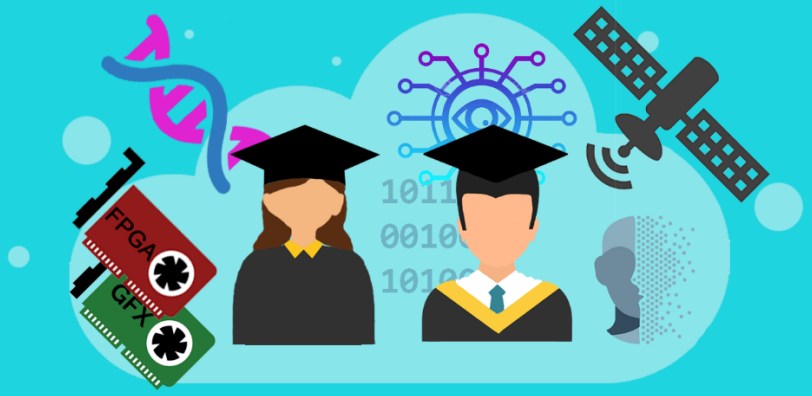


Diploma Thesis

Microprocessors and
Digital Systems
Laboratory



Agentic AI for Automated Deployment and Resource Management of LLM Services

Large language models (LLMs) are increasingly adopted across domains such as software development, information retrieval, conversational assistance, and content generation. Models supporting these applications differ in their capabilities and resource requirements, while the services built around them impose varying latency and throughput targets. Efficiently serving these models therefore requires deployment and resource management strategies that account for both model characteristics and application requirements.

Serving multiple LLMs introduces complex decisions concerning deployment configurations, resource allocation, and request scheduling. These decisions depend on the available **computing infrastructure**, **workload characteristics**, and **performance targets**, and must accommodate changes in demand. Existing approaches often require substantial manual configuration or specialized solutions tailored to particular setups, making efficient operation difficult to maintain as workloads and infrastructure evolve.

This thesis will explore the use of AI agents to automate the deployment and resource management of LLM services. It will investigate how agents can interpret service requirements, monitor system behavior, and interact with deployment and profiling tools to **identify effective configurations and guide resource allocation and scheduling decisions**. The objective is to assess whether agent-driven management can reduce manual effort and improve resource utilization while satisfying service performance targets.

RELATED MATERIAL:

- [1] Z. Li et al., "AlpaServe: Statistical Multiplexing with Model Parallelism for Deep Learning Serving," in *17th USENIX Symposium on Operating Systems Design and Implementation (OSDI 23)*, 2023, pp. 663–679.
- [2] B. Sun, Z. Huang, H. Zhao, W. Xiao, X. Zhang, Y. Li, and W. Lin, "Llumnix: Dynamic Scheduling for Large Language Model Serving," in *18th USENIX Symposium on Operating Systems Design and Implementation (OSDI 24)*, 2024, pp. 173–191.
- [3] Chen, Y., et al, "AIOpsLab: A Holistic Framework to Evaluate AI Agents for Enabling Autonomous Clouds," in *Proceedings of Machine Learning and Systems*, 2025.

CONTACT INFORMATION:

Christos Kavallaris, Ph.D. Student, NTUA: (ckavallaris@microlab.ntua.gr)
Achilleas Tzenetopoulos, Post Doc, NTUA
Dimosthenis Masouros, Post Doc NTUA
Sotirios Xydis, Ass. Prof., Microlab NTUA
Dimitrios Soudris, Professor Microlab NTUA