

SCML 2026 – From LLM-Assisted Reasoning to Verified Systems

Approach for the Network Configuration of Wireless Systems using RAG, Fine-Tuning and Formal Verification Component

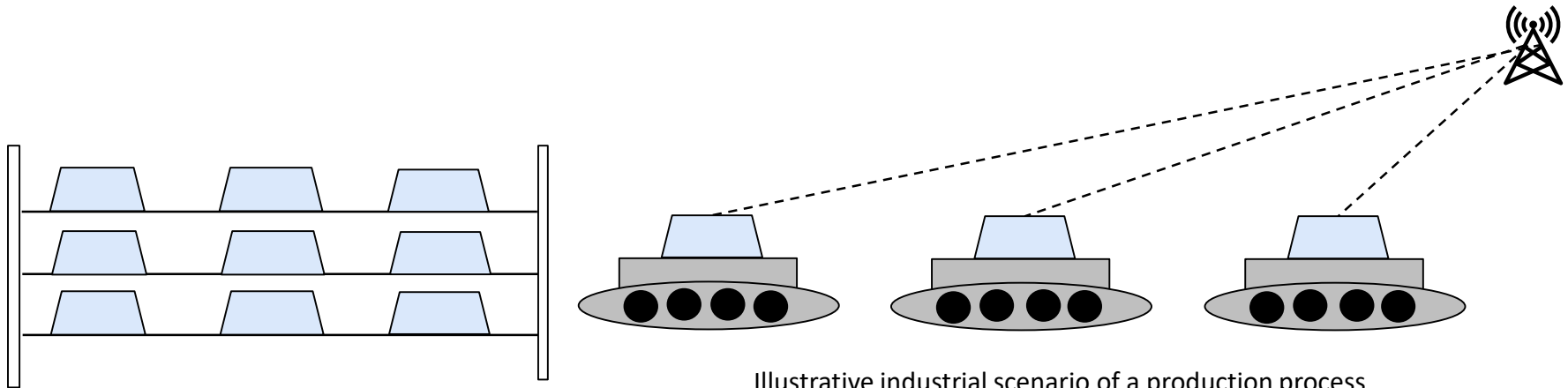
Steffen Fricke, Jürgen Jasperneite

Agenda

- Introduction
- Motivation
- Related Work
- Our Research
- Concept/Current Implementation
- Initial Evaluation and Verification of 5G System Configuration
- Conclusion and Outlook

Introduction

- **Trends and Challenges in Future Networks like 5G, 6G and TSN :**
 - **High-accuracy, low latency**, synchronization, and **reliability**
 - Increasing **softwarization, decentralization, and automation**
 - **Information Technology** and **Operational Technology** convergence



Motivation

- **Challenges in network configuration and management:**
 - **Increasing heterogeneity** and **complexity** of networks
 - Less experts and **expert knowledge**
 - Existing network management approaches **not sufficiently** handle the **growing heterogeneity, dynamic behavior**, and **service requirements** of modern wireless networks

- **Approach:**
 - Automated and **AI-driven decision-making** mechanisms
 - **Co-Pilot** for the **translation** of **high-level user intents** into **low-level network configurations**

Related Work

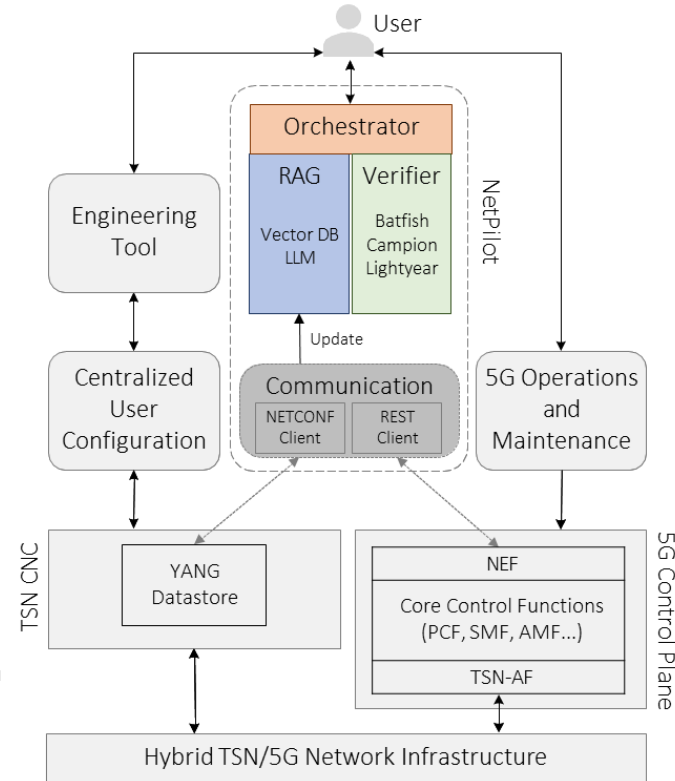
- **Current state, trends of using LLMs in Network Engineering [1]:**
 - Analyze technologies like 5G, 6G, Time-Sensitive Networking (TSN), and technology-independent approaches
 - **Substantial works** in the **5G domain** for the use cases of network configuration, network fault management, and network analysis
 - Limited work for TSN and 6G
- **Challenges:**
 - Most of the approaches are **simulation-based**
 - **well-known LLMs** like GPT, Llama and Gemini instead of specialized **fine-tuned models**
 - **Hallucinations** of LLMs which **produce incorrect output**, verification necessary
 - **Lack of datasets**

[1]: Steffen Fricke, Maxim Friesen, and Jürgen Jasperneite. "LLMs for Network Engineering Automation: State of the Art and Research Directions". In: Network Operations and Management Symposium, 7th IEEE/IFIP International Workshop on Management for Industry 5.0 (MFI5.0 2026). to appear. IEEE, May 2026

Related Work

- **Theoretical** and high-level **concept** for the configuration of hybrid TSN / 5G network
- High-level **user intent**
- **Orchestration** Layer for mangement
- **Retrival Augemented Generation (RAG)**
 - TSN configurations, 5G QoS parameter
- **Technology-specific verifiers** to validate configurations
- **Application Programming Interfaces**
 - NEF for 5G, NETCONF/YANG for TSN
- **Limitations:** Practical implementation is pending

NEF: Network Exposure Function
NETCONF: Network Configuration Protocol
YANG: Yet Another Next Generation
QoS: Quality of Service



[1]: Stefan Windmann et al. "NetPilot - Towards LLM-Assisted Configuration of Hybrid TSN/5G Networks". In: 2025 IEEE 30th International Conference on Emerging Technologies and Factory Automation(ETFA). 2025, pp. 1–4

Related Work

- **Two main approaches** for domain specific tasks are commonly used: **RAG and Low-Rank-Adaptation (LoRA) fine-tuning**

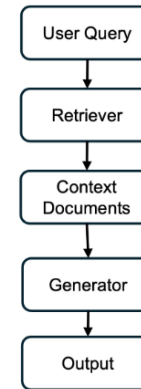
- **Traditional RAG:**

- Single-pass pipeline
- Offer domain knowledge
- Low inference latency

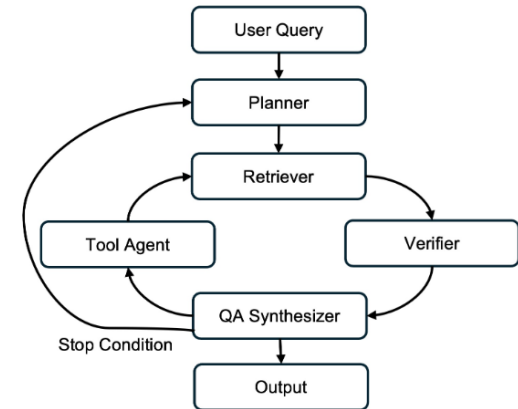
- **Agentic RAG:**

- Integrating external tools and iterative refinement mechanisms
- Structured reasoning and constraint-aware decision-making
- Increased latency

Traditional RAG



Agentic RAG



Fnu Neha and Deepshikha Bhatti. Traditional rag vs. agentic rag: A comparative study of retrieval-augmented systems. In 2025 IEEE International Conference on Future Machine Learning and Data Science (FMLDS), pages 383–387, 202

Related Work

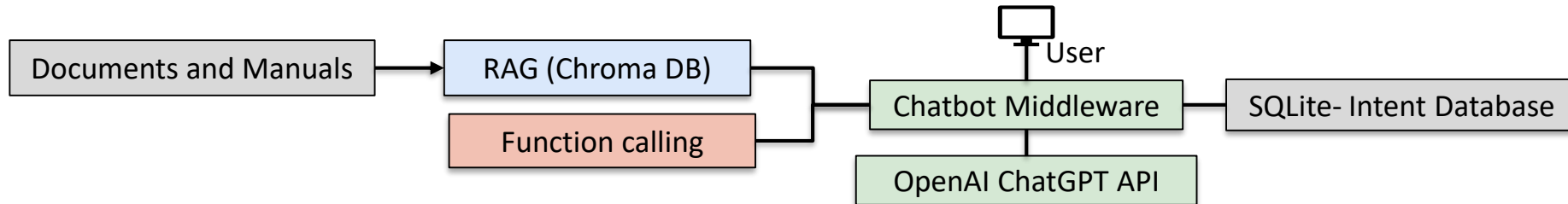
- **Current RAG and Fine-tuning approaches:**
 - Evaluation of **10000 Q & A multiple-choice questions** regarding general **understanding of telecommunication systems** [1]
 - **Vector database ChromaDB** to store 3GPP Release 18 dataset
 - Small and compact **model Phi-2** with **2.7 billion parameter**
 - More **specialized Q & A** in **O-RAN domain** [2]
 - Use of models such as **Llama 3.2, Qwen2.5-7B, and Qwen3.0-4B**
- **Limitations:**
 - **Q & A tasks is insufficient** for validating **real-world 5G system configuration**
 - **Missing open-end questions for reasoning**

[1]: Muhammad Ahmed Mohsin et al. Retrieval Augmented Generation with Multi-Modal LLM Framework for Wireless Environments. 2025.arXiv: 2503.07670.

[2]: Nathan Conger et al. "How to Discover Knowledge for FutureG: Contextual RAG and LLM Prompting for O-RAN". In: 2026 IEEE23rd Consumer Communications Networking Conference (CCNC).2026, pp. 1–6

Related Work

- Chatbot for the **reconfiguration of user equipment** using GPT-4
- Translate **high-level user intents** into low-level AT and **modem-specific commands**
- Use **RAG-system ChromaDB** to provide information about modem-specific AT commands
- Generated Python script is **verified using regular expressions**
- Practical test and evaluation using 5G network
- **Limitations:** relies on a general-purpose LLM, lacks real-world network measurements



1: Mattia Fontana et al. "Leveraging LLM-Powered Intelligent Chatbots for Intent-Based Networking in 5G Modem Reconfiguration". In: 2025 IEEE 11th International Conference on Network Softwarization (NetSoft). 2025, pp. 209–213.

Related Work

- **Fine-tuning of LLM [1]:**
 - **5G-systems** are controlled through APIs and **Network Exposure Function (NEF)**
 - **GPT-4** is used to generate a **dataset of 700 valid configurations**
 - **Quantized-Low-Rank Adaptation** to fine-tune a model Phi-2

- **Retrieval Optimization [2]:**
 - Improve efficiency by **strategic context reordering** to mitigate the **“lost-in-the-middle” problem**
 - **Less attention** to information located in the middle of the input
 - Most relevant passages at the beginning the query at the end, retrieval effectiveness

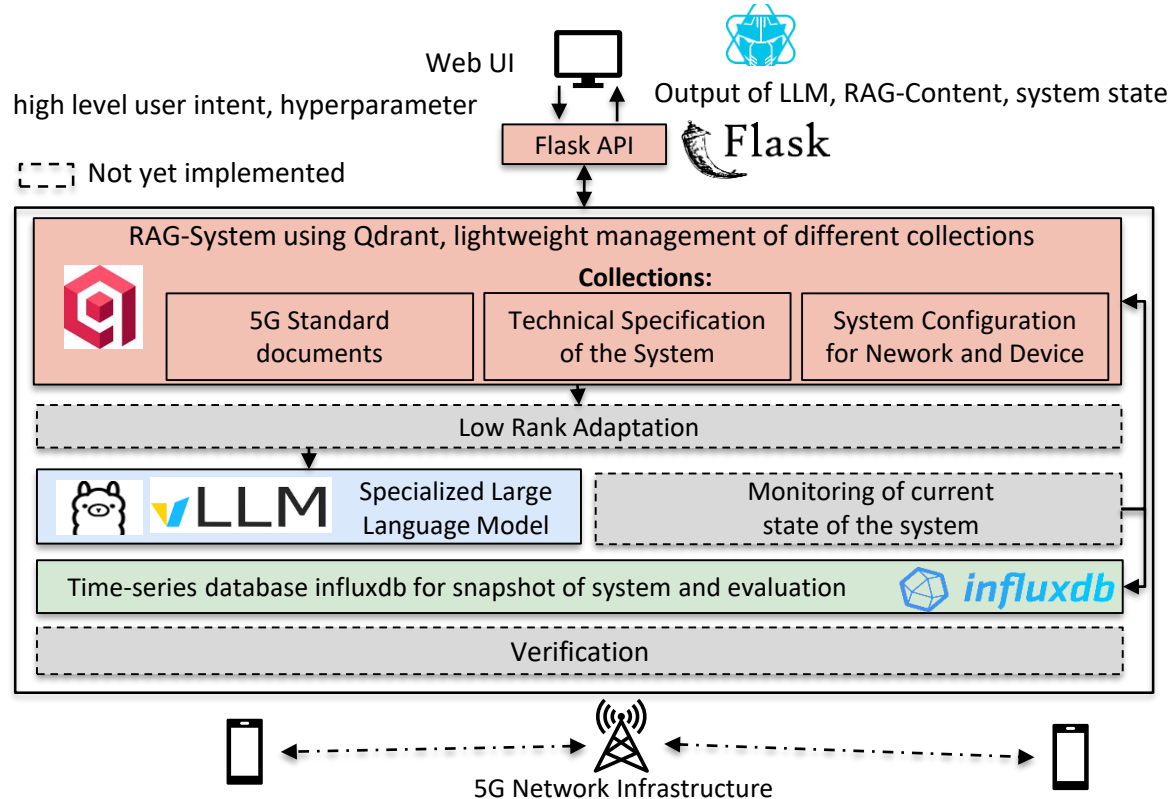
[1]: Zainab Khan, Ahmed Hussain, Mukesh Thakur, Arto Hellas, and Panos Papadimitratos. Nefmind:Parameter-efficient fine-tuning of open-source llms for telecom apis automation, 2025

[2]: Yufei Chen, Zhiliang Huang, and Youyan Zhang. Efficient-rag: Competitive single-pass performance in rag via strategic context reordering. In 2025 18th International Conference on Advanced Computer Theory and Engineering (ICACTE), pages 188–194, 2025

Our Research

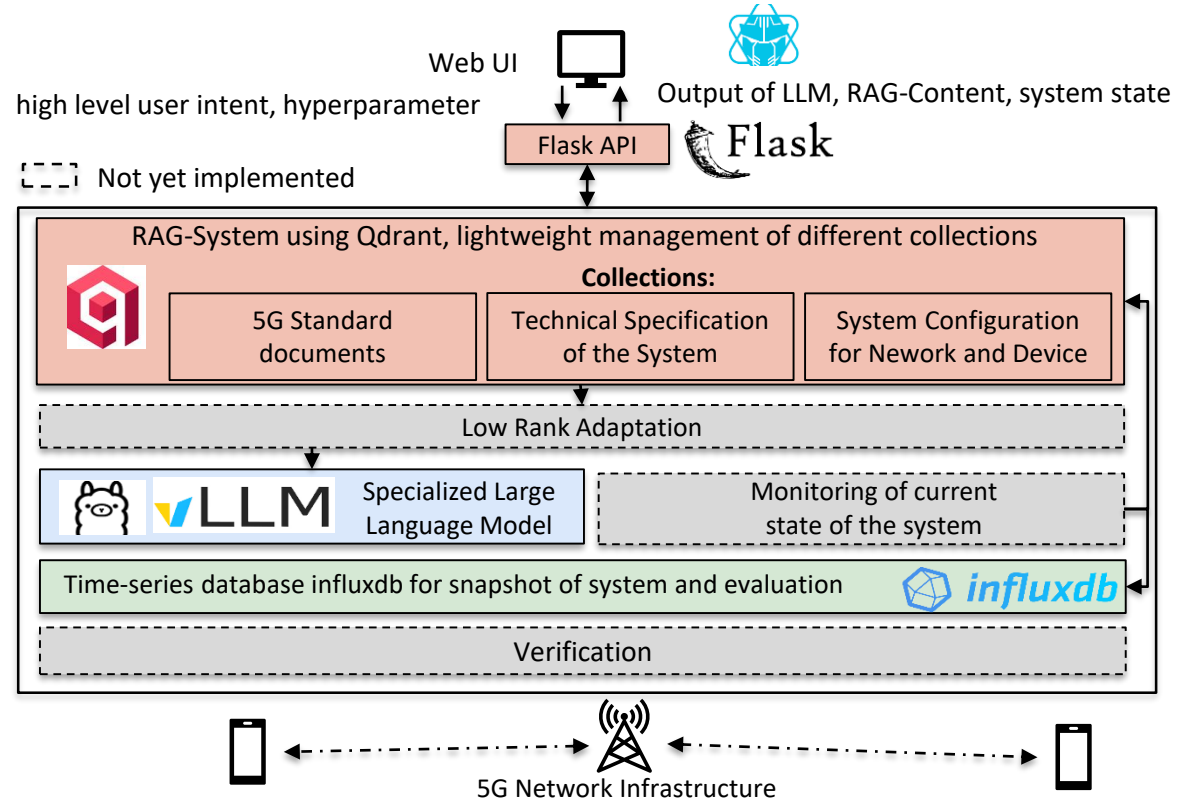
- **Implementation of a Co-pilot for the configuration of a 5G-system:**
 - Web UI for the interaction with the user
 - **RAG-system** for contextual knowledge
 - **Specialized and Fine-tuned model trained** on technical data
 - Verification of generated output
 - **Practical evaluation of commercial and open source 5G-systems**

Concept/Current Implementation



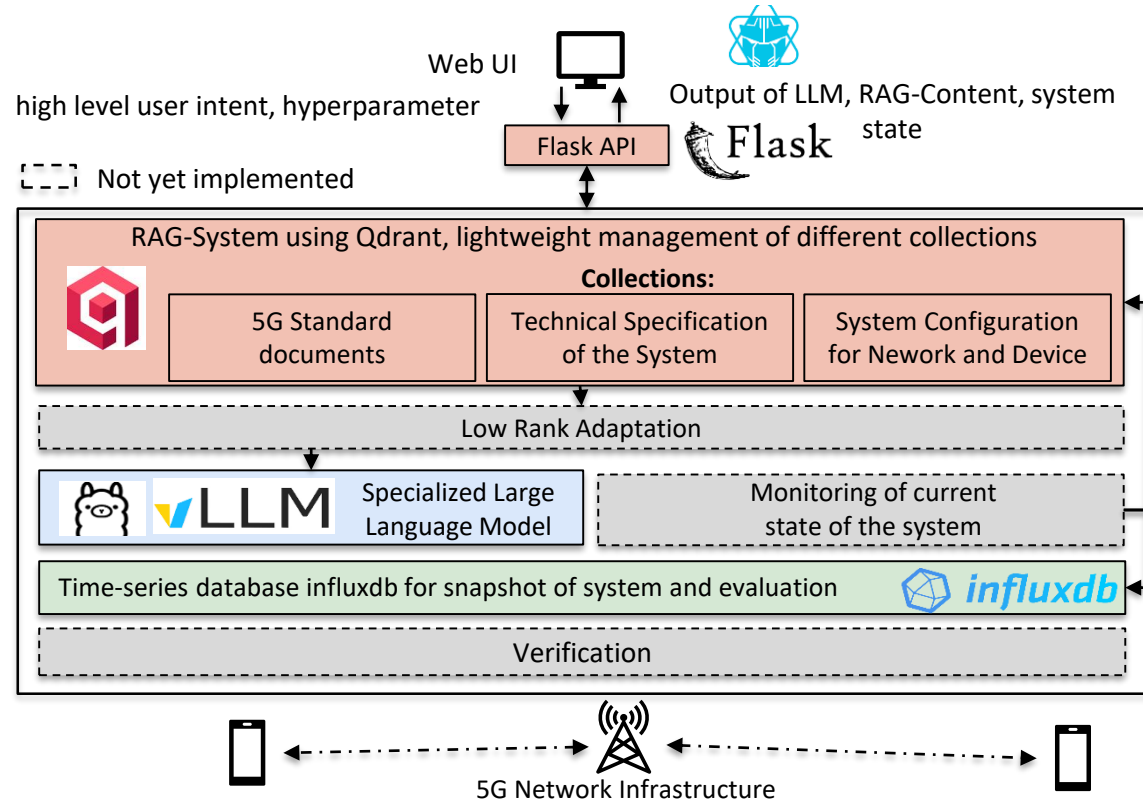
Concept/Current Implementation

- **WebUI**
 - Framework PrimeReact
 - Input /Output of LLM
 - Display system state
- **Flask API**
 - Bidirectional exchange between Web UI and backend
- **RAG system**
 - Contextual knowledge
 - Standards
 - User manuals
 - System parameter



Concept/Current Implementation

- **Fine-tune model**
 - Low-Rank Adaptation
- **Specialized LLMs**
 - Framework Ollama
 - Qwen-3-8b-Instruct
 - Phi-2
- **Verification**
 - Check syntax and semantic
- **Time series database**
 - Influx DB
- **Commercial and Open Source 5G systems**



Initial Evaluation and Verification of 5G System Configuration

- **Generation of curl-commands** for addition, deletion, and modification of system configurations of **commercial 5G System MECsWare** through the API
- **Store user manual** in the **vector database Qdrant**
- Evaluation using 50 Test prompts
 - Model Qwen-3-8b-Instruct



```
{
  "id": 1,
  "category": "Base Station Management - Add",
  "platform": "Linux",
  "prompt": "Create a curl command for Linux to add a new
    base station named 'hall 1', type 'SCE5164', IP
    '10.0.2.110', and MAC address '3C:62:F0:17:24:AA'.
    The description should be 'production hall 1'."
  "ground_truth": "curl -X POST https://10.0.1.2:6327/5g
    /bs/conf -k -u admin:admin -d '{\"Name\": \"hall
    1\", \"Type\": \"SCE5164\", \"Description\": \"
    production hall 1\", \"IP\": \"10.0.2.110\", \"MAC
    \": \"3C:62:F0:17:24:AA\"}' -H 'Content-Type:
    application/json' "
}
```

Initial Evaluation and Verification of 5G System Configuration

- **Example for the configuration of a 5G system:**

- **Prompt:** "Create a curl command for Linux to add a new base station named 'production hall 1', type 'SCE5164', IP '10.0.2.110', and MAC address '3C:62:F0:17:24:AA'. The description should be 'production hall 1' "
- **Ground Truth:** "curl -X POST https://10.0.1.2:6327/5g/bs/conf -k -u admin:admin -d '{"Name": " production hall 1 ", "Type": "SCE5164", "Description": "production hall 1", "IP": "10.0.2.110", "MAC": "3C:62:F0:17:24:AA"}' -H 'Content-Type: application/json'"
- **Output of the LLM:** curl -X POST http://localhost:8080/5g/bs/conf -H ""Content-Type: application/json"" -d '{"name":""," production hall 1","", "type":"","SCE5164", "ip":"","10.0.2.110", "mac":"","3C:62:F0:17:24:AA", "description":""," production hall 1"}' -k -u admin:admin

Initial Evaluation and Verification of 5G System Configuration

- **Verification of the generated output to avoid unintended system behavior:**
 - **Formal Verification:**
 - Check curl and JSON format
 - **Guardrail & Value Range Validation:**
 - Check output compliance with guardrails and constraints
 - **Prompt Coverage Validation:**
 - Check whether the response addresses all prompt requirements

- **Verification fails:**
 - **Single-pass pipeline:** No configuration is performed
 - **Agentic implementation:** Iterative refinement of the generated output based on the feedback using external tools

Conclusion and Outlook

- **Concept and partially implemented single-pass-pipeline** for the configuration of 5G-systems
 - **Web-user-interface** interaction with the system
 - **RAG-component** using Qdrant easy to use for collection management
 - **Specialized models:**
 - Qwen-3-8b-Instruct
 - Phi-2
 - **Time-series data base** to store snapshots of the system for the evaluation
 - **Fine-tuning** using LoRA
 - **Verification component**

Conclusion and Outlook

- **Further work:**
 - Various **commercial** and **open-source 5G systems** will be evaluated
 - Extend single-pass pipeline to **agentic implementation** and a **multi-agent system**
 - **Fine-tuning** of the model using **LoRA** to evaluate and refine the quality of the generated output



Thank you

M. Sc. Steffen Fricke
WG Industrial Communication
Institute Industrial IT
OWL University of Applied Sciences and Arts
Campusallee 6,
D-32657 Lemgo
steffen.fricke@th-owl.de