

Nordic AI Meet 2026

P4: Machine Learning, Intelligent Systems & AI for Science and Engineering

6. Deployed LLM Semantic Parser Pipeline for Drone Allocation System on Embedded RISC-V Hardware

Authors: Adam Laurell^{1,2}; Elias Kettunen^{1,2}

Affiliations: 1 Mälardalen University, Västerås, Sweden; 2 Cicor Nordic Engineering AB, Uppsala, Sweden

Abstract: Large language models (LLM) let warehouse operators dispatch and re-task an autonomous drone fleet in plain natural language — but their probabilistic, opaque behaviour makes them unsafe to wire directly into actuation. We present an LLM semantic parser implementation in a drone allocation system for warehouse monitoring and inventory in which a compact, quantised LLM, running on-device with llama.cpp, parses natural-language instructions into structured allocation commands.

Between the LLM's structured output and the allocation logic we insert a validation layer that passes safe commands and corrects or rejects any command that would drive the system into an unsafe state. This decouples safety from the LLM's internals: the shield does not need to trust or interpret the model's output, only check it against formally verified constraints.

We build an abstraction of the allocation system in Rebeca, an actor-based modelling language with model-checking support, express the system's safety and consistency requirements as assertions, and use model checking to verify they hold for any command sequence in the abstraction. The shield is then constructed to enforce these assertions by design, decoupling the safety case from the unverifiable internals of the model.

We deploy the full pipeline — LLM parser and validation layer — which interacts with a simulated allocation system. The parser runs on a RISC-V board under embedded Linux. Since the RVV 1.0 vector extension is effectively a fixed-function accelerator for the LLM's matrix-vector operations, we use this deployment to study how well the model's compute pattern matches the vector unit, profiling latency and energy with RVV enabled versus disabled and comparing this compute-unit-assignment effect against that of changing the LLM itself.

Index Terms: large language models, autonomous systems, RISC-V, model checking, safe AI, Rebeca, embedded systems.