ALMA MATER STUDIORUM · UNIVERSITÀ DI BOLOGNA

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

ARTIFICIAL INTELLIGENCE

Optimizing Small Language Models: An Experimental Investigation For Compressing Distilled LLaMa-based Models

Supervisor:
Prof. Francesco Conti

Presented by:
Angelo Galavotti

Co-supervisor:
Luca Bompani

Session I
Academic Year 2024/2025