Project Proposals

Amit LeVi
amitlevi@campus.technion.ac.il

Registration Form

Short Introduction

Large Language Models (LLMs) are at the forefront of AI research, enabling advanced capabilities in text generation, understanding, and reasoning. However, they are also susceptible to adversarial attacks and misalignment issues. These projects will allow you to explore cutting-edge methodologies, from building robust defenses and new optimization strategies to crafting innovative applications such as educational platforms and deepfake detection. Each project is designed to enrich both your research acumen and your practical engineering skills.

Requirements

- Basic familiarity with LLM frameworks (e.g., Hugging Face) and general deep learning principles.
- Ability to read and analyze academic papers to inform your research direction.
- Some experience running ML experiments is helpful, but not strictly required.

Team Structure

- Projects are intended for teams of two; individual or three-person teams may be approved based on need.

Project Proposals

Project 1: Planning Attacks on LLM Agents

Topic: Adversarial Attacks in Planning and Reasoning **Abstract:** In this project, you will explore how well-coordinated adversarial prompts can manipulate the decision-making processes of LLMs, particularly in multi-step or multi-turn planning scenarios. By leveraging modern NLP libraries (e.g., Hugging Face) and GPU-accelerated environments, you will craft experiments that expose subtle model vulnerabilities. You will then systematically analyze how LLMs agents response and use vulnerabilities for control theirs pipeline. This study aims to deepen our understanding of how LLMs Agents reason and plan, for offering insights into better securing them against targeted manipulation.

Project 2: Embedding Jailbreak Defense for LLMs

Topic: Jailbreak Attacks and Internal Model Defenses **Abstract:** This project focuses on proactively protecting LLMs from jailbreak attacks by embedding protective mechanisms within the model architecture. You will investigate how strategic modifications to embedding layers, attention mechanisms, or hidden states can detect malicious prompts when it circumvent alignment filters. Through extensive testing, you will evaluate the model's behavior under increasingly sophisticated attacks, aiming to identify where vulnerabilities

lie and how they can be mitigated at a structural level. The outcome will include both theoretical analysis and practical demonstrations of improved robustness by creating injection's prompt classifier.

Project 3: Irrational and Rational Reasoning Patterns in LLMs

Topic: Reward Signals and Causal Inference **Abstract:** Here, you will examine how LLMs sometimes learn irrational associations due to misaligned or noisy reward signals. By creating controlled training environments with reinforcement learning, you will trace the emergence of spurious correlations—such as linking performance success to irrelevant tokens or contextual cues. Through causal analysis and refined reward design, you will propose strategies to evaluate the models focus on truly relevant factors.

Project 4: Do LLMs Comply Together?

Topic: Collaborative Outputs and Adversarial Prompts **Abstract:** This project investigates whether multiple LLMs, each trained on potentially different datasets or with unique hyperparameters, respond consistently to similar adversarial prompts. You will run parallel experiments to collect embedding representations and output distributions, seeking to understand why some models might resist manipulation while others succumb. By identifying patterns in how they diverge or converge, you will offer insights into the collective dynamics of LLM ensembles. These findings could inform strategies for building more unified, adversarially robust AI systems.

Project 5: Hidden Attacks in Retrieval-Augmented Generation (RAG) Models

Topic: External Data Manipulation and Adversarial Embedding **Abstract:** RAG models enhance their outputs by pulling information from external sources, such as databases or web pages. This project examines how adversaries can embed harmful or misleading content in these external repositories to subtly alter the model's final responses. You will design and implement a suite of adversarial techniques—ranging from keyword manipulation to semantic obfuscation—to showcase the vulnerabilities present in the retrieval stage.

Project 6: Exploring Cybersecurity Vulnerabilities through Jailbreak Attacks

Topic: Cybersecurity Risks and Multi-Modal Threats **Abstract:** Beyond just generating disallowed content, jailbreak attacks can be leveraged to expose sensitive data or compromise connected services. In this project, you will simulate advanced attack vectors that target both textual and potentially multi-modal systems, aiming to highlight real-world cybersecurity implications. You will investigate how these attacks interface with authentication systems, data storage layers, or other networked components.

Project 7: Defense from Audio Attacks

Topic: Audio-Driven Adversarial Noise and LLM Security **Abstract:** This project explores how malicious audio cues can hijack speech recognition or voice-based interfaces, ultimately affecting LLM outputs. You will replicate scenarios where adversarial noise injects harmful text prompts through automated transcription pipelines. By implementing noise filtering or transformation layers, you will propose a defense strategy that can be integrated into voice-activated LLM systems.

Project 8: Segments Models in Vision Transformers (ViTs)

Topic: Semantic Segmentation and Human-Like Visual Processing **Abstract:** Inspired by how humans parse scenes into meaningful segments rather than individual pixels, this project aims to introduce segment-based processing into Vision Transformer architectures. You will investigate techniques for incorporating semantic or region-based grouping into the model's attention mechanism, possibly integrating with diffusion-based models or multi-modal LLMs, Adversarial Attacks or Defenses.

Project 9: Deepfake Detection

Topic: Identifying AI-Generated Media **Abstract:** As deepfakes become more convincing, reliable detection methods are critical for maintaining trust in digital media. In this project, you will first survey the current landscape of deep-fake detection techniques, ranging from CNN-based image classifiers to transformer-based approaches—and pinpoint their advantages and weaknesses. You will then build a proof-of-concept system, using a curated set of real and AI-generated images or videos, to build a detection model relying on Image Processing features.

Project 11: LLM Optimization via Multi-Next-Tokens Generations

Topic: Adaptive Token Generation and Efficiency Gains **Abstract:** Traditional LLMs output one token at a time, assessing probabilities for each step. In this project, you will add a mechanism that allows the model to generate multiple tokens simultaneously when sufficiently confident. Through quantitative evaluations on coherence, latency, and resource usage, you will determine how effectively this expanded vocabulary approach accelerates text production without compromising clarity or accuracy.