TASK 1: Membership Inference Attack

Franziska Boenisch & Adam Dziedzic

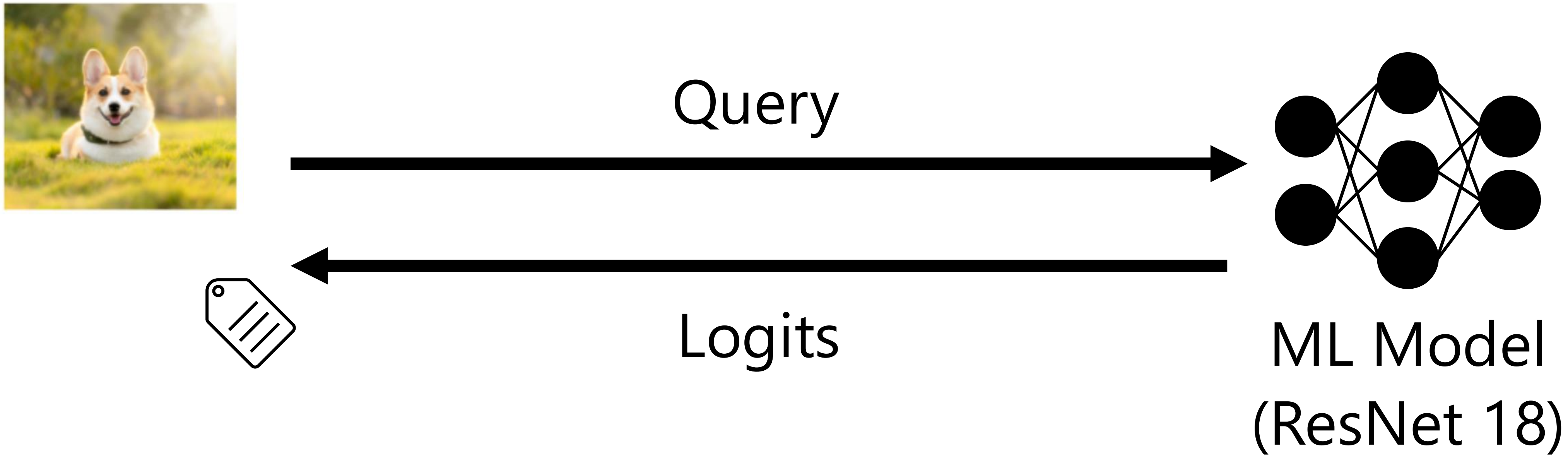
April 30th 2025

CISPA HELMHOLTZ CENTER FOR INFORMATION SECURITY

SprintML

Task 1: Membership Inference Attack

Was the image used to train the ML Model?

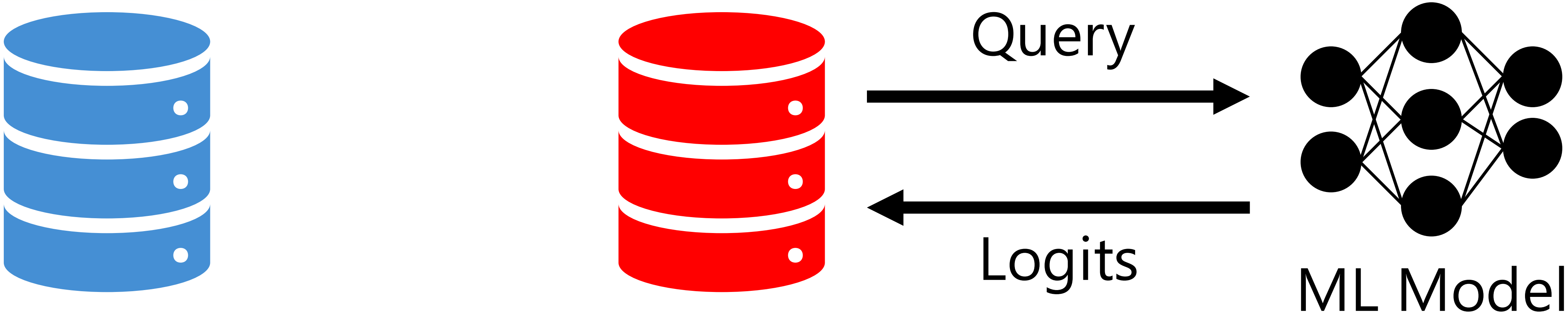


Query

Logits

ML Model
(ResNet 18)

Task 1: Membership Inference Attack



Public Dataset

Private Dataset

ML Model (Resnet18)

Id:1 image: <bits> label: 1, member: 1

Id:2 image: <bits> label: 2, member: 0

• • •

Id:n image: <bits> label: 4, member: 1

Id:1 image: <bits> label: 1, member: ?

Id:2 image: <bits> label: 2, member: ?

• • •

Id:n image: <bits> label: 4, member: ?

Task 1: Membership Inference Attack

Public Dataset

Id:1 image: <bits> label: 1, member: 1

Id:2 image: <bits> label: 2, member: 0

• • •

Id:n image: <bits> label: 4, member: 1

Private Dataset

Id:1 image: <bits> label: 1, member: ?

Id:2 image: <bits> label: 2, member: ?

• • •

Id:n image: <bits> label: 4, member: ?

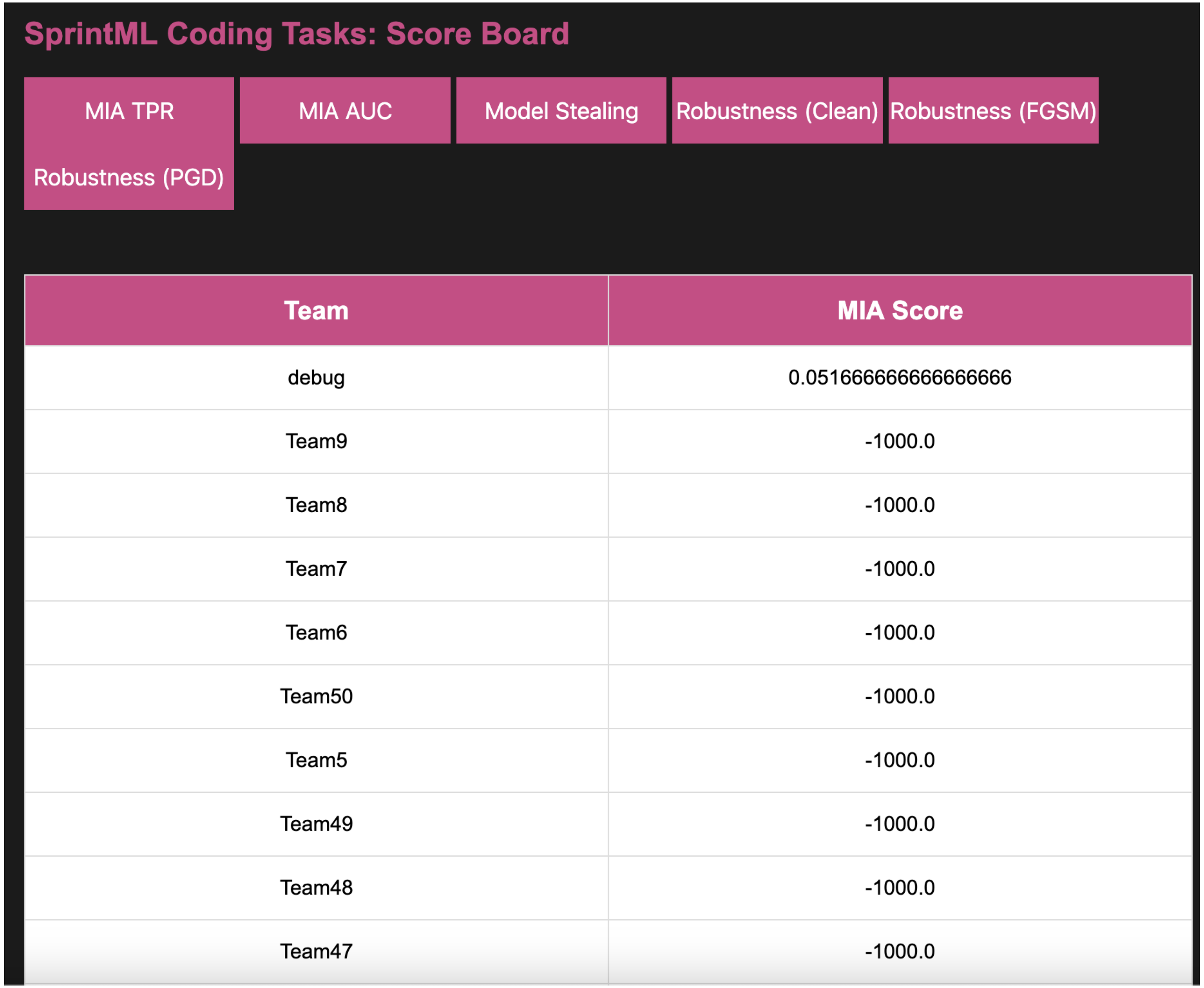
GOAL: Assign Continuous Membership Scores in range [0, 1].

Task 1: Membership Inference Attack

As an attacker, you get access to:

- Resnet18 model **trained on an *undisclosed* dataset**
- Public dataset that contains the **ids, images, class labels, and membership information** (1 == is member, 0 == not a member). Members are data points that were used in the training dataset to train the model, and non-members were not
- Private dataset that contains the **ids, images, and class labels, with membership set to None**, which means *the membership is unknown*. Note that some of the data points in this dataset were used to train the model (members) and some were not (non-members).

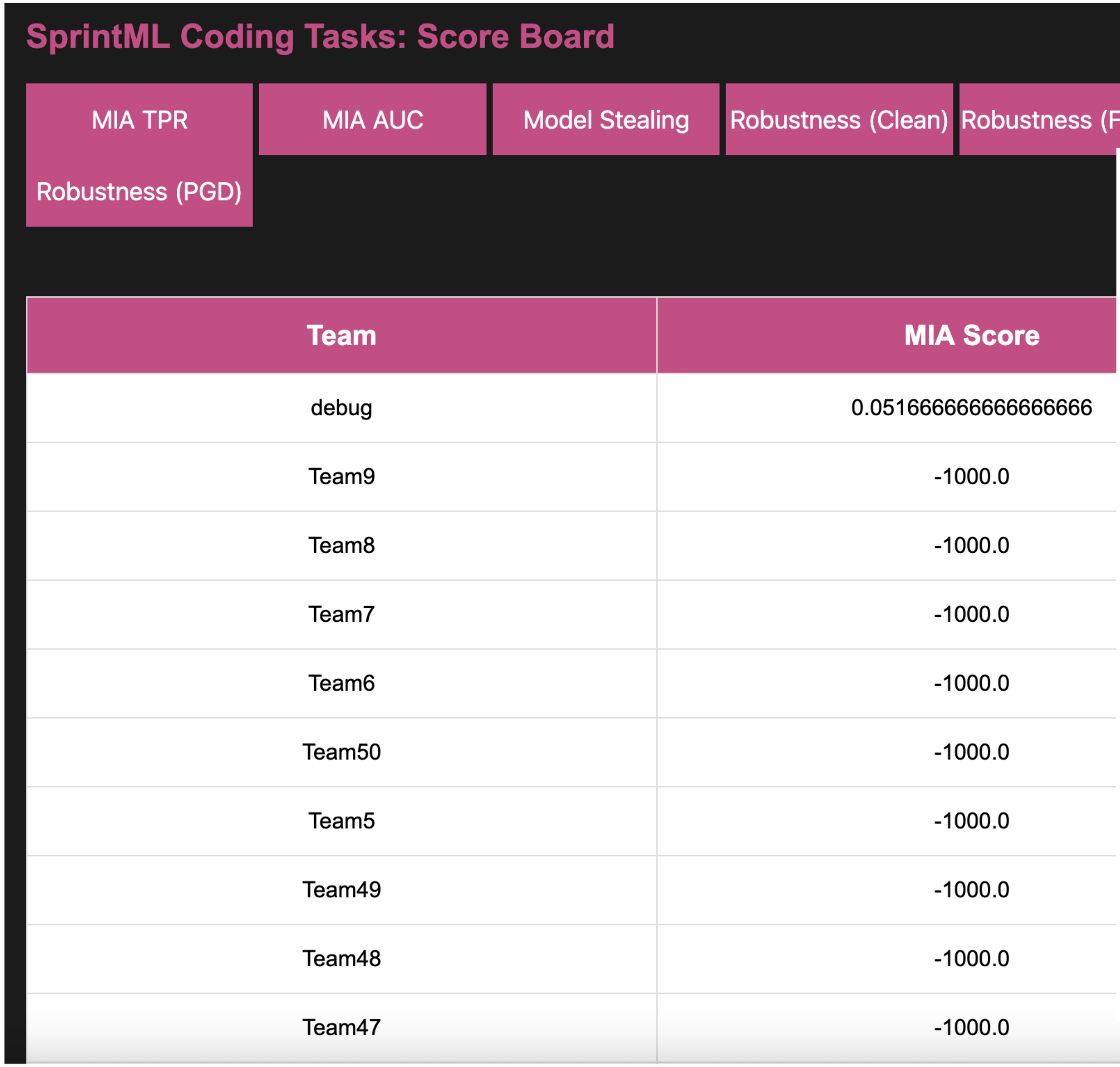
Task 1: Evaluation TPR @ low FPR



SprintML Coding Tasks: Score Board

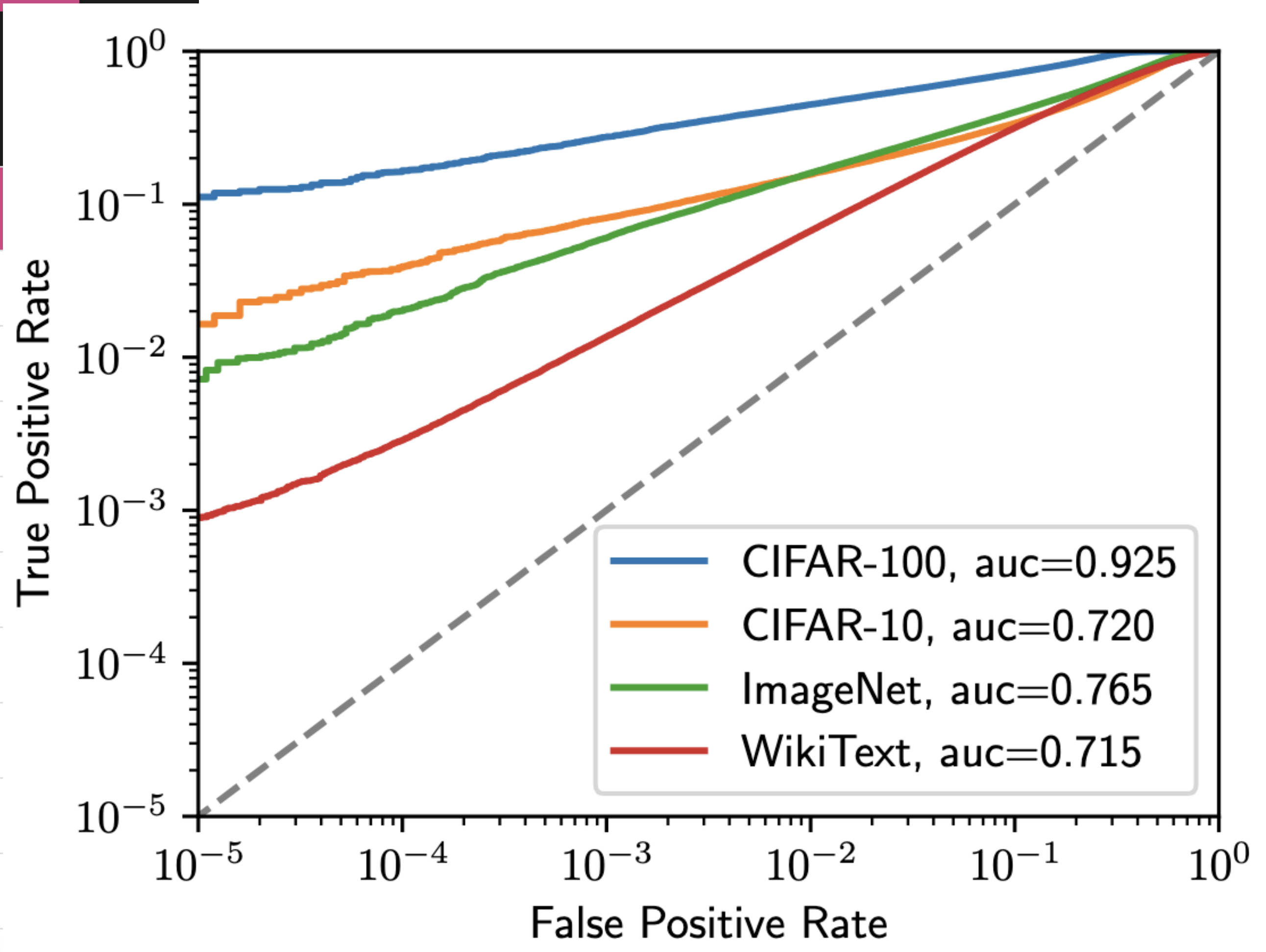
Team	MIA Score
debug	0.051666666666666666
Team9	-1000.0
Team8	-1000.0
Team7	-1000.0
Team6	-1000.0
Team50	-1000.0
Team5	-1000.0
Team49	-1000.0
Team48	-1000.0
Team47	-1000.0

Task 1: Evaluation TPR @ low FPR

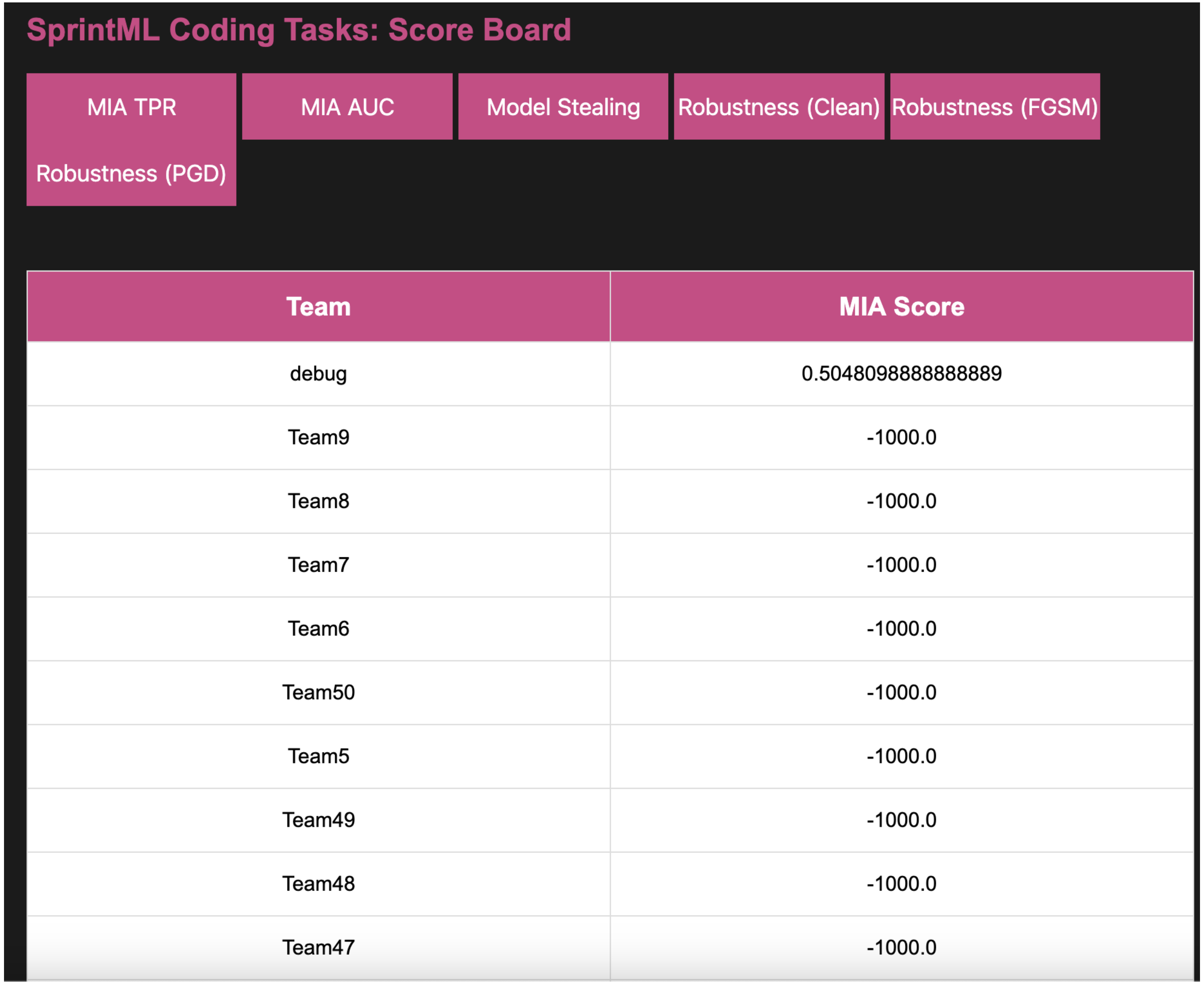


SprintML Coding Tasks: Score Board

Team	MIA Score
debug	0.05166666666666666
Team9	-1000.0
Team8	-1000.0
Team7	-1000.0
Team6	-1000.0
Team50	-1000.0
Team5	-1000.0
Team49	-1000.0
Team48	-1000.0
Team47	-1000.0



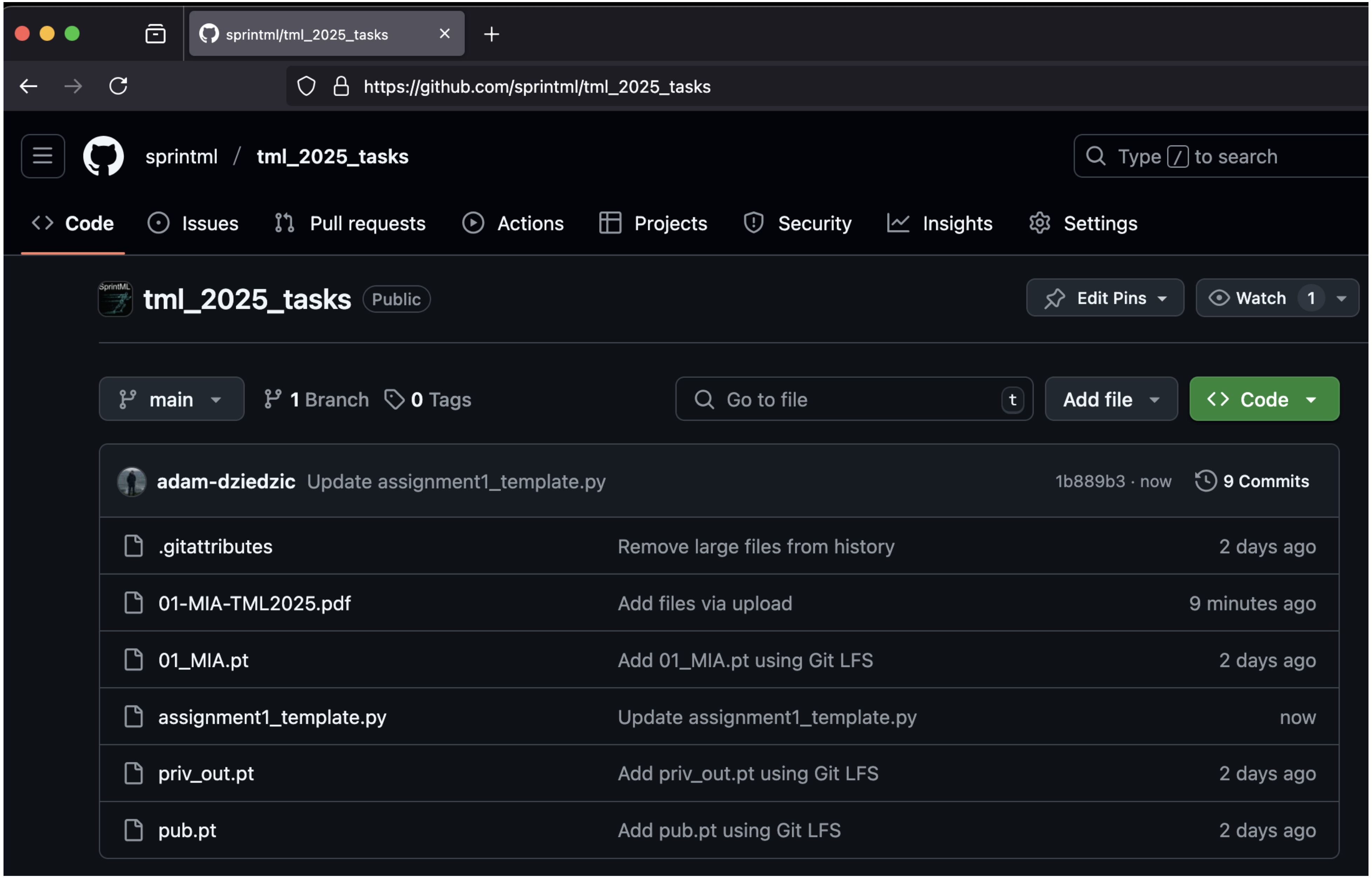
Task 1: Evaluation AUC



SprintML Coding Tasks: Score Board

Team	MIA Score
debug	0.5048098888888889
Team9	-1000.0
Team8	-1000.0
Team7	-1000.0
Team6	-1000.0
Team50	-1000.0
Team5	-1000.0
Team49	-1000.0
Team48	-1000.0
Team47	-1000.0

Task 1: Membership Inference Attack



adam-dziedzic Update assignment1_template.py 1b889b3 · now 9 Commits

.gitattributes	Remove large files from history	2 days ago
01-MIA-TML2025.pdf	Add files via upload	9 minutes ago
01_MIA.pt	Add 01_MIA.pt using Git LFS	2 days ago
assignment1_template.py	Update assignment1_template.py	now
priv_out.pt	Add priv_out.pt using Git LFS	2 days ago
pub.pt	Add pub.pt using Git LFS	2 days ago

Task 1: Membership Inference Attack

1. Deadline: May 28th, 2025
2. 10% of the grade for each Assignment
3. Work in pairs
4. Register your team to obtain the token
5. Link to the repository for the tasks:
   https://github.com/sprintml/tml_2025_tasks

Questions?