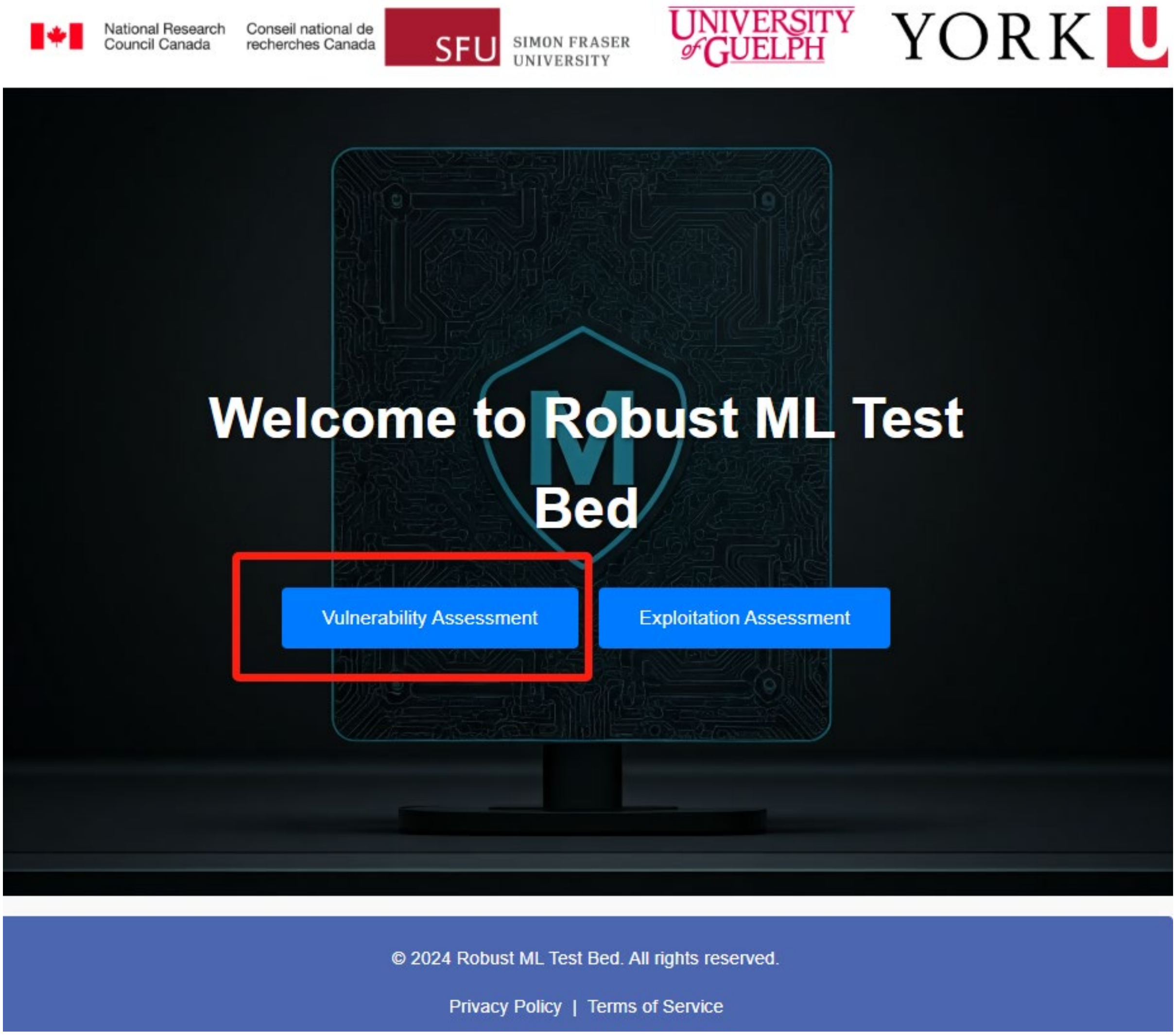
Robust ML Testbed

User Guide

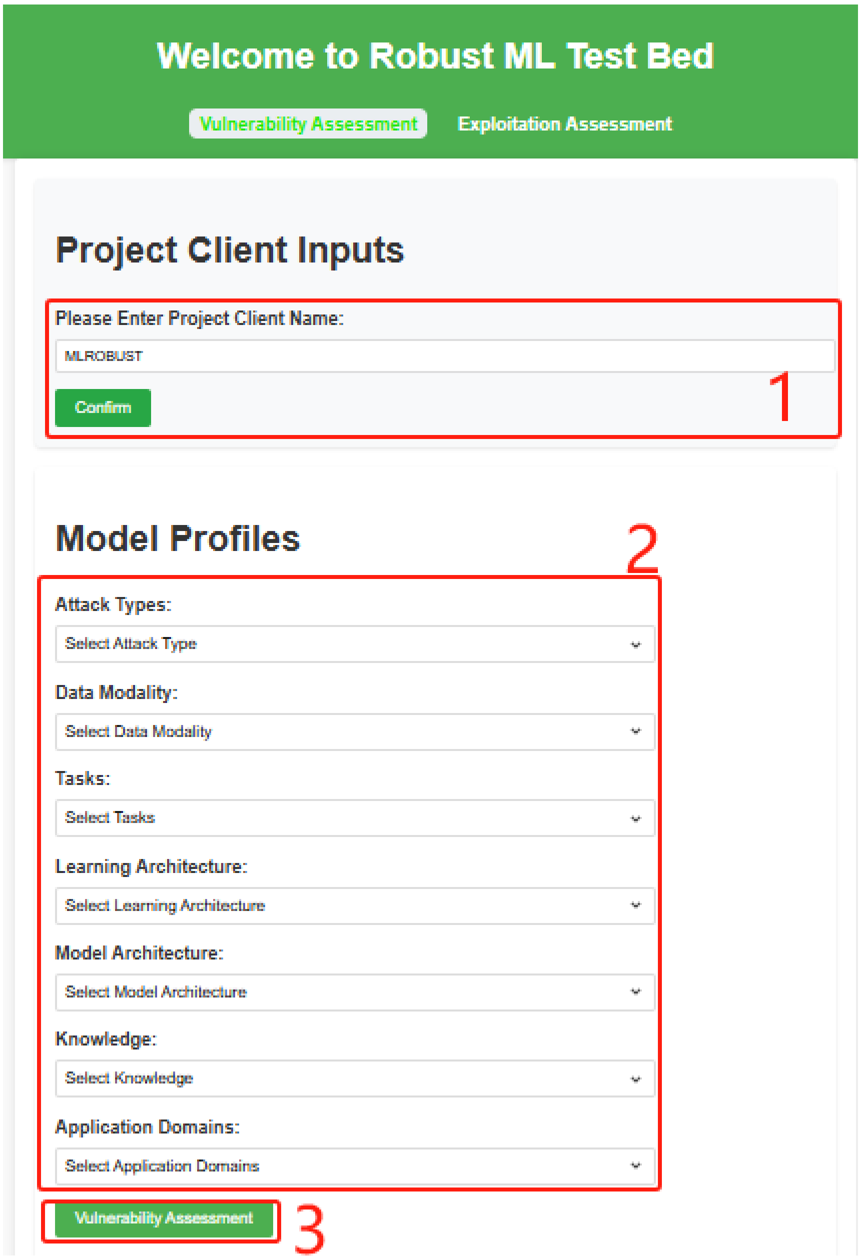
1. Access the website: open a browser, input the address: http://132.246.129.156:3000/, then you will enter the following website.


National Research Council Canada
Conseil national de recherches Canada
SFU SIMON FRASER UNIVERSITY
UNIVERSITY of GUELPH
YORK U
Vulnerability Assessment
Exploitation Assessment
Privacy Policy | Terms of Service

Welcome to Robust ML Test Bed

© 2024 Robust ML Test Bed. All rights reserved.

2. Click the “Vulnerability Assessment” Button, then enter the “Vulnerability Assessment”Page.



Welcome to Robust ML Test Bed

Project Client Inputs

Please Enter Project Client Name: MLROBUST

Model Profiles

Attack Types: Select Attack Type

Data Modality: Select Data Modality

Tasks: Select Tasks

Learning Architecture: Select Learning Architecture

Model Architecture: Select Model Architecture

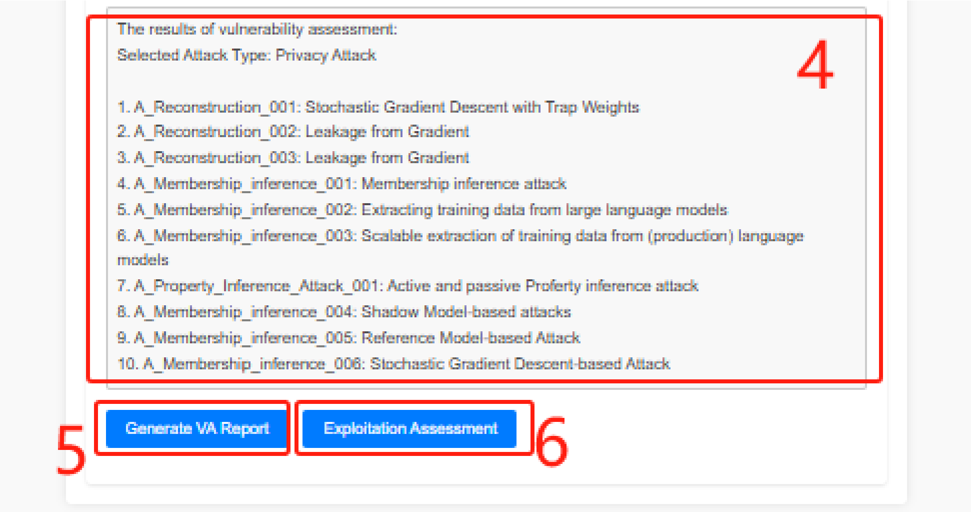
Knowledge: Select Knowledge

Application Domains: Select Application Domains

3 **Step 1:** Input the "Project client name" and press the "confirm" button;

Step 2: Select the Model Profiles based on the following options(including Attack types, data modality, tasks, learning architecture, model architecture, knowledge, application domains);

Step 3: Click the "vulnerability assessment" button,

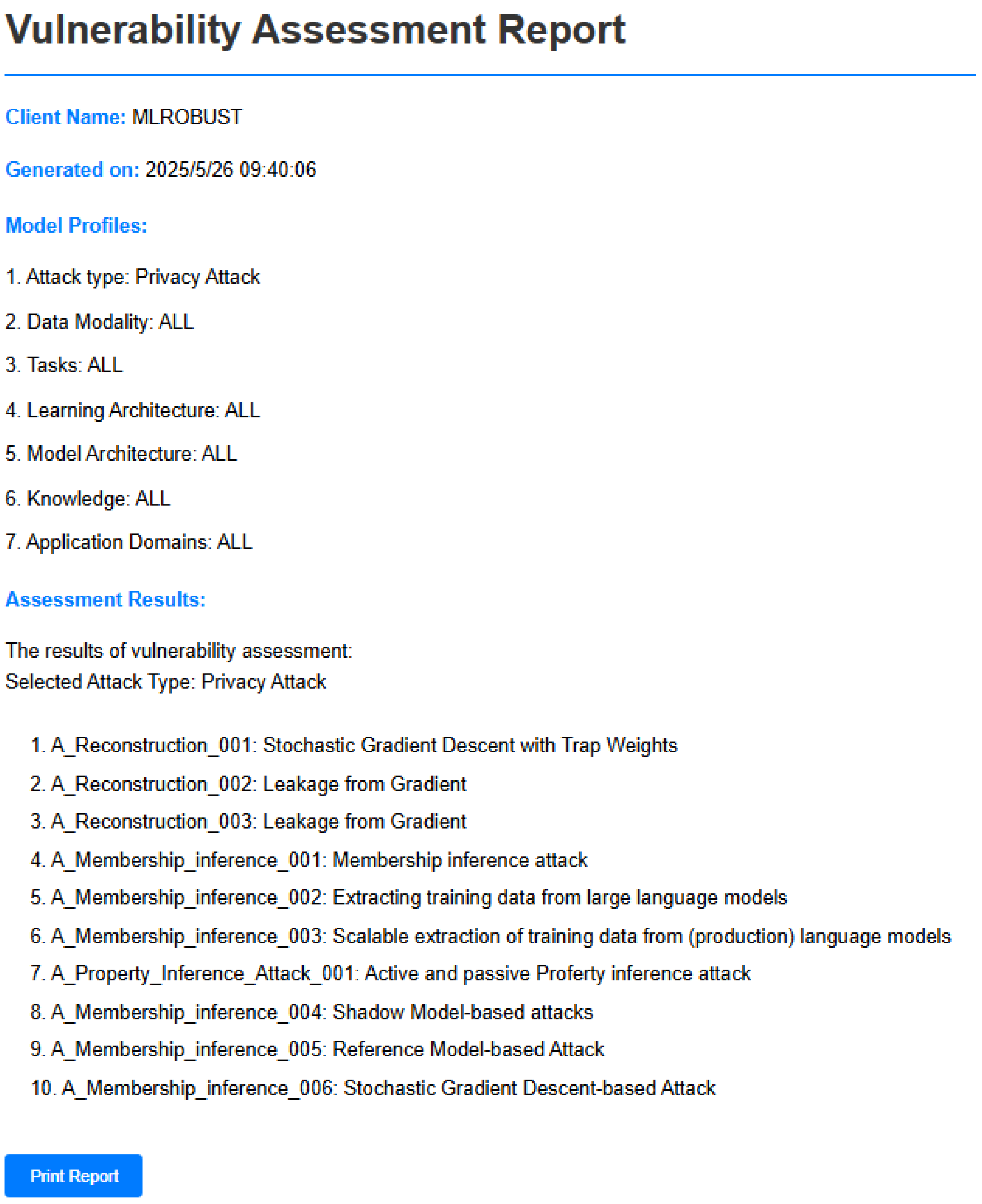


The results of vulnerability assessment:
Selected Attack Type: Privacy Attack

1. A_Reconstruction_001: Stochastic Gradient Descent with Trap Weights
2. A_Reconstruction_002: Leakage from Gradient
3. A_Reconstruction_003: Leakage from Gradient
4. A_Membership_inference_001: Membership inference attack
5. A_Membership_inference_002: Extracting training data from large language models
6. A_Membership_inference_003: Scalable extraction of training data from (production) language models
7. A_Property_Inference_Attack_001: Active and passive Proferty inference attack
8. A_Membership_inference_004: Shadow Model-based attacks
9. A_Membership_inference_005: Reference Model-based Attack
10. A_Membership_inference_006: Stochastic Gradient Descent-based Attack

Step 4: Then you will get the results in Area 4.

Step 5: If you want to save the above information, then you can click the "Generate VA Report" button, and get a report which can be printed.



Vulnerability Assessment Report

Client Name: MLROBUST

Generated on: 2025/5/26 09:40:06

Model Profiles:

1. Attack type: Privacy Attack
2. Data Modality: ALL
3. Tasks: ALL
4. Learning Architecture: ALL
5. Model Architecture: ALL
6. Knowledge: ALL
7. Application Domains: ALL

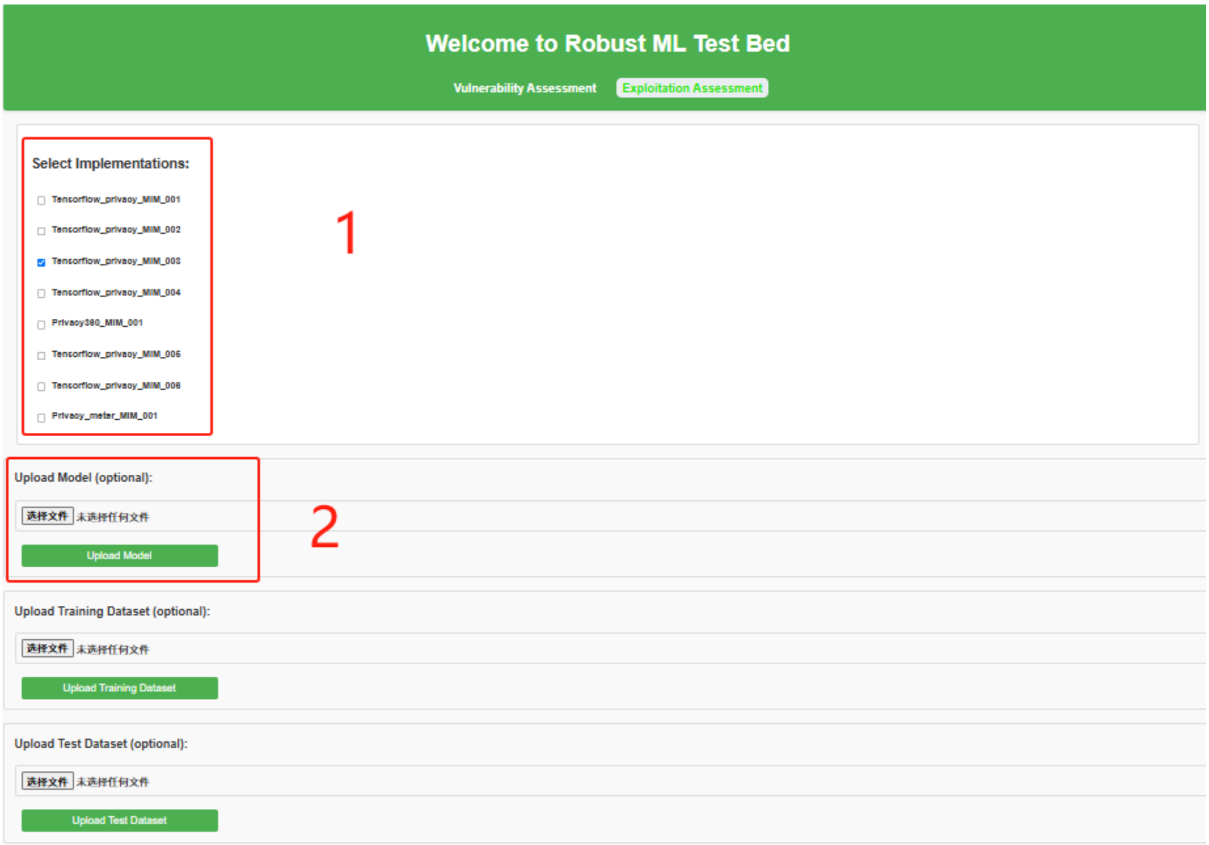
Assessment Results:

The results of vulnerability assessment:
Selected Attack Type: Privacy Attack

1. A_Reconstruction_001: Stochastic Gradient Descent with Trap Weights
2. A_Reconstruction_002: Leakage from Gradient
3. A_Reconstruction_003: Leakage from Gradient
4. A_Membership_inference_001: Membership inference attack
5. A_Membership_inference_002: Extracting training data from large language models
6. A_Membership_inference_003: Scalable extraction of training data from (production) language models
7. A_Property_Inference_Attack_001: Active and passive Proferty inference attack
8. A_Membership_inference_004: Shadow Model-based attacks
9. A_Membership_inference_005: Reference Model-based Attack
10. A_Membership_inference_006: Stochastic Gradient Descent-based Attack

Step 6: Click the “Exploitation Assessment” button, then go to the further analysis page.

4. At the Exploitation Assessment Page:



Select Implementations:

- □ Tensorflow_privacy_MIM_001
- □ Tensorflow_privacy_MIM_002
- ☑ Tensorflow_privacy_MIM_003
- □ Tensorflow_privacy_MIM_004
- □ Privacy380_MIM_001
- □ Tensorflow_privacy_MIM_006
- □ Tensorflow_privacy_MIM_008
- □ Privacy_meter_MIM_001

Upload Model (optional): 选择文件 未选择任何文件

Upload Training Dataset (optional): 选择文件 未选择任何文件

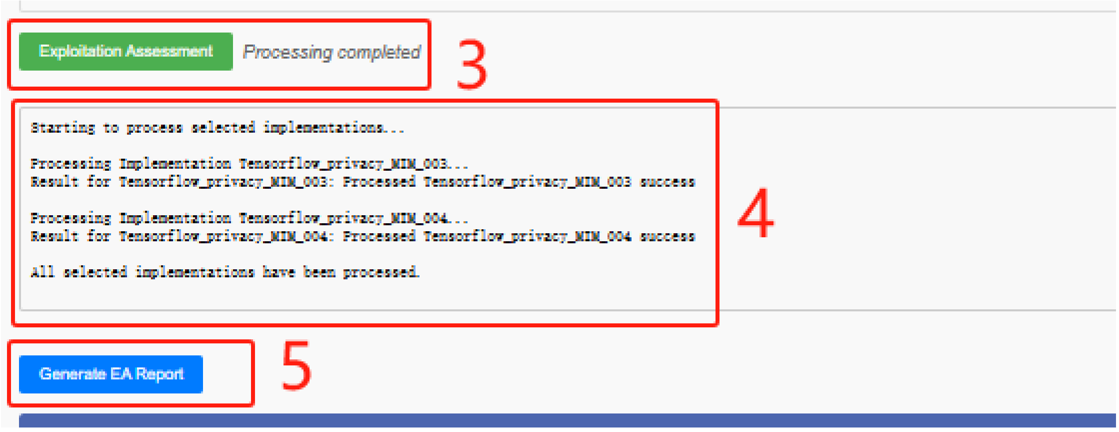
Upload Test Dataset (optional): 选择文件 未选择任何文件

Step 1: Select implementation in this area (you can choose one or more options)

Step 2: Select the upload model from the libraries following the requirements below

Attack types	**Implementation ID**	**Model**
Privacy Attack	Tensorflow_privacy_	Tensorflow_privacy_model.h5
	Privacy_meter_	Privacy_meter_model.h5
Evasion Attack	ART_EA_	Evasion_model.pt
Poison Attack	ART_PA_	**No need to upload the model**

Note: As for the present, there is no need to upload the datasets.



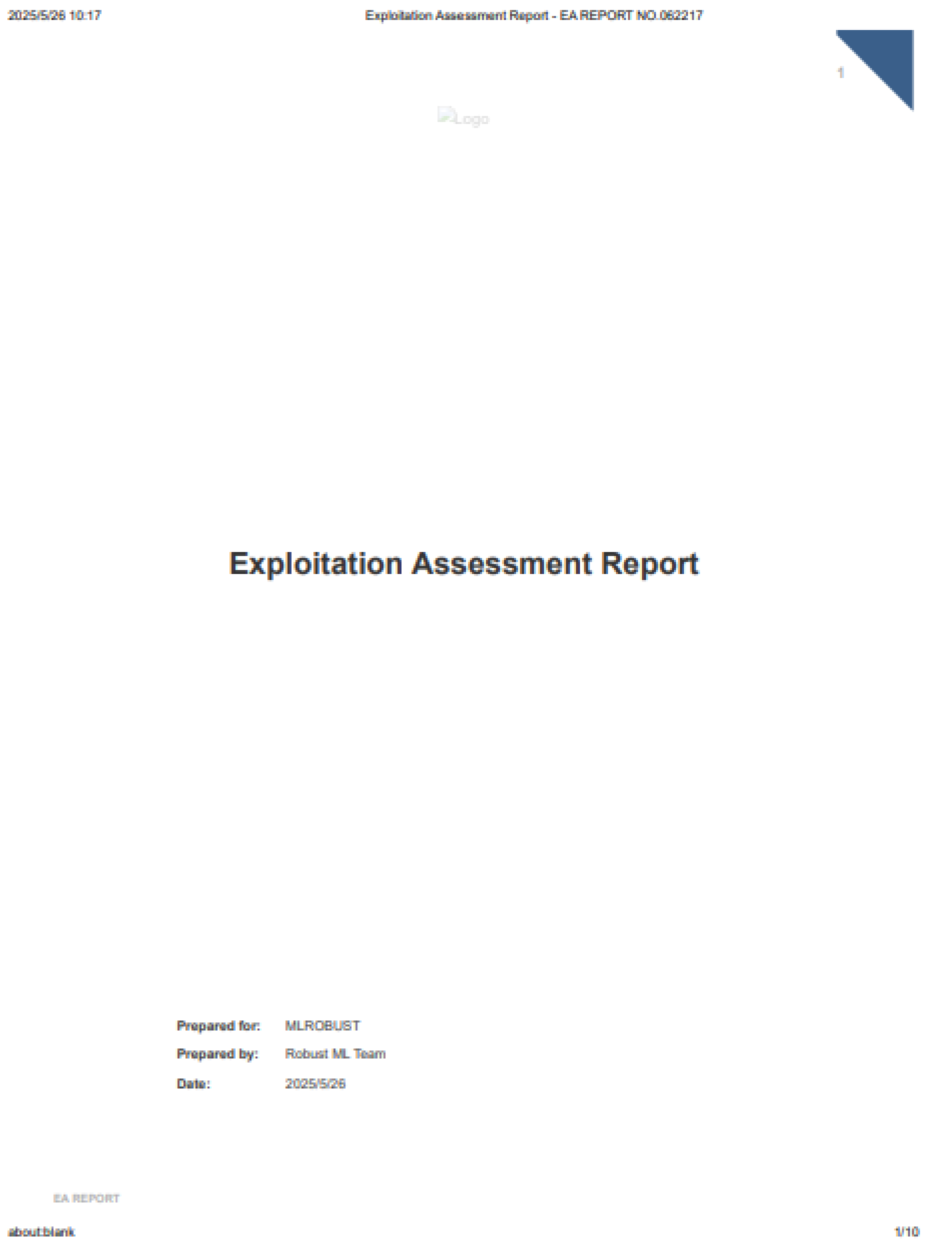
Processing completed

```
Starting to process selected implementations...

Processing Implementation Tensorflow_privacy_MIM_003...
Result for Tensorflow_privacy_MIM_003: Processed Tensorflow_privacy_MIM_003 success

Processing Implementation Tensorflow_privacy_MIM_004...
Result for Tensorflow_privacy_MIM_004: Processed Tensorflow_privacy_MIM_004 success

All selected implementations have been processed.
```

Step 3: Click the “Exploitation Assessment” button, then the testbed will start analyzing

Step 4: Wait for a while, and Area 4 will state what is processing.

Step 5: If the information indicates “All selected implementations have been processed”, then you can click the “Generate EA Report” button and get a final report as follows.



2025/5/26 10:17 Exploitation Assessment Report - EA REPORT NO.062217

1

Logo

Exploitation Assessment Report

Prepared for:	MLROBUST
Prepared by:	Robust ML Team
Date:	2025/5/26

EA REPORT

about:blank 1/10