ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Institut für Integrierte Systeme
Integrated Systems Laboratory

Semester's Thesis at the Department of
Information Technology and Electrical Engineering

Autumn Semester 2025

Garcia Cardenas Facundo

A RISC-V-based Hybrid MPC–LLM Framework for Adaptive Autonomous Driving

September 30, 2025

Advisers: Dr. Lorenzo Lamberti (IIS), ETZ J 68.2, llamberti@iis.ee.ethz.ch
Liam Boyle (PBL), ETZ, liam.boyle@pbl.ee.ethz.ch
Dr. Alessio Burrello, PoliTo, alessio.burrello@polito.it
Handout: Oct 1, 2025
Due: Jan 7, 2026

1 Project description

Recent advances in Large Language Models (LLMs) have shown promise for enhancing Autonomous Driving Systems (ADS) by complementing purely data-driven neural networks with knowledge-driven reasoning models [1, 2]. While data-driven models often fail to generalize to underrepresented edge-case scenarios, LLMs can incorporate human intent and contextual cues through natural language, effectively introducing human decision-making in the loop.

Building on this vision, prior work has integrated Model Predictive Control (MPC) for low-level control with LLM-based reasoning for decision-making and control adaptability in ADS [1]. This hybrid framework demonstrated how LLMs can adapt MPC cost functions and constraints in response to high-level instructions.

The goal of this thesis is to port and extend this hybrid MPC–LLM framework onto the RISC-V-based Axelera AI accelerator[1], a high-performance edge computing platform optimized for deep learning inference. Together with a compact ARM-based Mini-ITX single board computer (SBC), this setup aims to deliver a power-efficient, embedded solution for embodied AI in robotics.

This semester project will explore: deployment of pre-compiled and pre-trained LLMs on Axelera hardware; integration of Retrieval-Augmented Generation (RAG) to enhance the LLM output; benchmarking accuracy, latency, and power efficiency compared to NVIDIA Orin; validation of the system on a small dataset [1] and in closed-loop withing a driving simulator.

2 Project Goals

The project aims to achieve the following milestones (M):

1. **M1:** Familiarize with the Axelera sdk, pre-compiled LLM inference on the Axelera PCIe board, and familiarize with the existing LLM-enhanced autonomous driving system for car racing [1] (1W).
2. **M2:** Using Axelera's pre-trained LLM on a intel host to enable MPC tuning of the reference LLM-enhanced ADS system[2] and compare the accuracy with the dataset of the reference paper [1] (2W).
3. **M3:** Port the intel-host Axelera setup to an embedded ARM-based single computer board (SBC), the Firefly ITX-3588J with a RK3588 Rockchip (2W).
4. **M4:** Deployment and integration of RAG on the Firefly embedded setup using lama.cpp[3] (2W).
5. **M5:** power/latency profiling of the ITX-3588J, and comparison to the Nvidia Orin AGX setup (1W).

[1] `https://axelera.ai/metis-aipu-benchmarks`
[2] `https://github.com/ForzaETH/LLMxRobot`
[3] `https://github.com/ggml-org/llama.cpp`

6. **M6:** install ROS and the simulator for autonomous car racing (follow GitHub[4] or the tutorial videos[5]), and replicate the current setup for car racing [1] (1W).

7. **M7:** in the simulator, evaluate in closed-loop the pre-trained Axelera LLM for ADS and compare its performance to the LoRA finetuned LLM of [1] (2W).

8. **M8:** Perform code refactoring, documentation, and release (1W).

9. **M9:** Final report writing and thesis presentation (2W).

2.1 Project Plan

Within the first month of the project, the student will be asked to prepare a project plan. This plan should identify the tasks to be performed during the project and individual tasks' deadlines. The prepared plan will be discussed with the supervisors during the first week's meeting. Note that the project plan should be updated constantly depending on the project's status.

2.2 Meetings

Weekly meetings will be held between the student and the supervisors. These meetings' exact time and location will be determined during the project's first week. These meetings will be used to evaluate the status and progress of the project and to advise and support the student. Besides these regular meetings, additional meetings can also be organized to address urgent issues.

2.3 Report

Documentation is an important and often overlooked aspect of engineers. One final report has to be completed within this project. The final report will be written in English. Any form of word processing software is allowed for writing the reports, nevertheless the use of LaTeX with Tgif[6] or any other vector drawing software (for block diagrams) is strongly encouraged by the IIS staff.

Final Report The final report has to be presented at the end of the project, and a digital copy needs to be handed in. Note that this task description is part of your report and has to be attached to your final report.

[4] `https://github.com/ForzaETH/race_stack`

[5] `https://www.youtube.com/playlist?list=PLMzSGo5LtaW9cgdwHB_FnX3qlAYx7P6JI`

[6] See: `http://bourbon.usc.edu:8001/tgif/index.html` and `http://www.dz.ee.ethz.ch/en/information/how-to/drawing-schematics.html`.

2.4 Presentation

There will be a presentation at the end of this project to present your results to a wider audience: Master's Thesis have 20 min presentation and 5 min Q&A, Semester/Bachelor's thesis have (15 min presentation and 5 min Q&A). The exact date will be determined towards the end of the work.

References

[1] Nicolas Baumann et al. *Enhancing Autonomous Driving Systems with On-Board Deployed Large Language Models*. 2025. arXiv: 2504.11514 [cs.AI]. URL: https://arxiv.org/abs/2504.11514.

[2] Liam Boyle et al. *RobotxR1: Enabling Embodied Robotic Intelligence on Large Language Models through Closed-Loop Reinforcement Learning*. 2025. arXiv: 2505.03238 [cs.RO]. URL: https://arxiv.org/abs/2505.03238.

Zürich, September 30, 2025 Prof. Dr. Luca Benini