1 Evaluating Unlearning in Modern Large Language Models

Supervisor: Liran Cohen, M.Sc. Student

liranc6@campus.technion.ac.il

1.1 Project Description:

This project offers a deep dive into the fascinating and increasingly critical field of machine unlearning. In today's world, it's vital for AI models to be able to "forget" specific pieces of data they were trained on, whether for privacy, security, or error correction. However, a major challenge is determining whether a model has truly forgotten the information.

In this project, you will be introduced to a new method for evaluating unlearning. The method generates embedding-proximity perturbations by replacing tokens with their neighbors in embedding space and analyzes the resulting Input Loss Landscape (ILL) features as a sensitive indicator of residual memorization. The project will involve applying this evaluation framework to powerful Large Language Models (LLMs) and comparing its effectiveness with existing approaches.

1.2 Project Goals & Research Questions

The objectives of this project are to:

1. **Evaluate on Next-Generation Models:** Apply the proposed unlearning evaluation method to state-of-the-art LLMs such as Llama 3, Mixtral, or other recent models. Understand how unlearning works in the latest models.
2. **Perform Comparative Analysis:** Implement existing evaluation methods from recent research as benchmarks and compare results.
3. **Generate New Insights:** Based on the experiments, write a detailed report of the results. This is the core research component of the project: documenting what you learned about the unlearning process from the given evaluation method.
4. **Ensure Reproducibility:** Develop well-organized and clean code so that all experiments can be easily replicated by others.

1.3 Methodology & Timeline

The project will be structured into three main phases, with a total duration of 12 weeks.

Phase 1: Research & Setup (Weeks 1-3) - Literature Review: The student will read relevant literature to gain a solid understanding of machine unlearning

and related work in the field. - Baselines Research: Review academic literature to find unlearning evaluation methods to use for comparison. - Model Research: Identify and select open-source LLMs based on their suitability for the experiments. - Environment Setup: Set up the necessary software and libraries (like PyTorch and Hugging Face transformers) to create a stable environment for running the experiments.

Phase 2: Experimentation (Weeks 4-8) - Experiment Execution: Apply the proposed evaluation framework to the selected LLMs. - Baseline Implementation: Implement and run the chosen baseline methods.

Phase 3: Analysis & Reporting (Weeks 9-12) - Data Analysis: Analyze experimental results, identify key findings, and create visualizations. - Report Writing: Write a final report covering background, methodology, and conclusions.

1.4 Prerequisites & Learning Outcomes

Prerequisites: - Required: Successful completion of a deep learning course. - Recommended: Strong skills in Python, experience with a deep learning framework (e.g., PyTorch).

Learning Outcomes By completing this project, you will gain: - Hands-On AI Expertise: Practical experience with cutting-edge LLMs and deep learning tools. - Research Competence: Skills in designing, running, and analyzing scientific experiments. - Domain Knowledge: A deep understanding of machine unlearning, an important area in AI safety and ethics. - Technical Proficiency: Enhanced programming and data analysis abilities in Python.