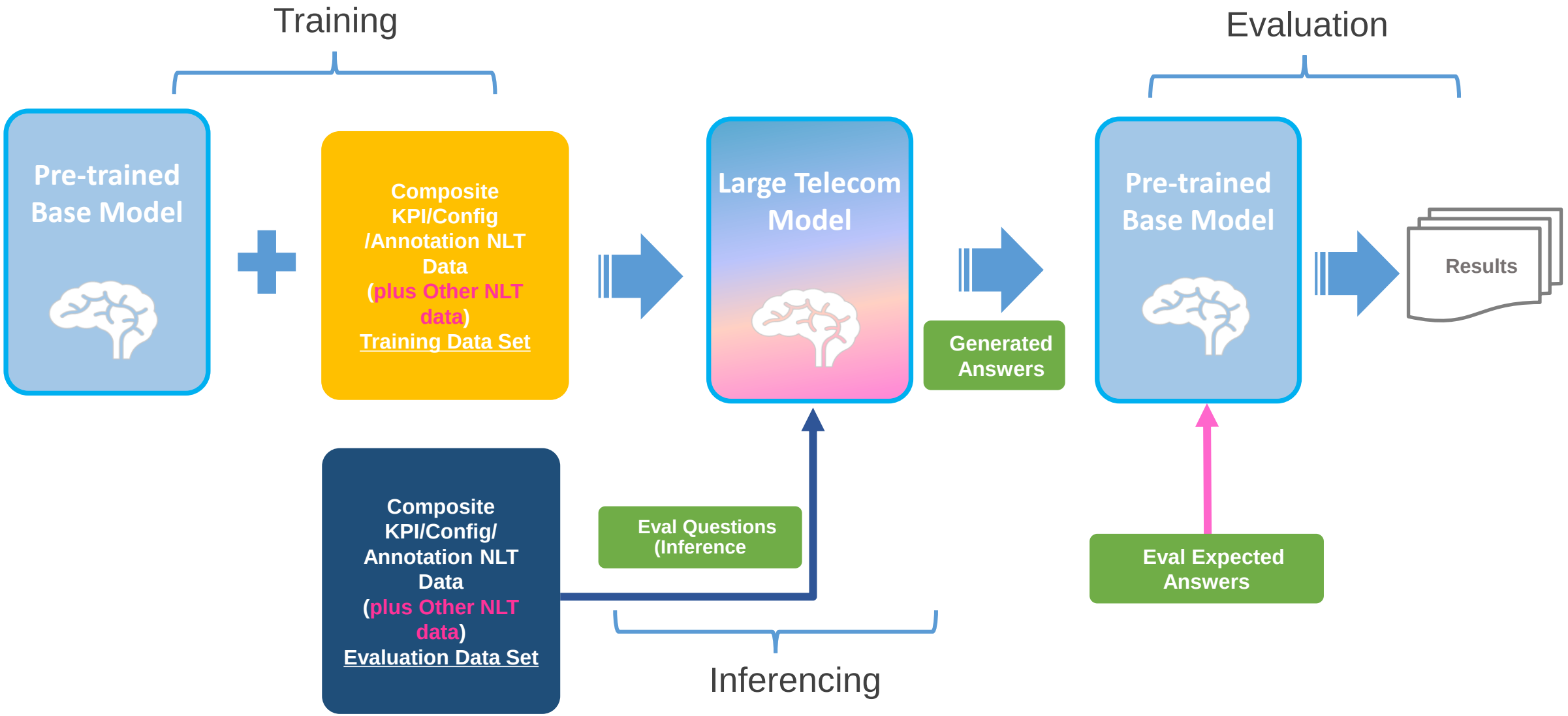
Expert Annotation

Parameters used when reselection to another cell during RRC Idle. You can set the reselectionPriority parameter for each transition destination LTE frequency, notified in SIB5.

The relative evaluation of the priority of the own cell notified by SIB1 and the priority of the frequency notified by SIB5 is separated into HigherPriority, EqualPriority, and LowPriority, and the Idle transition operation is performed in accordance with the 3GPP regulations.

LTM training

Training & Evaluation Pipeline



Current Model Capabilities

The current version 1 of the model is trained to perform natural language processing tasks (below). It is also finetuned for business applications (such as base station config management)

- Open-ended Q&A
- Trend analysis
- Relation extraction
- Summarization
- Recommendation

General capabilities: What Do They Look Like

SoftBank

RAN KPI

```
"question": "Describe the time-based trend of `nr_setup_att` values."
"expected": "Increasing in morning, peaking in afternoon, decreasing at night"
"generated": "Values tend to increase from midnight to afternoon and decrease from evening to night."
```

Relationship between RAN KPI parameters

```
"question": "Identify the dependency pattern between `dl_active_user_max` and `cu_max` values."
"expected": "The values tend to move together, but with varying intensities."
"generated": "The values tend to be dependent, with changes in one often affecting the other."
```

RAN Config

```
"question": "Description of the typical configuration of the `isHoAllowed` parameter within the network configuration."
"expected": "The `isHoAllowed` parameter is typically set to true."
"generated": "The typical configuration of the `isHoAllowed` parameter is true."
```

Relationship between RAN KPI and RAN Config

```
"question": "Are `nr_setup_sr` and `GNBDUFunction/Rrc/t304` values dependent on each other during midday?"
"expected": "No, `nr_setup_sr` and `GNBDUFunction/Rrc/t304` values do not show dependency on each other during midday."
"generated": "No, `nr_setup_sr` and `GNBDUFunction/Rrc/t304` values do not appear to be dependent on each other during midday."
```

Annotation

```
"question": "What does the `cell_state` parameter represent in the network configuration?"
"expected": "The `cell_state` parameter indicates whether a cell is locked or not."
"generated": "The `cell_state` parameter represents the cell's current state, with possible values including locked and unlocked."
```

Summarization of specific cells (Urban, 1.7GHz, etc.)

```
"question": "What is the most common `frequency_band` among DenseUrban cells?"
"expected": "Most common `frequency_band` among DenseUrban cells is 1.7GHz."
"generated": "1.7GHz"
```

Lessons Learned

SoftBank

Technical

- Slicing the different data types (KPI, Config, Annotation) into chunks is crucial for training. Diverse data is important for general capabilities
- Organizing different data types into Model context length can be challenging
- Both data and model hyper-parameter tuning does take time
- Model evaluation is tricky considering the reliance on an evaluation model for judgement. Semantic ambiguity is to be expected in answers and may need human inspection.
- Tabular time-series data handling can be improved with dedicated encoders

Others

- Data is seldom in one place or org. Access times vary
- Setting up internal infrastructure for securely handling data requires proper governance (aka time-consuming, rightfully)
- Finding the right mix of domain and AI expertise and getting them to work together is fun!


SoftBank

Speaker

Shun Tamura

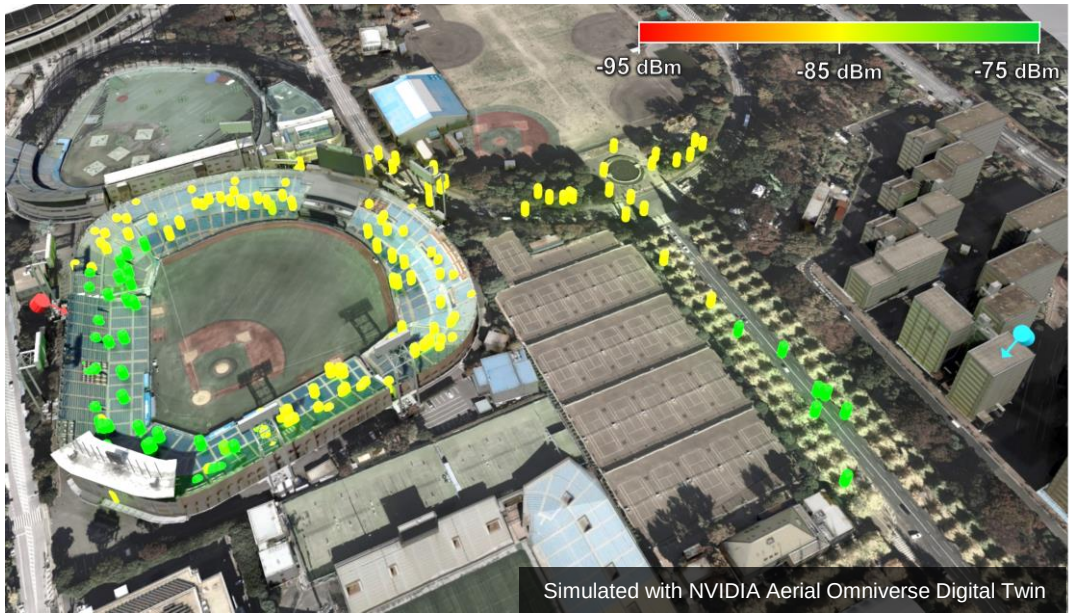
AI Engineer

Research Institute of Advanced Technology

SoftBank Corp.

MNO Use Cases

Specific UC Example: RAN Config Optimization



Scenario

- Handling congestions during events

What LTM does

- Optimize RAN config parameters of base stations within targeted area

Input

- Order for congestion handling (Intent)
- RAN configs of neighboring BSs
- RAN performance (KPI) of neighboring BSs

Fine Tuned LTM

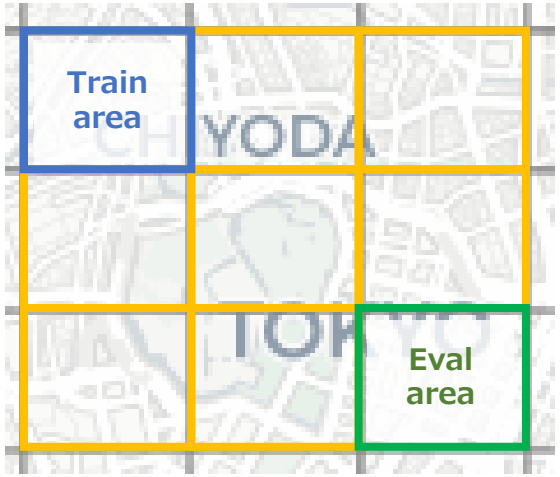
Output

- RAN config parameters of BSs that need to be updated

Fine-Tuning Methodology for Specific Use Cases

SoftBank

Divide whole Japan into training areas and evaluation areas



I want to optimize RAN configs of this base station in this area.

Current RAN configs:

Cell ID	RAN Config 1	RAN Config 2
AAA	True	0
BBB	False	1

Current RAN KPIs:

Cell ID	RAN KPI1	RAN KPI2
AAA	100	10
BBB	50	20

Optimized

Prediction

You can update RAN configs of the base station as follows:

Updated RAN configs:

Cell ID	RAN Config 1	RAN Config 2
AAA	True	0
BBB	**True**	1

Expected RAN KPIs after update:

Cell ID	RAN KPI1	RAN KPI2
AAA	100	10
BBB	**80**	20

FT LTM

Number of synthetic optimization scenarios based on SoftBank's actual data

Evaluation Results

1. Easier

User specifies the parameter to update, model only infers its value

I want to optimize RAN configs of this base station "BBB" in this area. **What would be the best config value for the parameter "RAN config 1"?**

Current RAN configs:

Cell ID	RAN Config 1	RAN Config 2
AAA	True	0
BBB	False	1

Current RAN KPIs: **Specify which parameter to update**

Cell ID	RAN KPI1	RAN KPI2
AAA	100	10
BBB	50	20

99% of value accuracy

2. More difficult

Model infers both parameter to update and its value

I want to optimize RAN configs of this base station "BBB" in this area. **What would be the best config parameter to update and its value?**

Current RAN configs:

Cell ID	RAN Config 1	RAN Config 2
AAA	True	0
BBB	False	1

Current RAN KPIs: **Don't specify which parameter to update**

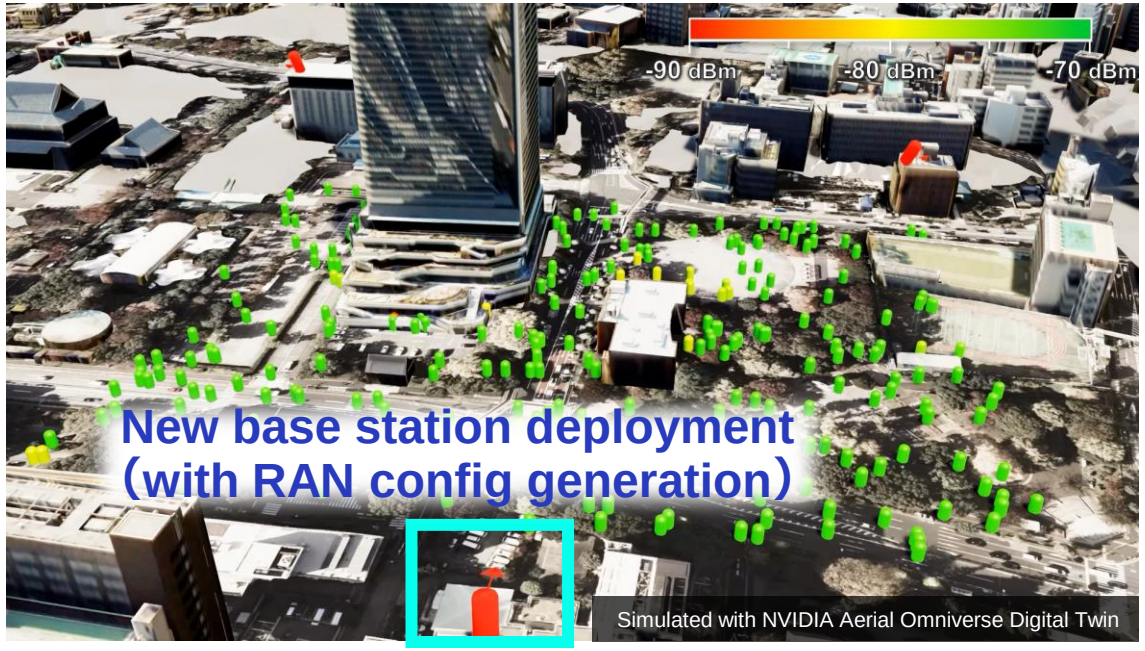
Cell ID	RAN KPI1	RAN KPI2
AAA	100	10
BBB	50	20

94% of param selection + value accuracy

LTM works even in more difficult settings w/ param identification

SoftBank

Advanced: Full RAN Config Generation



Scenario

- Countermeasures for weak signal areas caused by new construction

What LTM does

- Suggests location and RAN configs of the new base station

Input

- Order for weak signal area countermeasures
- RAN configs
- RAN KPIs

FT LTM

Output

- Full RAN configs for the new BS

Evaluation Results: Full Generation

Model infers the value of all the parameters

I want to generate RAN configs of a new base station "CCC" in this area. **What would be the best config value for each parameter?**

Current RAN configs:

Cell ID	RAN Config 1	RAN Config 2
AAA	True	0
BBB	False	1
CCC	???	???

Current RAN KPIs:

Generate all parameter values

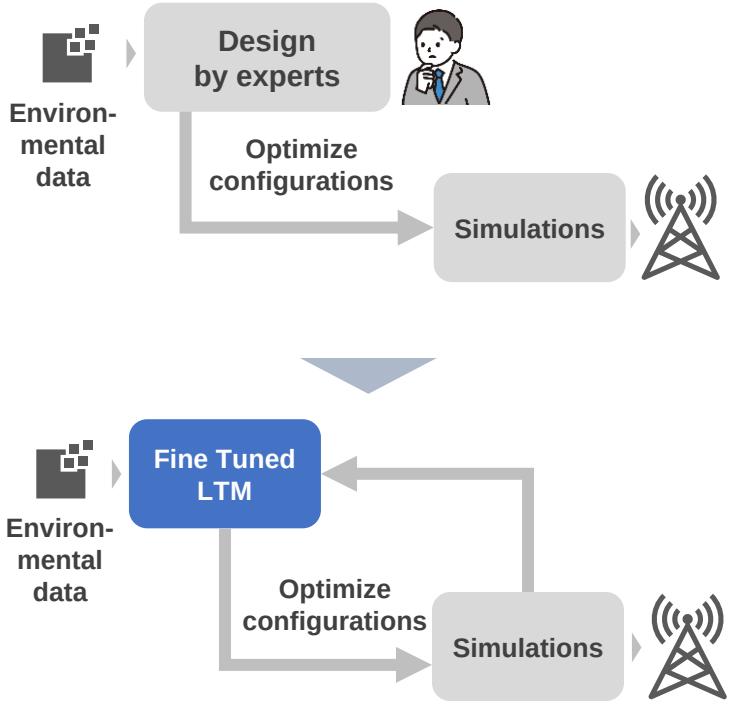
Cell ID	RAN KPI1	RAN KPI2
AAA	100	10
BBB	50	20
CCC	???	???

91% of value accuracy across all parameters

LTM even capable of full RAN config generation

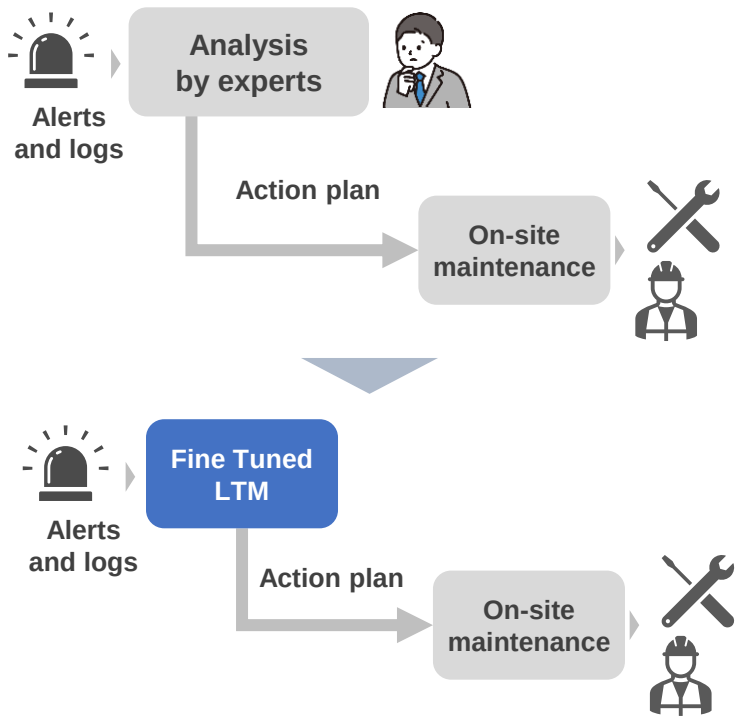
Examples of MNO Use Cases

Base Station Configurations



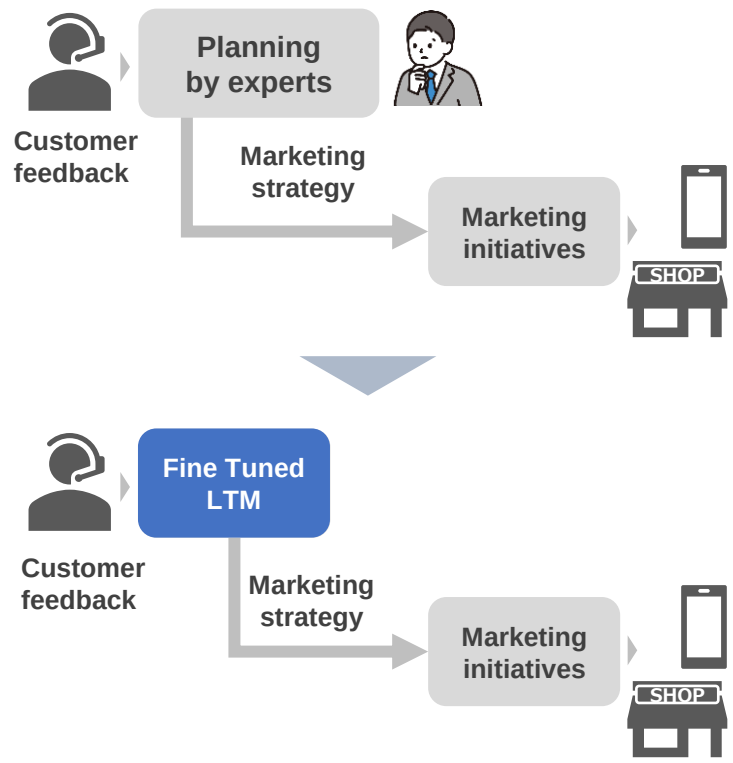
LTM analyzes data,
generates/optimizes configs,
deploys after simulation

Maintenance



LTM analyzes alerts and logs,
suggests causes and actions,
then on-site maintenance

Sales



LTM analyzes customer voice,
develops strategies,
executes promotions

Lessons Learned

SoftBank

Good

LTM's applicability to very specific downstream tasks i.e., use cases

Normally, using specific external tools is an option in such cases, but we proved possibility to solve these specific tasks with fine tuning

Good reproducibility of real-world optimization

Directly suggests the possibility of replacing existing business operations

No need to worry too much about the output format

Fine tuning is usually good at keeping format-level consistency

Could be better

Treating tabular data with LLM is computationally not efficient

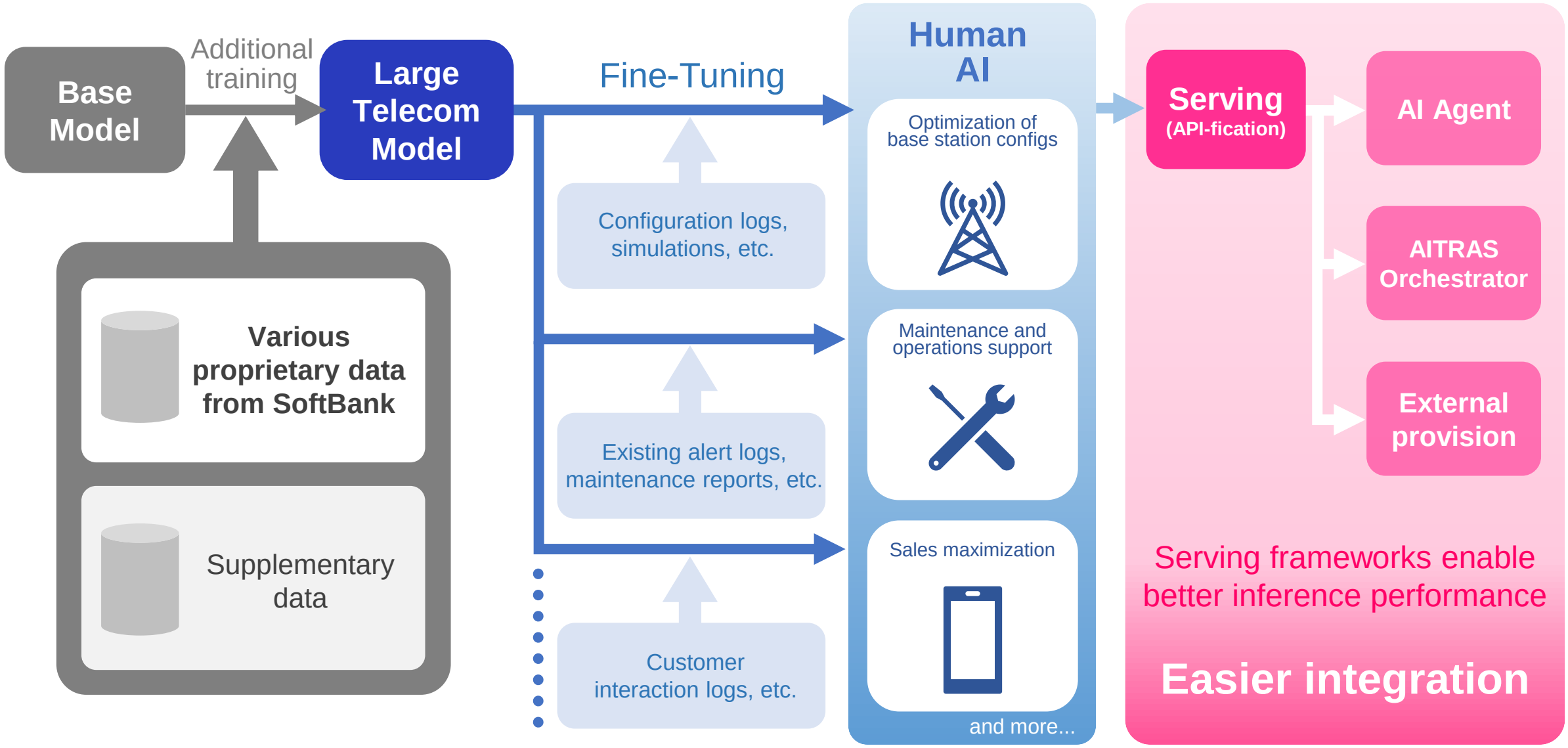
Specific table encoder can be a solution

NLT data is effective, but still needs more sampling

Reinforcement Learning will be needed

SoftBank

Integration into Actual Operations



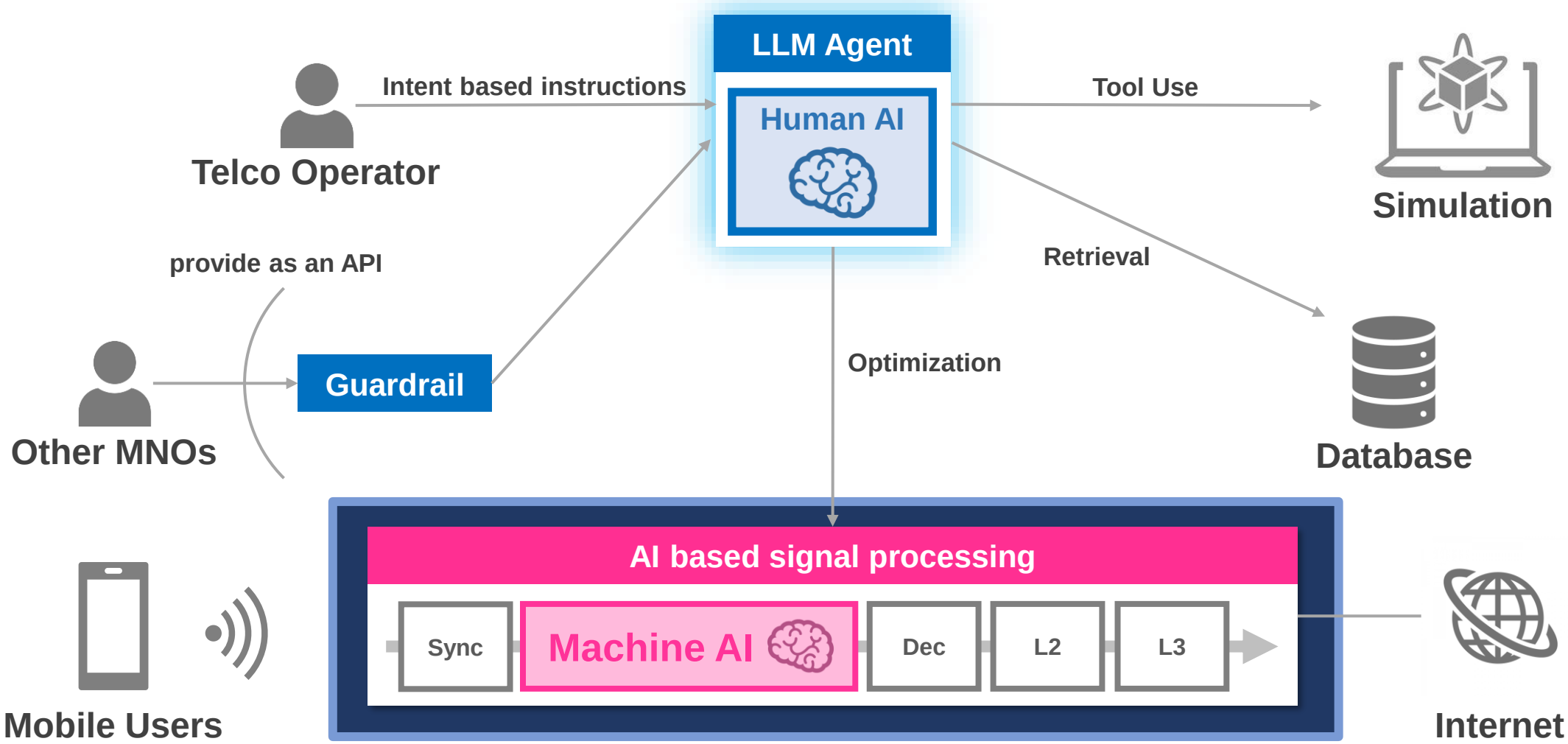
Future Outlook

Summary

- A foundation model provides a unified approach to adopting GenAI across an organization without requiring specialized skill sets
- A foundation model is also the basis for transfer learning to a variety of use cases in operations
- Telcos have enormous amounts of invaluable multi-modal data. Creating and managing quality data sets is half the problem
- Training and Evaluation require non-trivial experimentation involving AI and Mobile human expertise. And infrastructure!
- A foundation model may be a good investment if a Telco wishes to holistically evolve with GenAI
- SoftBank is upgrading LTM to take advantage of recent advances

Next Steps

SoftBank will replace and optimize various operational use cases with AI




SoftBank

Thank you!