

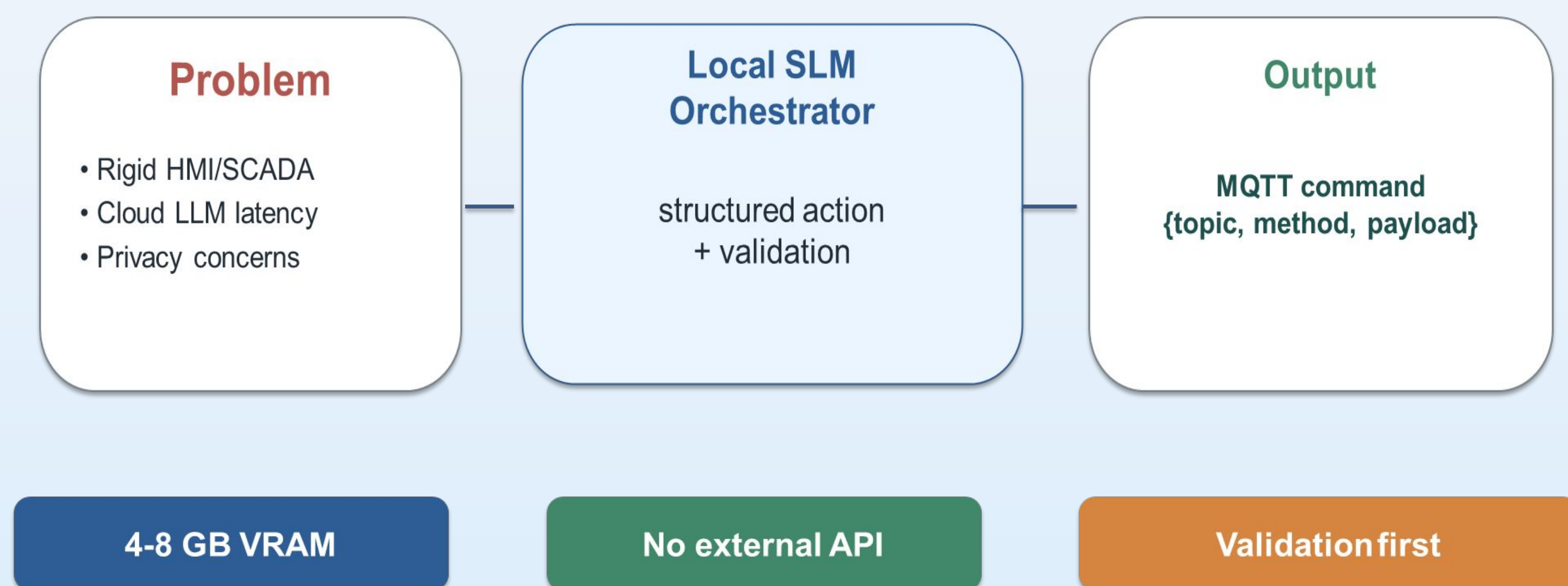
Student(s)

Dilara Alkanalka, Hüseyin
Eren Yıldız, Mert Seçen,
Neşenaz Yalçın

Faculty Member(s)

Dilara Keküllüoğlu

Company Advisor(s)

Erdem Çimenoglu
Turgut Helvacioğlu**ABSTRACT**

This project develops a resource-efficient AI Action Orchestration Platform that enables industrial operators to control edge devices using natural language commands. The system translates operator inputs into validated MQTT protocol messages, running entirely on local hardware [2].

Problem Statement

- **Cloud Dependency:** Latency and bandwidth constraints in industrial settings.
- **Security:** Operational data must remain on-premise.
- **Resource Constraints:** Edge devices limited to 4-8 GB VRAM.
- **Rigid Interfaces:** HMIs lack flexibility for complex operator queries.

Motivation

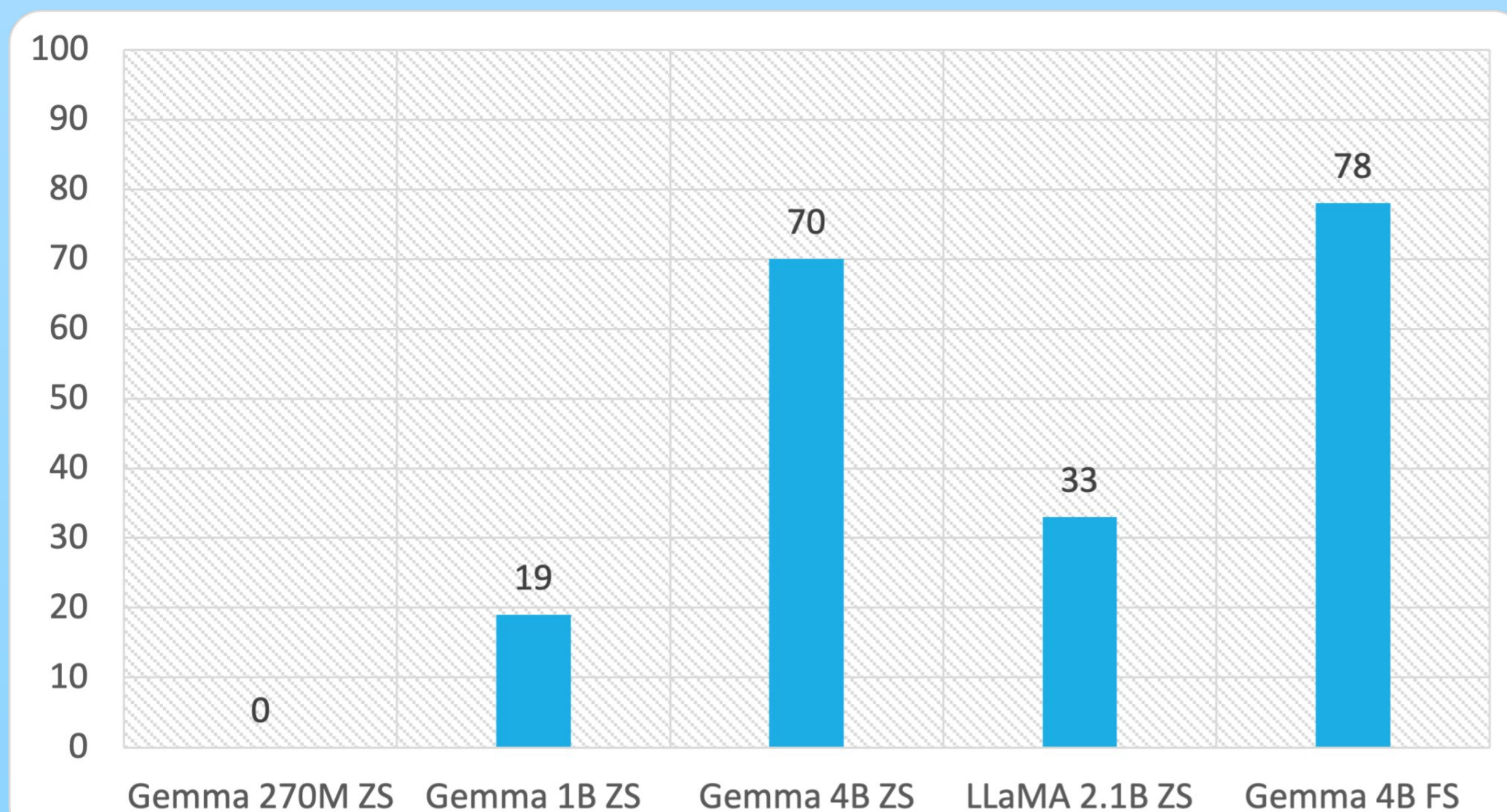
Enabling local execution of SLM models to reduce training requirements and human error while ensuring secure, low-latency machine control.

Approach

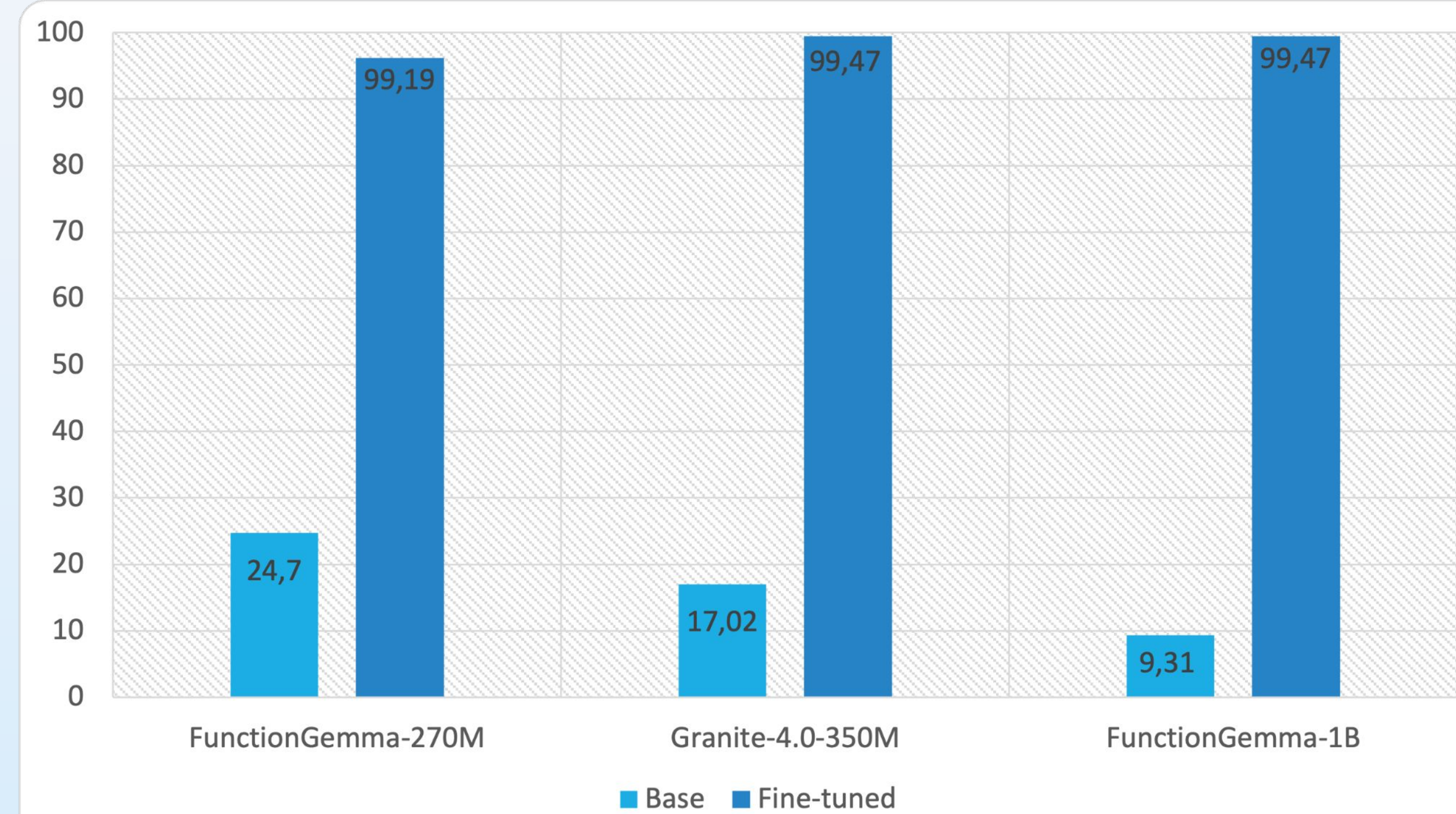
The system uses a fine-tuned SLM to generate structured outputs in topic/method/payload format, followed by a two-stage validation layer. Structural validation checks whether the output is parseable and complete. Semantic validation checks whether the correct device and action were selected. Outputs that fail either stage are blocked before reaching the MQTT layer. The model is also trained on negative examples, teaching it to return null for unsupported or ambiguous inputs rather than generating a plausible-but-wrong command.

OBJECTIVES

1. Define a protocol-oriented command schema: topic (sensor ID), method (publish/subscribe), payload (action type)
2. Generate a synthetic dataset with 3753 examples: positive commands, neutral/ambiguous inputs, and negative (out-of-scope) examples
3. Benchmark Gemma and LLaMA model families under zero-shot and few-shot conditions
4. Fine-tune compact function-oriented models (FunctionGemma-270M, FunctionGemma-1B, Granite-4.0-350M)

BENCHMARKING RESULTS

Gemma and LLaMA families were tested without fine-tuning [1]. Best result: Gemma3:4B few-shot at 78% accuracy but with ~34 second average response time, too slow for interactive industrial use. Key finding: accuracy alone is not enough, latency must be evaluated alongside correctness for edge deployment.

DATASET & FINE-TUNING METHODOLOGY**DATASET**

The fine-tuning dataset contains 3753 English-language examples covering sensors s1–s20 with 25 supported payload types. Split: 80% train / 10% validation / 10% test.

- Positive (3180): valid MQTT commands with topic, method, payload
- Neutral (200): ambiguous or incomplete sensor-related inputs → null
- Negative (373): out-of-scope requests and hard negatives → null

FINE-TUNING CONFIGURATION

Three models were selected based on their parameter efficiency and function-calling orientation [3].

FunctionGemma-270M — Google's 270 million parameter model designed specifically for function calling.

FunctionGemma-1B — The 1 billion parameter version of the same architecture.

Granite-4.0-350M — IBM's 350 million parameter model.

FINE-TUNING RESULTS

Evaluation used exact match scoring; a prediction was only counted as correct if all three fields (topic, method, and payload) exactly matched the reference output.

The results show dramatic improvement across all three models. Before fine-tuning, Granite's null accuracy was 0% it attempted to generate a tool call for every single input but after fine-tuning on our negative examples, null accuracy reached 96.77%. This is the core safety contribution of the project: the model learned not just how to generate commands, but also when not to.

CONCLUSIONS

This project demonstrates secure, local natural language control of industrial IoT on resource-constrained edge devices using Small Language Models.

- Fine-tuned compact SLMs provide the best accuracy–latency balance compared to larger models.
- Larger models showed better reasoning potential but suffered from impractical latency.
- Safety is improved through out-of-scope rejection and two-stage validation.
- Only structurally and semantically valid MQTT commands are allowed to reach the hardware.

FUTURE WORK

Future work targets advanced quantization to heavily compress SLMs, ensuring high token throughput and a minimal memory footprint without accuracy loss.

Additionally, the system will evolve into a hybrid architecture integrating Rasa as a preemptive layer. This deterministic layer will use Named Entity Recognition (NER) for fast pre-processing and similarity mapping, dynamically invoking SLMs only when rule-based matching fails.

Finally, evaluating these quantized models and hybrid workflows on physical hardware, alongside fine-tuning on expanded synthetic datasets, will maximize domain-message compliance and establish a scalable, reliable standard for industrial IoT.

REFERENCES

- [1] D. Alkanalka, M. Seçen, N. Yalçın, H. E. Yıldız, E. Çimenoglu, T. Helvacioğlu, and D. Keküllüoğlu, “Uç Yapay Zekanın Nesnelerin İnternetine Entegrasyon Mimarisi / Edge AI Integration Architecture for the Internet of Things,” SIU 2026, 2026.
- [2] OASIS MQTT Technical Committee, “MQTT Version 5.0,” OASIS Standard, 2019.
- [3] W. Chen, Z. Li, and M. Ma, “Octopus: On-device language model for function calling of software APIs,” Proc. NAACL-HLT Industry Track, 2025.