

ALMA MATER STUDIORUM · UNIVERSITÀ DI BOLOGNA

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
ARTIFICIAL INTELLIGENCE

MASTER THESIS

IN

ARCHITECTURES AND PLATFORMS FOR ARTIFICIAL INTELLIGENCE

**Optimizing Small Language Models: An
Experimental Investigation in Compressing
Distilled LLaMA Architectures**

Supervisor:
Prof. Francesco Conti

Candidate:
Angelo Galavotti

Co-supervisor:
Luca Bompani

Session I
Academic Year 2024/2025