YEDITEPE UNIVERSITY
Department of Computer Engineering

SEMINAR

April 7, 2026 (Tuesday)

11:00 – 12:00

Engineering Faculty
Room: B-217

DIFFERENT TECHNIQUES FOR PREVENTING JAILBREAK ATTEMPTS IN LARGE LANGUAGE MODELS

Samir Ganbarli

Software Engineer
Orion Innovation Turkey

Abstract

Today, large language models (LLMs) have become widely adopted across many domains, bringing both significant opportunities and critical security concerns. One of the most important of these concerns is the emergence of "jailbreak" attacks, where malicious inputs are used to bypass model safeguards and manipulate outputs. Therefore, understanding how these attacks work and how to prevent them has become essential for building reliable AI systems. In this context, different types of jailbreak attacks such as character-level manipulation, token-level encoding, and role-play techniques, along with prevention methods including input/output filtering, supervised fine-tuning (SFT), and reinforcement learning from human feedback (RLHF), constitute the main topics of this seminar.

Biography

Graduated with a bachelor's degree from ADA University in Baku in computer engineering department (2021). I started my master's degree in the same field at Yıldız Technical University in the same year and graduated in 2024. In the same year, I started my PhD studies at the same university and currently I am on thesis phase. Currently, I am researching RLHF method and some advancements in it to prevent jailbreak attacks using low computational power. In addition, I have 3 years of experience in backend development field.