**Section 1 : Topic Submission Form**

This form should be submitted by the mentioned deadline.

Name:           Deepak Goyal

Student Number:  **1180979**

Course:              [LJMU Masters in Machine Learning and AI](https://learn.upgrad.com/)\_\_\_\_ \_\_\_\_\_\_\_\_\_

**Fill your topic/s below**

Project Title/Area 1:    Redefining Code Security: Harnessing Large Language Models for Enhanced Vulnerability Detection                                                                        \_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Dataset:

[NIST Software Assurance Reference Dataset](https://samate.nist.gov/SARD/)

  (The dataset provides vulnerable code examples that serve as my baseline. I will employ various prompting methodologies with LLMs to generate corresponding code, then conduct comparative analysis to determine which prompting approaches yield code with reduced security vulnerabilities.)                                                       \_\_ \_

Description:     Explore various Techniques of prompt engineering to produce a code that is less vulnerable to security attacks, In this study I will explore various LLM models and how do they perform to produce a secure code with various prompting techniques like Tip setting, Chain of Thought Step-by Step etc and suggest \_a novel prompt technique to produce less vulnerable code.\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Project Title/Area 2:   Protecting PII Data in LLM’s by using Encoding/ introducing noise at client end                                                                               \_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Dataset:             [PII | External Dataset](https://www.kaggle.com/datasets/alejopaullier/pii-external-dataset)

(The dataset includes extensive samples containing personally identifiable information (PII) such as names, contact details, and addresses. Our evaluation will assess whether the Split-Privacy Architecture effectively prevents sensitive data from reaching the LLM server).

                                                                \_\_ \_

Description:        Split-Privacy Architecture - Divide LLM processing between client and server with sensitive computations happening locally at client side, Develop hybrid architectures where sensitive tokens never leave user devices, Research area: Creating efficient partitioning algorithm that minimize performance impact and still protect PII data bu masking or converting at client end.                                                                    \_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Project Title/Area 3:    **Bias Mitigation in Generative Models:** Study techniques to reduce biases in generative outputs, ensuring fairness and inclusivity.                                                                              \_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Dataset:                  [Adult - UCI Machine Learning Repository](https://archive.ics.uci.edu/dataset/2/adult)

       (This dataset is used to predict the salary of a person (below 50K$or above 50k$). The dataset contains two Sensitive Attributes (SA), i.e., Ethnicity and Gender . This dataset has a bias for Caucasians and males while predicting income, so we will use some models/techniques to augment the dataset to reduce the bias.)                                               \_\_ \_

Description:         Present a framework for mitigating biases, e.g. Prompt Engineering for datatset augmentation(bias-aware prompts), by using a technique of adversarial learning i.e. including a variable for the group of interest and simultaneously learning a predictor and an adversary  . Create a framework that:

* Detects bias in generated synthetic data
* Adjusts augmentation parameters
* Re-generates data until bias metrics improve

               \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**Fill in this section if a member of staff has agreed to be your supervisor:**

Member of Staff:     Sushrut Dilip Shendre                                                                           \_\_\_\_

If you have found a supervisor then you and the member of staff who agreed to supervise your project should sign below.

\_\_Deepak Goyal\_                                                                 Sushrut Shendre      \_\_\_\_\_\_\_\_\_

Student Signature                                                                         Supervisor Signature

\_1st July\_\_                                                                            1st July\_\_\_\_\_\_\_\_\_\_\_\_

Date                                                                                               Date

**Section 2 : Topic Selection Research**

**Table 1 : Topic 1(**Redefining Code Security: Harnessing Large Language Models for Enhanced Vulnerability Detection )

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Title** | **Link to the Paper** | **Understanding of the Dataset** | **Understanding the Methodology Used** | **Dataset Link** |
| Prompting Techniques for Secure Code Generation: A Systematic Investigation | [Prompting Techniques for Secure Code Generation: A Systematic Investigation](https://dl.acm.org/doi/pdf/10.1145/3722108) | It contains 150 natural language prompts that can be leveraged for assessing the security performance of large language models. These prompts are natural language descriptions of code snippets prone to various security vulnerabilities listed in MITRE's Top 25 Common Weakness Enumeration | A systematic literature review to identify prompting techniques and an in-depth security evaluation of a subset of these techniques | [c01dsnap/LLM-Sec-Evaluation · Datasets at Hugging Face](https://huggingface.co/datasets/c01dsnap/LLM-Sec-Evaluation) |
| Enhancing Software Code Vulnerability Detection Using GPT-4o and Claude-3.5 Sonnet: A Study on Prompt Engineering Techniques | [Enhancing Software Code Vulnerability Detection Using GPT-4o and Claude-3.5 Sonnet: A Study on PromptEngineering Techniques](https://www.researchgate.net/publication/382099568_Enhancing_Software_Code_Vulnerability_Detection_Using_GPT-4o_and_Claude-35_Sonnet_A_Study_on_Prompt_Engineering_Techniques/fulltext/668d4d97af9e615a15d8d39d/Enhancing-Software-Code-Vulnerability-Detection-Using-GPT-4o-and-Claude-35-Sonnet-A-Study-on-Prompt-Engineering-Techniques.pdf) | The Software Assurance Reference Dataset (SARD) is a growing collection of test programs with documented weaknesses. Test cases vary from small synthetic programs to large applications. The programs are in C, C++, Java, PHP, and C#, and cover over 150 classes of weaknesses | This focuses on evaluating the effectiveness of advanced large language models (LLMs) in detecting software vulnerabilities using various prompt engineering techniques | [NIST Software Assurance Reference Dataset](https://samate.nist.gov/SARD/) |
| Principled Instructions Are All You Need for Questioning LLaMA-1/2, GPT-3.5/4 | [2312.16171](https://arxiv.org/pdf/2312.16171) | This dataset, comprised of 13k data points, supports the study of LLM prompting principles. The data is curated to facilitate understanding and application of the [26 principles](https://github.com/VILA-Lab/ATLAS/blob/main/data/README.md) | The evaluation of large language models (LLMs) for code generation extensively assesses their functional correctness, using specialized datasets like HumanEval and APPS with the pass@k metric1.... Prompt engineering techniques, such as "Step-by-Step" and "Recursive Criticism and Improvement," are crucial, significantly enhancing LLM performance in both code generation and vulnerability detection accuracy | [GitHub - VILA-Lab/ATLAS: A principled instruction benchmark on formulating effective queries and prompts for large language models (LLMs). Our paper: https://arxiv.org/abs/2312.16171](https://github.com/VILA-Lab/ATLAS) |
| Evaluating Large Language Models Trained on Code | [Evaluating Large Language Models Trained on Code](https://arxiv.org/pdf/2107.03374) | The HumanEval dataset released by OpenAI includes 164 programming problems with a function sig- nature, docstring, body, and several unit tests. They were handwritten to ensure not to be included in the training set of code generation models. | In this study, Author generated independent Python functions from descriptions written in natural language (docstrings). To check if the generated code is correct, we run automatic unit tests instead of relying on human evaluation, which is common in text generation tasks | [openai/openai\_humaneval · Datasets at Hugging Face](https://huggingface.co/datasets/openai/openai_humaneval) |
| Is Your AI-Generated Code Really Secure? Evaluating Large Language Models on Secure Code Generation with CodeSecEval | https://arxiv.org/pdf/2407.02395v1 | This paper uses CodeSecEval dataset. It includes a broad spectrum of critical vulnerability types and provides detailed attributes for each data instance, enabling precise automatic evaluations. By utilizing CodeSecEval, we aim to more accurately investigate the capabilities of state-of-the-art LLMs in code generation and repair, while also proposing effective strategies to enhance security in both tasks | Comprehensive study aimed at precisely evaluating and enhancing the security aspects of code LLMs | [XuanwuAI/SecEval · Datasets at Hugging Face](https://huggingface.co/datasets/XuanwuAI/SecEval) |
|  |  |  |  |  |

**Table 2 : Topic 2(**Protecting PII Data in LLM’s by using Encoding/ introducing noise at client end)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Title** | **Link to the Paper** | **Understanding of the Dataset** | **Understanding the Methodology Used** | **Dataset Link** |
| Split-and-Denoise: Protect Large Language Model Inference with Local Differential Privacy | [Split-and-Denoise: Protect Large Language Model Inference with Local Differential Privacy](https://arxiv.org/html/2310.09130v4#A1.SS3) | Various Datasets are used in this paper that includes Customer Queries for a Bank and Amazon Reviews revealing some of the PII information | This paper introduces Split-N-Denoise (SnD), an private inference framework that splits the model to execute the token embedding layer on the client side at minimal computational cost. | [PolyAI/banking77 · Datasets at Hugging Face](https://huggingface.co/datasets/PolyAI/banking77)  [takala/financial\_phrasebank · Datasets at Hugging Face](https://huggingface.co/datasets/takala/financial_phrasebank)  [mteb/amazon\_polarity · Datasets at Hugging Face](https://huggingface.co/datasets/mteb/amazon_polarity) |
| Hide and Seek (HaS): A Lightweight Framework for Prompt Privacy Protection | [[2309.03057] Hide and Seek (HaS): A Lightweight Framework for Prompt Privacy Protection](https://ar5iv.labs.arxiv.org/html/2309.03057#:~:text=To%20address%20this%20challenge%2C%20we%20propose%20a%20novel,for%20anonymization%20and%20seeking%20private%20entities%20for%20de-anonymization.) | Uses Synthetic dataset as well as BBC News classification dataset | In this paper, Author expand the application scenarios of anonymization techniques by training a small local model to de-anonymize the LLM’s returned results with minimal computational overhead. | [BBC News Classification | Kaggle](https://www.kaggle.com/c/learn-ai-bbc/data) |
| A privacy-preserving algorithm for distributed training of neural network ensembles | [A privacy-preserving algorithm for distributed training of neural network ensembles | Neural Computing and Applications](https://link.springer.com/article/10.1007/s00521-012-1000-8) |  | In this paper, the author studies the privacy protection in distributed training of neural network ensembles. We design a privacy-preserving distributed algorithm for training neural network ensembles using AdaBoost.M2. The author also analyses the security and complexity of his algorithm. |  |
| Privacy-Preserving Backpropagation Neural Network Learning | [Privacy-Preserving Backpropagation Neural Network Learning | IEEE Journals & Magazine | IEEE Xplore](https://ieeexplore.ieee.org/abstract/document/5223520) | This dataset is originally from the National Institute of Diabetes and Digestive and Kidney Diseases. The objective of the dataset is to diagnostically predict whether or not a patient has diabetes, based on certain diagnostic measurements included in the dataset. | In this paper, Author present a privacy-preserving algorithm for backpropagation neural network learning. The algorithm guarantees privacy in a standard cryptographic model, the semihonest model. | [Pima Indians Diabetes Database](https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database) |
|  |  |  |  |  |
|  |  |  |  |  |

**Table 3 : Topic 3(Bias Mitigation in Generative Models:** Study techniques to reduce biases in generative outputs, ensuring fairness and inclusivity.)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Title** | **Link to the Paper** | **Understanding of the Dataset** | **Understanding the Methodology Used** | **Dataset Link** |
| Bias Mitigation for Machine Learning Classifiers: A Comprehensive Survey | [Bias Mitigation for Machine Learning Classifiers: A Comprehensive Survey | ACM Journal on Responsible Computing](https://dl.acm.org/doi/10.1145/3631326) | Predict whether annual income of an individual exceeds $50K/yr based on census data. Also known as "Census Income" dataset. Dataset has Bias for Gender and Enthnicity.  Identify patients who will be admitted to a hospital within the next year using historical claims data. | Author provides a comprehensive overview of the research on bias mitigation methods for ML classifiers;  Paper introduces the experimental design details for evaluating existing bias mitigation methods | [Adult - UCI Machine Learning Repository](https://archive.ics.uci.edu/dataset/2/adult)  [Medical Expenditure Panel Survey Public Use File Details](https://meps.ahrq.gov/mepsweb/data_stats/download_data_files_detail.jsp?cboPufNumber=HC-192)  [Heritage Health Prize | Kaggle](https://www.kaggle.com/c/hhp) |
|  |  |  |  |  |
| Targeted Data Augmentation for bias mitigation | [2308.11386](https://arxiv.org/pdf/2308.11386) | The dataset contains 33,126 dermoscopic training images of unique benign and malignant skin lesions from over 2,000 patients. It has Bias such as hair, frames, rulers, pen marks, or gel drops, are present in this dataset | In this study, Author introduces a novel and efficient approach for addressing biases called Targeted Data Augmentation (TDA), which leverages classical data augmentation techniques to tackle the pressing issue of bias in data and models. Unlike the laborious task of removing biases, our method proposes to insert bi ases instead, resulting in improved performance. | [SIIM-ISIC-2020](https://www.kaggle.com/datasets/prashantjeswani/siimisic2020) |
| Generative Adversarial Networks for Mitigating Biases in Machine Learning Systems | [1905.09972](https://arxiv.org/pdf/1905.09972) | Predict whether annual income of an individual exceeds $50K/yr based on census data. Also known as "Census Income" dataset | The proposed framework is based on conditional Generative Adversarial Networks (cGANs), which are used to generate new synthetic fair data with selec tive properties from the original data. | [Adult - UCI Machine Learning Repository](https://archive.ics.uci.edu/dataset/2/adult) |
| Mitigating Unwanted Biases with Adversarial Learning | [1801.07593](https://arxiv.org/pdf/1801.07593) | Predict whether annual income of an individual exceeds $50K/yr based on census data. Also known as "Census Income" dataset | The author demonstrated a general and powerful method for training unbiased machine learning models. | [Adult - UCI Machine Learning Repository](https://archive.ics.uci.edu/dataset/2/adult) |
| Unbiased-Diff: Analyzing and Mitigating Biases in Diffusion Model-Based Face Image Generation | [IEEE Xplore Full-Text PDF:](https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=10820122) | FairFace is a face image dataset which is race balanced. It contains 108,501 images from 7 different race groups: White, Black, Indian, East Asian, Southeast Asian, Middle Eastern, and Latino. Images were collected from the YFCC-100M Flickr dataset and labeled with race, gender, and age groups. | The author analyzed the presence of social biases in diffusion-based face generations and propose a novel sampling process guidance algorithm to mitigate these biases. Specifically, during the diffusion sampling process, we guide the generation to produce samples with attribute distributions that align with a balanced or desired attribute distribution. | [HuggingFaceM4/FairFace · Datasets at Hugging Face](https://huggingface.co/datasets/HuggingFaceM4/FairFace) |
| FAIRTL:A Transfer Learning Approach for Bias Mitigation in Deep Generative Models | [IEEE Xplore Full-Text PDF:](https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=10436577) | Flickr-Faces-HQ (FFHQ) is a high-quality image dataset of human faces, originally created as a benchmark for generative adversarial networks (GAN).The dataset consists of 70,000 high-quality PNG images at 1024×1024 resolution and contains considerable variation in terms of age, ethnicity and image background. | To address the biases, the author's main contribution is to propose novel methods to learn fair generative models via transfer learning. Specifically, first, we propose FAIRTL where wepre-train the generative model with a large biased dataset | [student/FFHQ · Datasets at Hugging Face](https://huggingface.co/datasets/student/FFHQ) |
|  |  |  |  |  |