Subsalt MIDS Kick-off

September 5th, 2024

The Team from Duke

Vijay Keswani (Mentor)

Post-Doc in AI Ethics

Simrun Sharma

Pre-med|Economics|Python|Gen AI

Antara Bhide

Law | Legaltech | Data Science

Minling Zhou

ML | SOAR | SIEM | CCSK

Problem Statement

Create predictive model(s) that reduce the compute cost of running Subsalt’s system by identifying which model configurations are likely to produce high-quality, private data for a given dataset before starting the training process

Project Goals

Goal 1

Model that can produce a score for each model configuration on an arbitrary dataset schema

Goal 2

Model that predicts the likelihood that synthetic data produced by a specific model configuration on a dataset schema will pass a known series of privacy tests.

Goal 3

Component (model, API service, etc) that can use the outputs from Goals 1 and 2 to produce a list of model configurations to train given a specific budget (either wall clock time or compute cost)

Understanding the problem

Questions

- Simplified overview - case study client
- Reduce computation cost
- Data structure/size, training process
- Privacy standard testing
- Current model para/eva
- High-quality, private data measurement
- Current optimization technique, model type

Onboarding

- Data classification
- Data/Model access
- Privacy regulation awareness
- Development environment
- Synthetic data life cycle/workflow

Expectation

- Benchmarks - Timeline
- Deliverables