**Thesis Proposal: Investigating the Vulnerability of Private Data in LLMs through Publicly Released Weights**

## Introduction

Large Language Models (LLMs) are trained on vast datasets and are computationally expensive [1], often containing sensitive, private information that is not publicly available. Once training is complete, the weights of these models are frequently made open source for public use. This raises significant privacy concerns [2]: if the weights of an LLM are publicly accessible, could they be used to retrace the private training data? My research aims to explore this critical privacy issue by determining whether it is possible to reconstruct private training data from LLM weights alone.

## Background and Significance

As LLMs become more widely adopted, their open-source nature presents a potential privacy vulnerability [3]. Although the weights of the models are anonymized and detached from their original training data, there is a theoretical risk that skilled individuals could reverse-engineer these weights to recover sensitive, private information. This issue is of immense importance in fields such as healthcare, finance, and law, where data privacy is paramount. Addressing this question is crucial for ensuring that LLMs can be used safely without compromising the privacy of individuals whose data was included in the training process.

## Objectives

The primary objective of this research is to investigate the possibility of retracing private data from the publicly released weights of an LLM. Specifically, the research will focus on three fundamental questions:

1. Can users retrace private training data from an LLM's weights alone?
2. If so, what specific conditions or requirements must the user satisfy to accomplish this?
3. What safeguards can be implemented to prevent users from accessing private data through LLM weights?

If the research finds that retracing private training data is not possible, questions two and three may become irrelevant, but the investigation will still provide valuable insights into the privacy implications of LLMs.

## Methods

The research will involve a combination of theoretical analysis and practical experimentation. First, a review of the current state of LLM weight extraction methods will be conducted to identify potential vulnerabilities. Then, various attack vectors will be tested on an open-source LLM to determine if private training data can be reconstructed from its weights. These experiments will be complemented by an investigation into the computational and technical requirements necessary to execute such attacks. Finally, if retracing is found to be feasible, the research will explore potential safeguards, such as differential privacy techniques or encryption, to mitigate these risks.

## Organization

The research will be divided into three phases: (1) a theoretical review and vulnerability assessment, (2) experimentation and testing of attack methods on LLMs, and (3) development of potential safeguards, if necessary. The timeline will span over the course of the thesis period, with key milestones at the end of each phase for review and adjustment.

**Citations**

[1] Sachdeva, Noveen, et al. "How to Train Data-Efficient LLMs." arXiv preprint arXiv:2402.09668 (2024).

[2] Choquet, Géraud, Aimée Aizier, and Gwenaëlle Bernollin. "Exploiting Privacy Vulnerabilities in Open Source LLMs Using Maliciously Crafted Prompts." (2024).

[3] Dunlap, Trevor, et al. "Pairing Security Advisories with Vulnerable Functions Using Open-Source LLMs." *International Conference on Detection of Intrusions and Malware, and Vulnerability Assessment*. Cham: Springer Nature Switzerland, 2024.