1st NIST GenAI Text Workshop

NIST GenAI (pilot)

Text-to-Text Evaluation Results

March 26, 2025

Yooyoung Lee, Information Access Division (IAD)
Hari Iyer, Statistical Engineering Division (SED)

NIST NATIONAL INSTITUTE OF STANDARDS AND TECHNOLOGY
U.S. DEPARTMENT OF COMMERCE

ITL INFORMATION TECHNOLOGY LABORATORY

Disclaimer

- Certain commercial equipment, instruments, software, or materials are identified in this article in order to specify the experimental procedure adequately. Such identification is not intended to imply recommendation or endorsement by NIST, nor is it intended to imply that the equipment, instruments, software or materials are necessarily the best available for the purpose.
- The views, opinions and/or findings expressed are those of the authors and should not be interpreted as representing the official views or policies of NIST or the U.S. Government.
- All videos, images, graphs, and charts are original works created for the NIST GenAI program.

NIST GenAI Team

Yooyoung Lee
(Computer Scientist)

Hari Iyer
(Statistician)

Seungmin Seo
(Computer Scientist, CTR)

Peter Fontana
(Computer Scientist)

Lukas Diduch
(Research Engineer)

Kay Peterson
(Social Scientist)

George Awad
(Computer Scientist)

Haiying Guan
(Computer Scientist)

Sonika Sharma
(Computer Scientist)

Nour Riman
(Computational Neuroscientist, FGR)

Aryan Mishra
(Robotics Engineer, PREP)

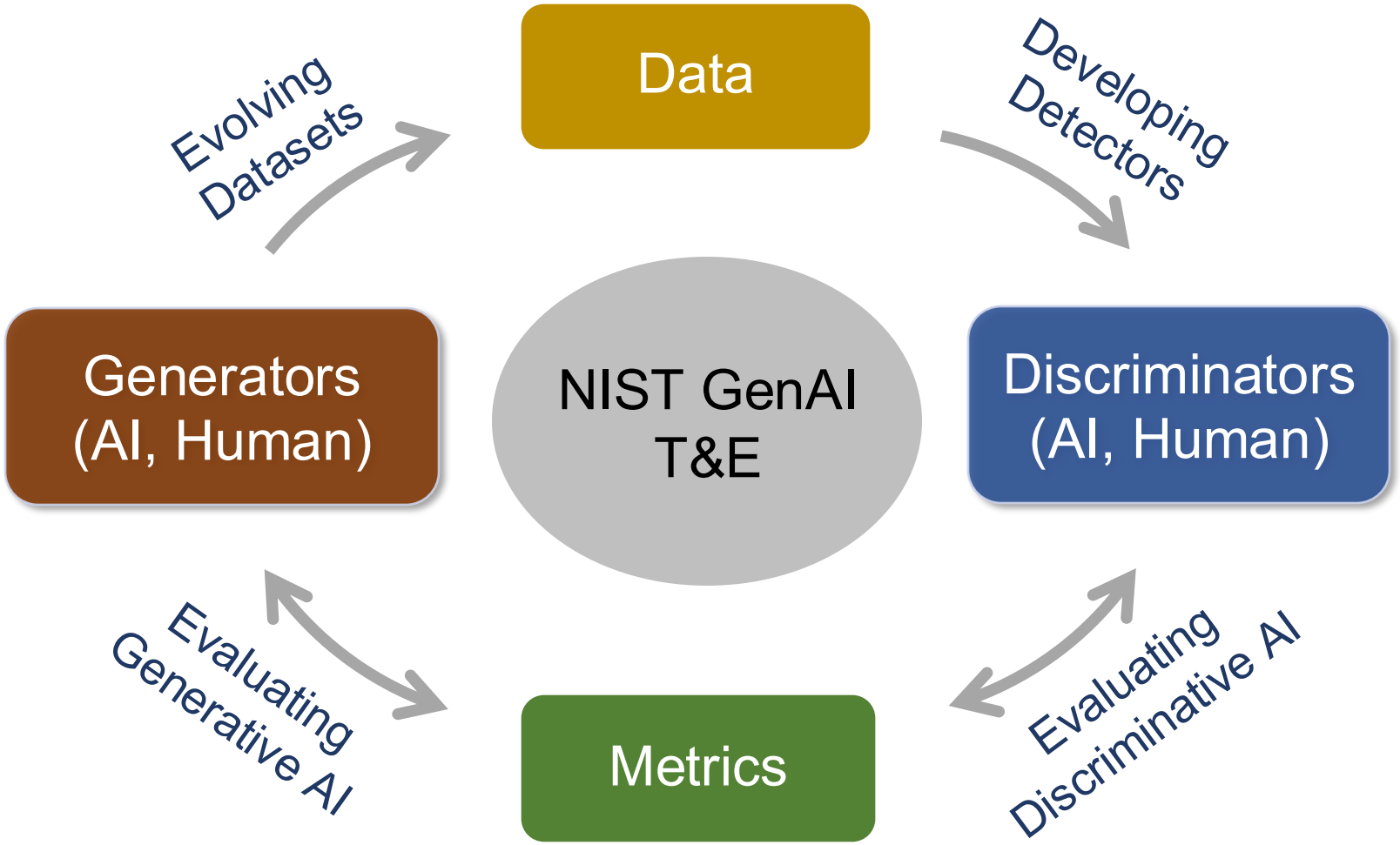
Outline

- NIST GenAI Introduction
- Text-to-Text (T2T) tasks & Datasets
- Participant & Submission Overview
- Performance Metrics (AUC, Brier Scores)
- Initial Data Analyses and Results
- Limitations, Challenges, and Conclusions
- Future Work

What is NIST GenAI?

- An umbrella program that supports various evaluations for research and measurement science in Generative AI across different modalities (text, image, video, audio, code)
  - A platform for testing and evaluation to measure and understand the capabilities and limitations of cutting-edge technologies
  - Generative adversarial framework: a cat-and-mouse game between generative AI and discriminative AI
  - Bridge the gap between research and real-world applications, and
  - Gather information to help ensure the trustworthiness of information and guide responsible use of digital content.

NIST GenAI Framework



NIST GenAI in Multiple Stages

First Stage Second Stage Subsequent Stages

- First stage (pilot): ”Indistinguishability” -> AI, Human, (AI+ Human)
- Second stage: “Indistinguishability” + ”Believability”
- Subsequent stages: feedback from participants and stakeholders (e.g., factual)

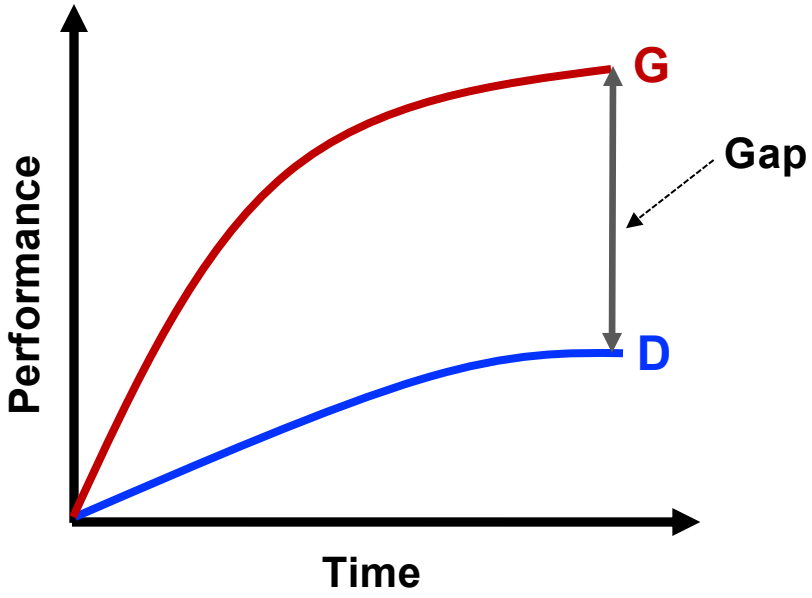
We want AI to be as effective as (or better than) Humans in generating content for improving:

- Quality of life
- Economic growth and innovation,
- Automation and efficiency …

These are Goals for the Generators

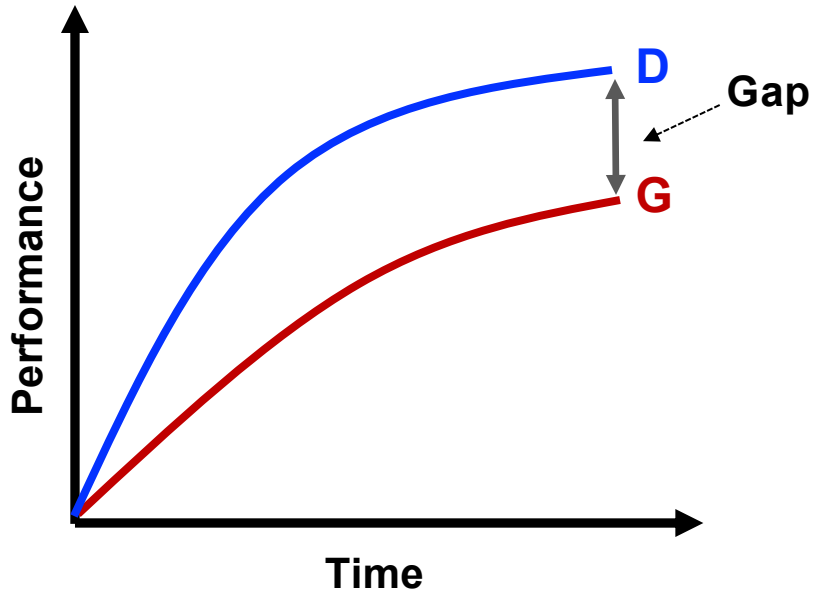
To safeguard against the negative consequences, it is important to have detectors that can recognize content generated by AI.

These are Goals for the Discriminators

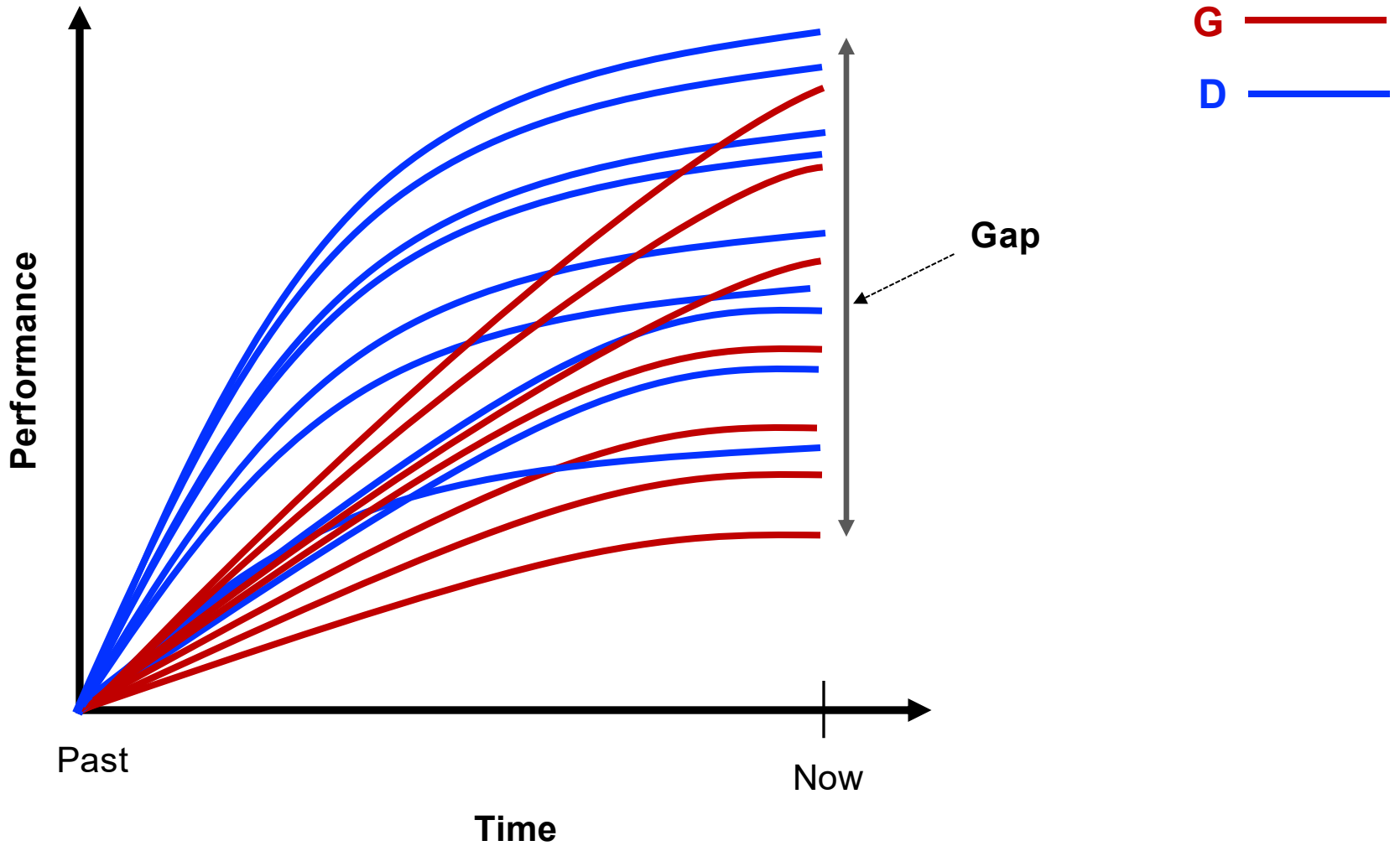


As Generators evolve and get better, Discriminators also need to get better at detecting synthetic (AI) content.

As Discriminators improve and get better, Generators also need to get better at generating content as good as human.



So, we need to understand the performance **gap** between Generators and Discriminators.
Our studies in *NIST GenAI* can help understand this **gap**.

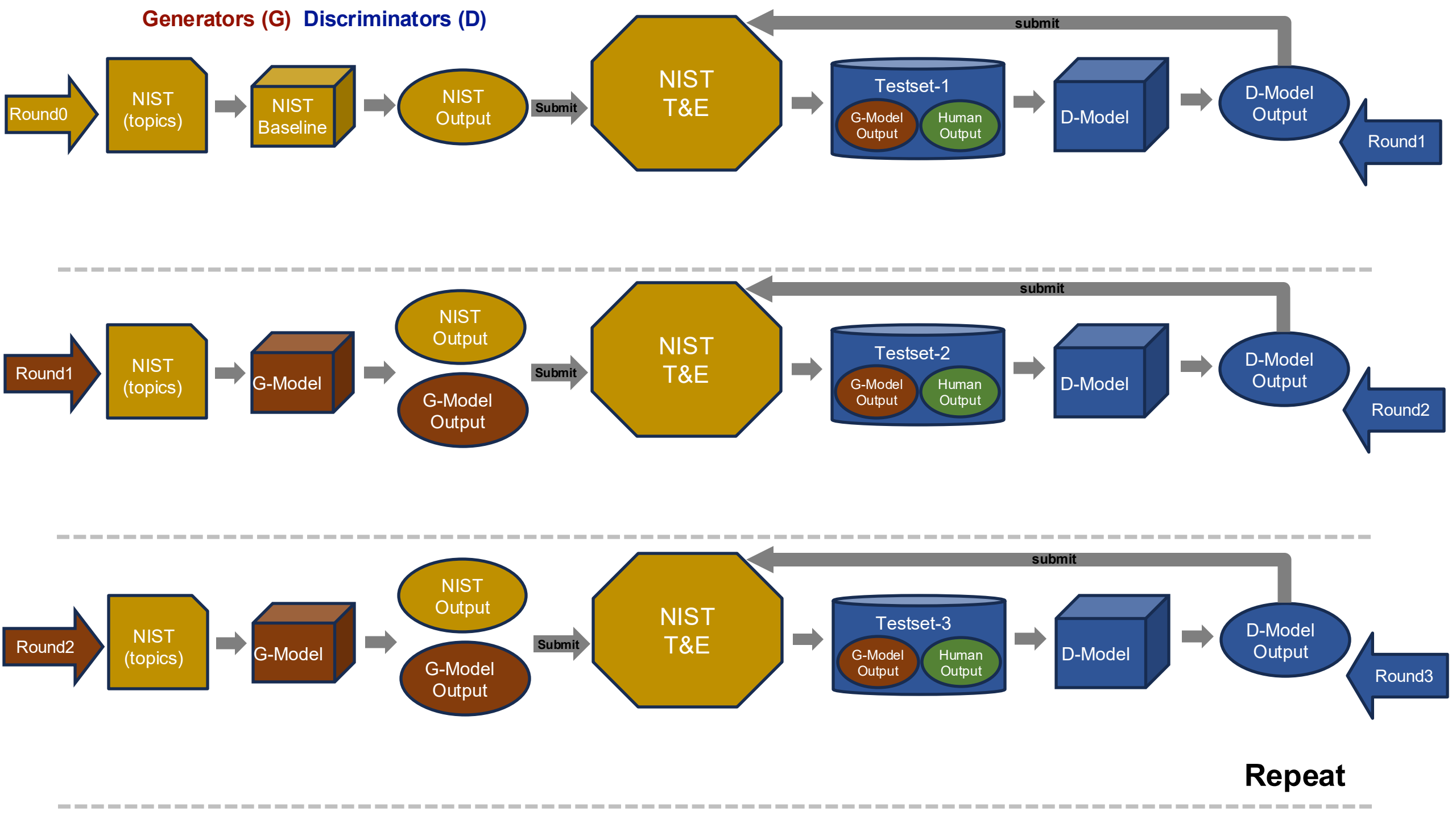

G
D
Performance
Gap
Past
Now
Time

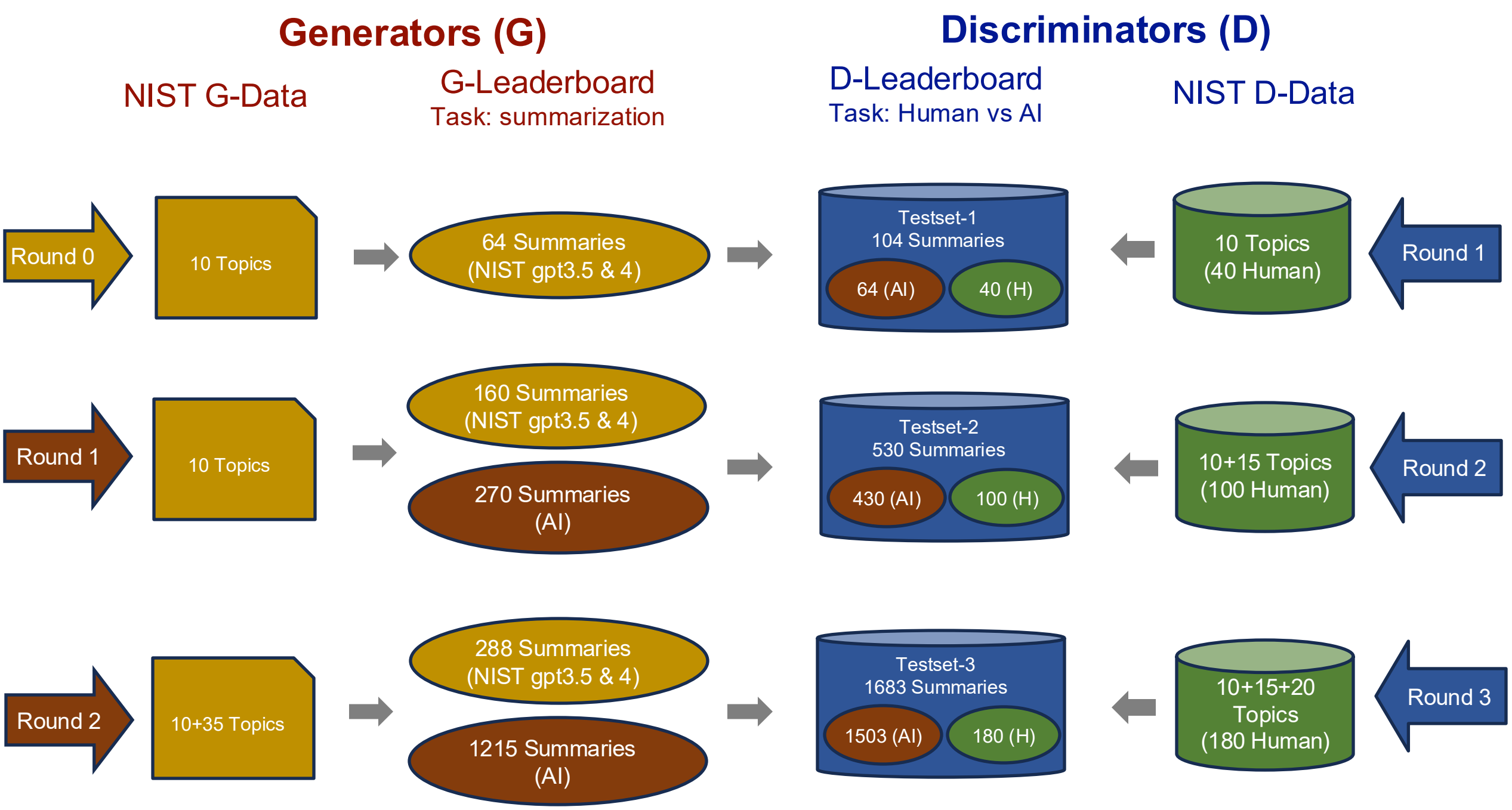
In this pilot, is there a performance **gap** between AI generators and AI discriminators?

The GenAI pilot is designed to gain sufficient knowledge to design a larger study that can make a realistic assessment of this **gap.**

Text-to-Text (T2T) Task

- T2T Generators (T2T-G) task
  - automatically generate summaries given a statement of information needed ("topic") and a set of source documents to summarize. The summaries should be indistinguishable from human created summaries as much as possible
- T2T Discriminators (T2T-D) task
  - automatically detect if a target summary has been generated using a Generative AI system or a Human


Generators (G) Discriminators (D)
Round0
NIST (topics)
NIST Baseline
NIST Output
Submit
NIST T&E
submit
Testset-1
G-Model Output
Human Output
D-Model
D-Model Output
Round1
Round1
NIST (topics)
G-Model
NIST Output
G-Model Output
Submit
NIST T&E
submit
Testset-2
G-Model Output
Human Output
D-Model
D-Model Output
Round2
Round2
NIST (topics)
G-Model
NIST Output
G-Model Output
Submit
NIST T&E
submit
Testset-3
G-Model Output
Human Output
D-Model
D-Model Output
Round3
Repeat

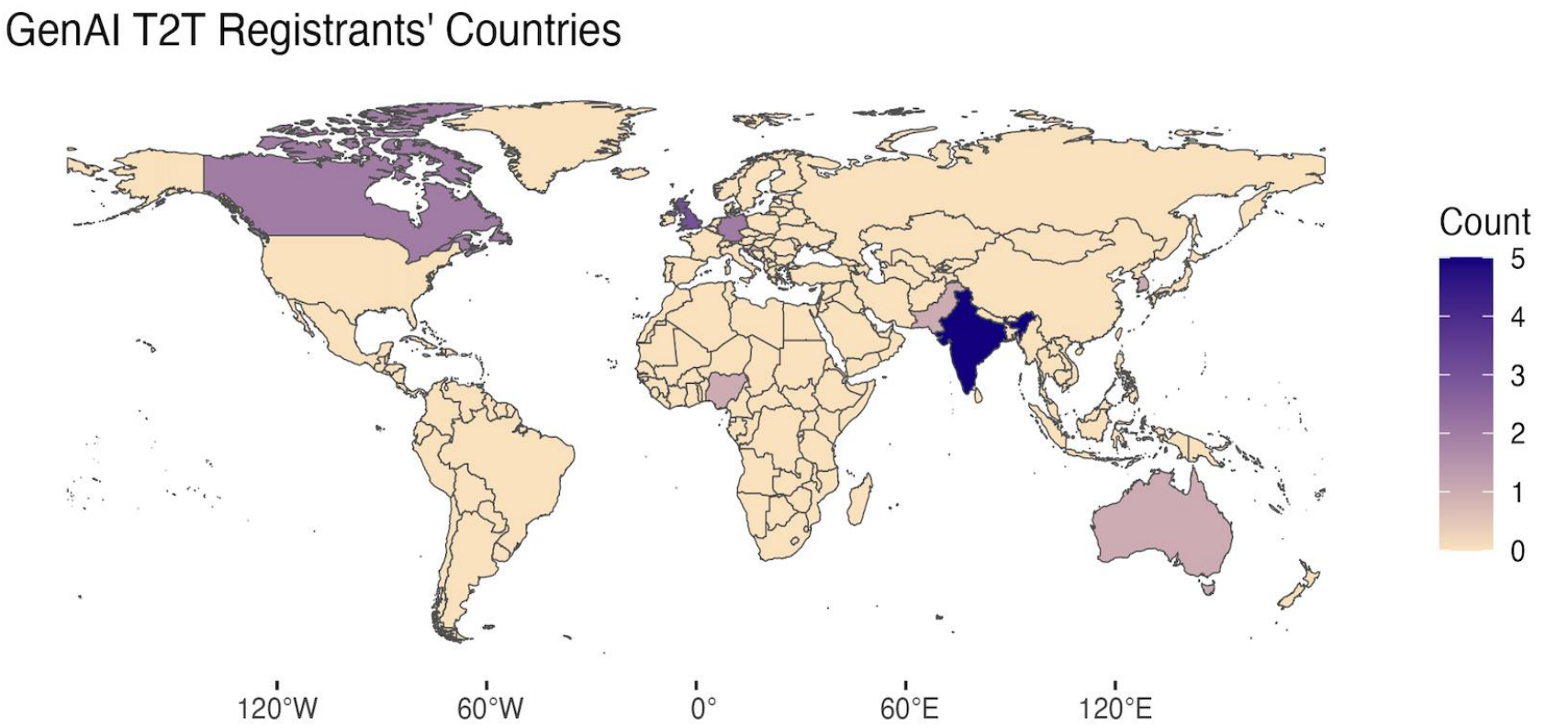

Generators (G)
Discriminators (D)
NIST G-Data
G-Leaderboard
Task: summarization
D-Leaderboard
Task: Human vs AI
NIST D-Data
Round 0
10 Topics
64 Summaries
(NIST gpt3.5 & 4)
Testset-1
104 Summaries
64 (AI)
40 (H)
10 Topics
(40 Human)
Round 1
Round 1
10 Topics
160 Summaries
(NIST gpt3.5 & 4)
270 Summaries
(AI)
Testset-2
530 Summaries
430 (AI)
100 (H)
10+15 Topics
(100 Human)
Round 2
Round 2
10+35 Topics
288 Summaries
(NIST gpt3.5 & 4)
1215 Summaries
(AI)
Testset-3
1683 Summaries
1503 (AI)
180 (H)
10+15+20
Topics
(180 Human)
Round 3

Dataset Overview

Total Summary Count			Discriminator			Generator			
Dataset	Human	AI	Round	Human	AI	Round	G-Participant	NIST-GPT3.5	NIST-GPT4
Testset1 104	40	64	**D-Round-1**	40	64	**G-Round-0**	0	24	40
Testset2 530	100	430	**D-Round-1**	40	64	**G-Round-0**	0	24	40
			D-Round-2	60	366	**G-Round-1**	270	36	60
Testset3 1683	180	1503	**D-Round-1**	40	64	**G-Round-0**	0	24	40
			D-Round-2	60	366	**G-Round-1**	270	36	60
			D-Round-3	80	1073	**G-Round-2**	945	48	80

T2T Registrations

- Registrants: 172
- Organizations: 83
- Countries: 14



Participants Overview

Generators (G)

Team ID	Team Type
6fc49	Academic
804fe	Academic
87a8c	Industry
0dea0	Academic
0782f	Industry
aa872	Industry

Discriminators (D)

Team ID	Team Type
6fc49	Academic
804fe	Academic
87a8c	Industry
0dea0	Academic
29d48	Industry
18126	Industry
6655b	Academic
9de37	Government
b3cd9	Industry
d718e	Industry
993ad	Industry

Submission Overview

Generators (G)

48 valid submissions from 6 teams

Round 1

27 SUBMISSIONS
6 ORGANIZATIONS

Round 2

21 SUBMISSIONS
3 ORGANIZATIONS

Discriminators (D)

348 valid submissions from 11 teams

50 SUBMISSIONS
11 ORGANIZATIONS

Round 1

137 SUBMISSIONS
8 ORGANIZATIONS

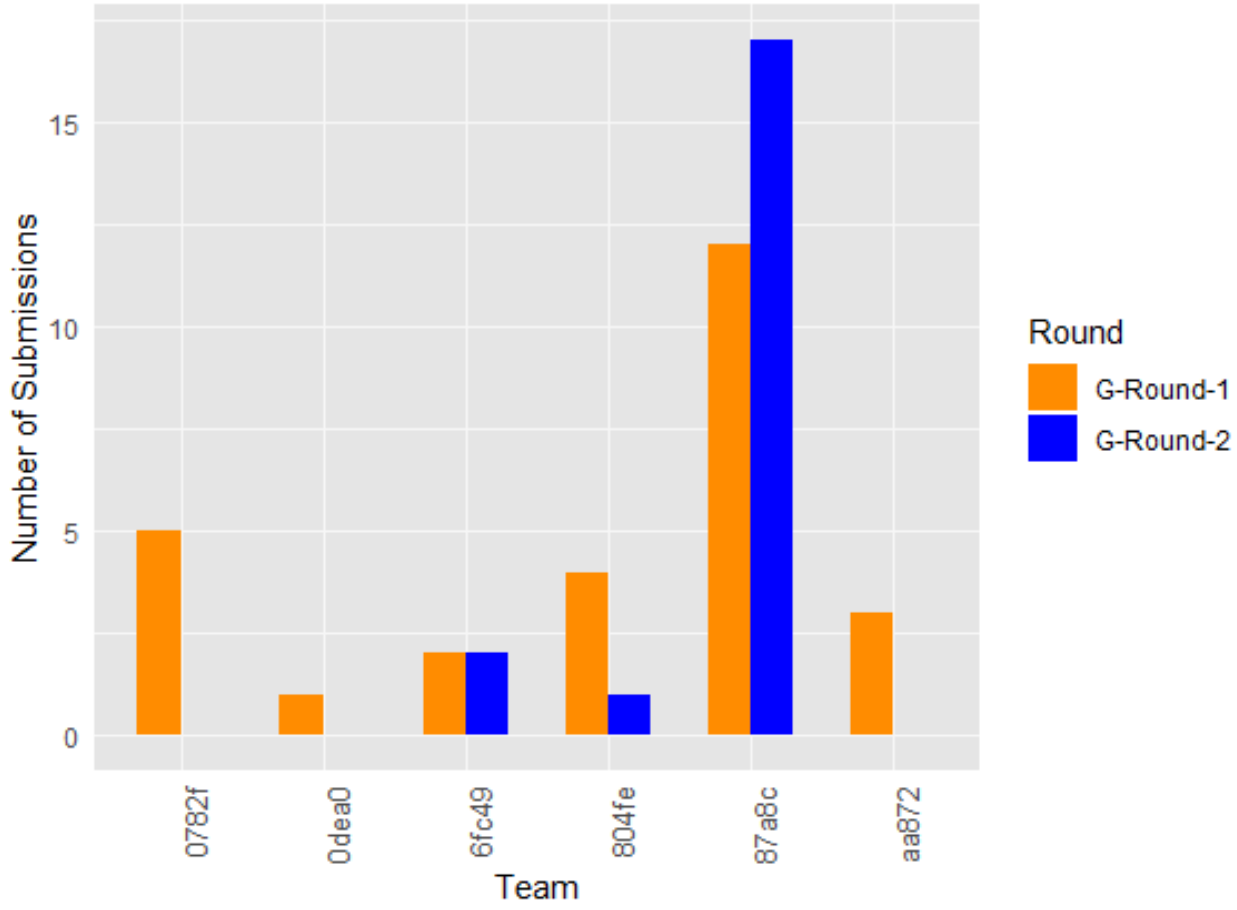
Round 2

161 SUBMISSIONS
7 ORGANIZATIONS

Round 3

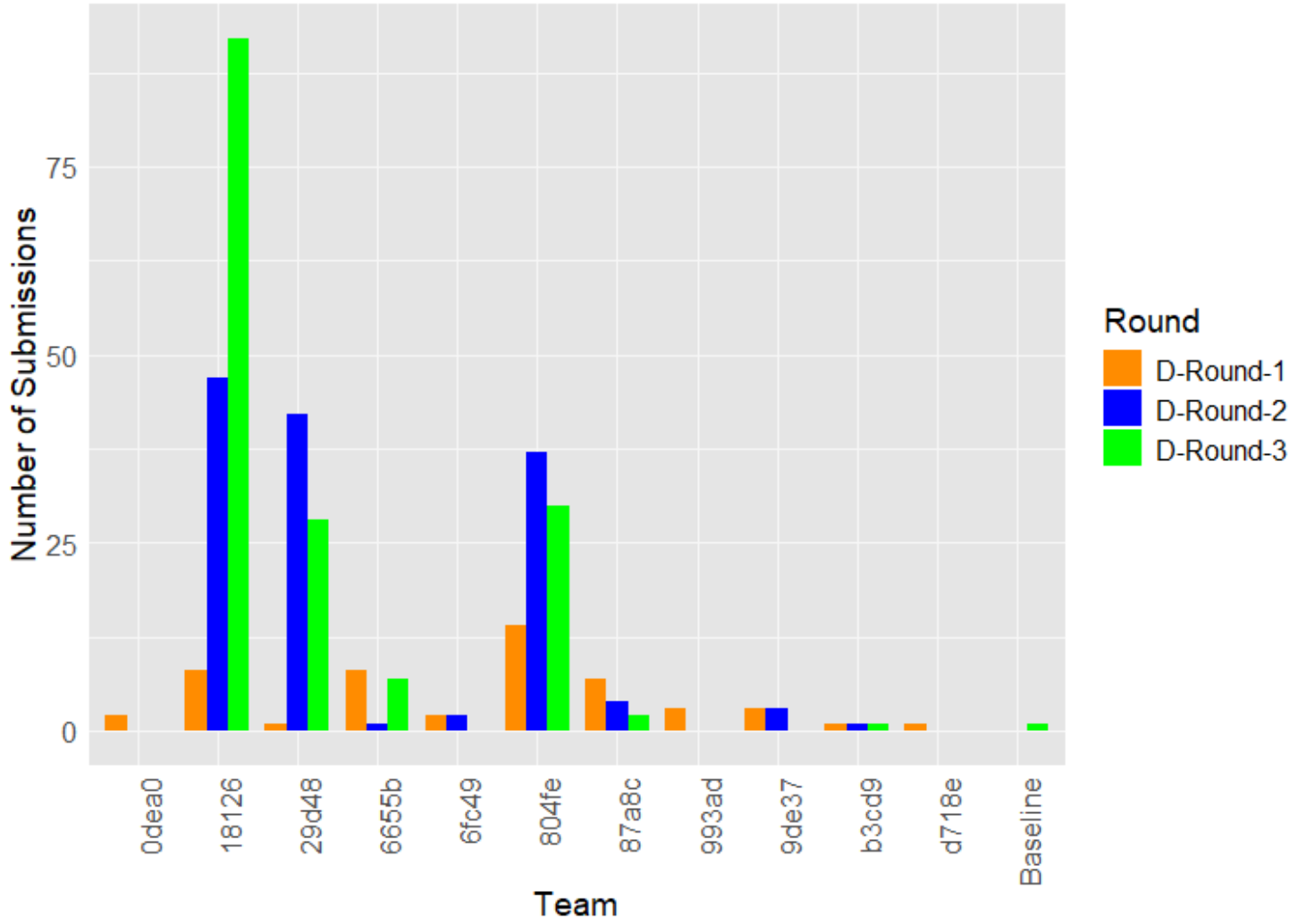
Generator Round Submissions

G-Team	G-round-1	G-round-2
0782f	Yes	No
0dea0	Yes	No
6fc49	Yes	Yes
804fe	Yes	Yes
87a8c	Yes	Yes
aa872	Yes	No



Discriminator Round Submissions

D-Team	D-round-1	D-round-2	D-round-3
0dea0	Yes	No	No
18126	Yes	Yes	Yes
29d48	Yes	Yes	Yes
6655b	Yes	Yes	Yes
6fc49	Yes	Yes	No
804fe	Yes	Yes	Yes
87a8c	Yes	Yes	Yes
993ad	Yes	No	No
9de37	Yes	Yes	No
b3cd9	Yes	Yes	Yes
d718e	Yes	No	No

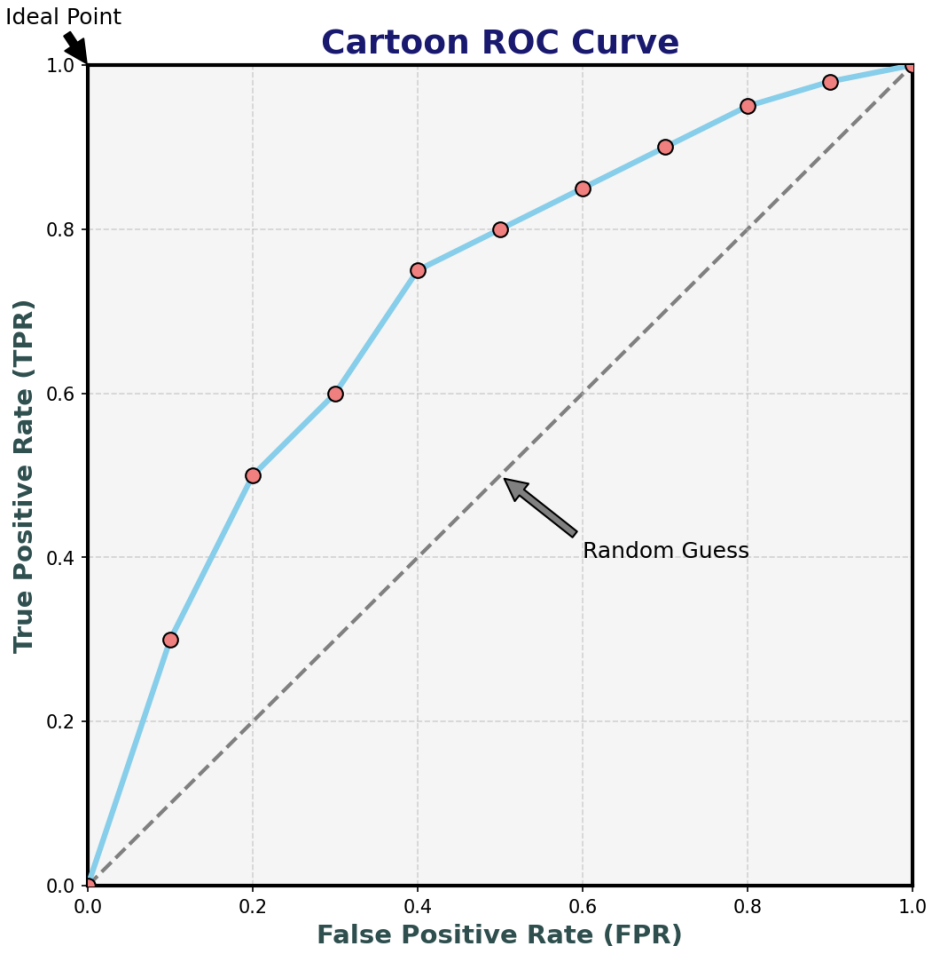


Outline

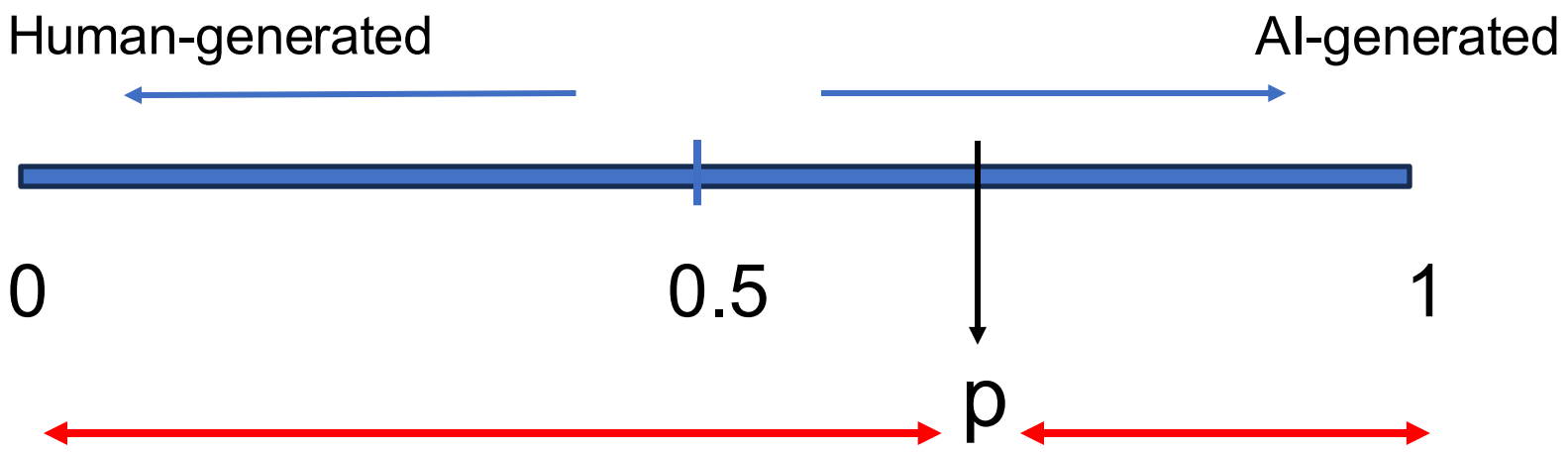
- NIST GenAI Introduction
- Text-to-Text (T2T) tasks & Datasets
- Participant & Submission Overview
- Performance Metrics (AUC, Brier Scores)
- Initial Data Analyses and Results
- Limitations, Challenges, and Conclusions
- Future Work

Performance Metrics - AUC

- Detectors are given a test-set with AI generated content and human generated content.
- Detection scores are real numbers between 0 and 1. Higher numbers suggest the content is AI-generated (target) and lower numbers suggest the content is Human-generated (nontarget)
- The detector's ability to discriminate between AI and Human is quantified using AUC (Area Under the ROC curve).



Detection Scores



For Human-generated content

Error = p-0 = p

Penalty = p^2

Error = 1-p

Penalty = $(1-p)^2$

For AI-generated content

Brier Score for Targets (BrierT) = Average Penalty = Average of $(1-p)^2$

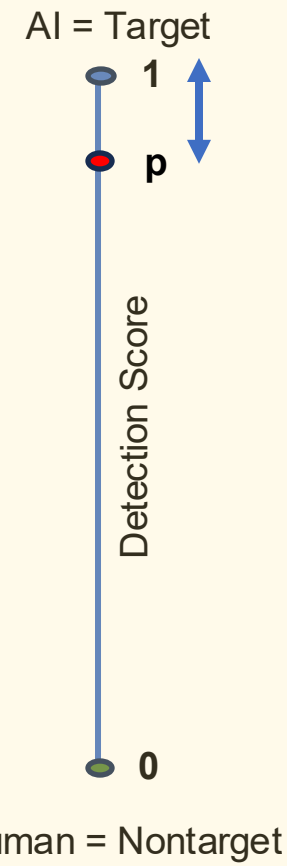
Brier Score for Nontargets (BrierN) = Average Penalty = Average of p^2

Performance Metrics – BrierT

- Suppose we have n_1 summaries that are AI-generated.
- $p_1, p_2, \ldots, p_{n_1}$ are the detection scores

$$\text{BrierT} = \frac{1}{n_1}\sum_{i=1}^{n_1}(p_i - 1)^2$$

(T represents '**T**arget')

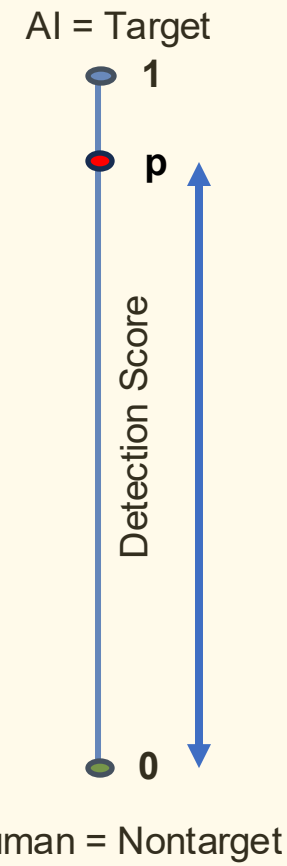


Performance Metrics – BrierN

- Suppose we have n_0 summaries that are Human-generated.
- $p_1, p_2, \ldots, p_{n_0}$ are the detection scores

$$\text{BrierN} = \frac{1}{n_0} \sum_{i=1}^{n_0} (p_i - 0)^2.$$

(N represents '**N**on-target')



Goal for Detector Systems

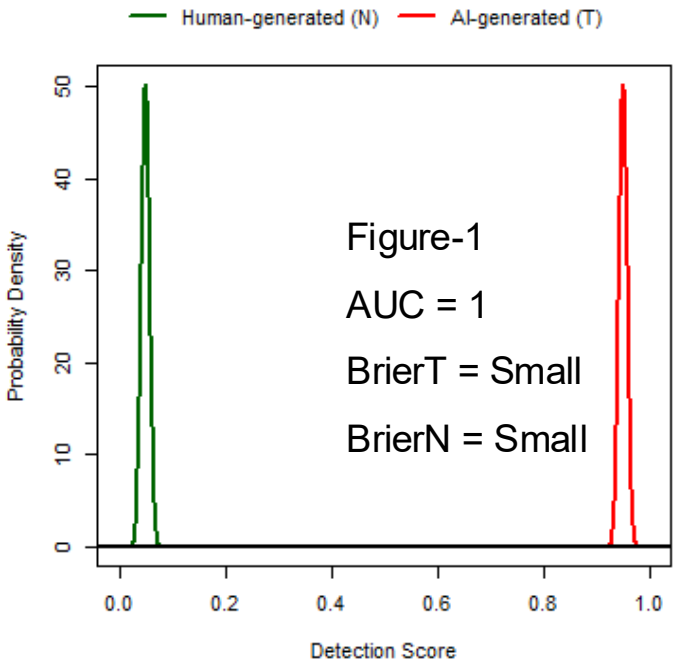
Good detectors should have

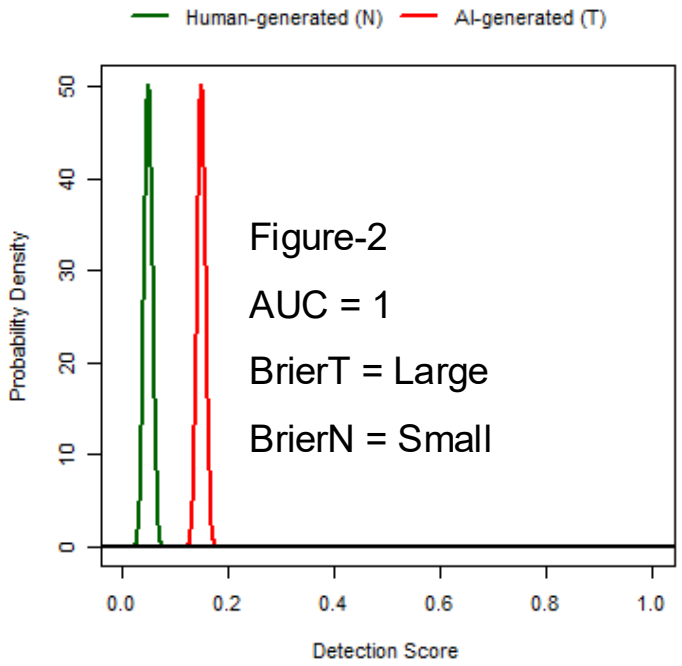
- AUC scores close to 1
- BrierT scores close to 0 (i.e. Detection scores close to 1 for Targets)
- BrierN scores close to 0 (i.e. Detection scores close to 0 for Nontargets)

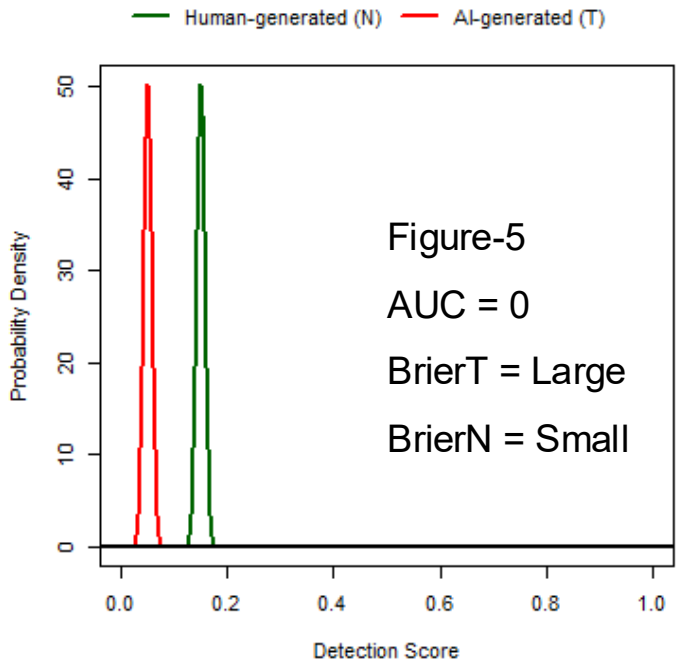
- BrierT and BrierN are reported in “squared” scale (similar to mean square error).
- They may be easier to interpret in the original scale, that is sqrt(BrierT) and sqrt(BrierN) (similar to Root Mean Square Error)

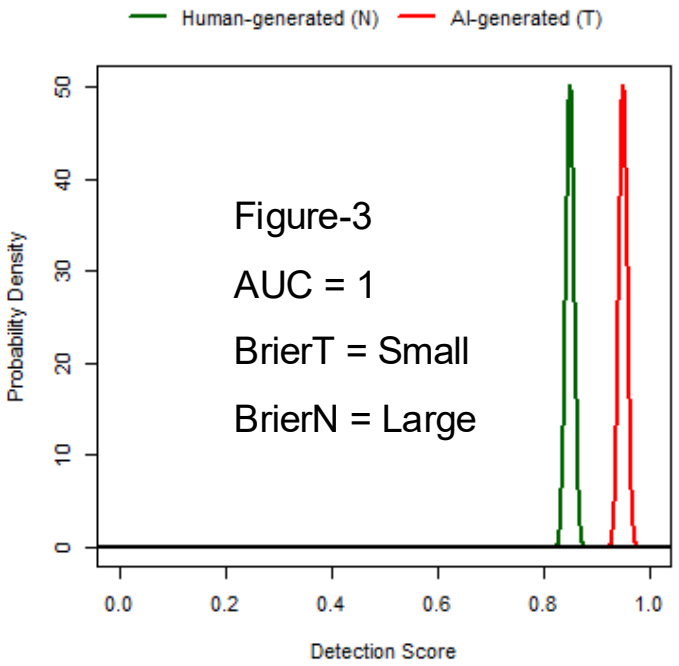
Goal for Generator Systems

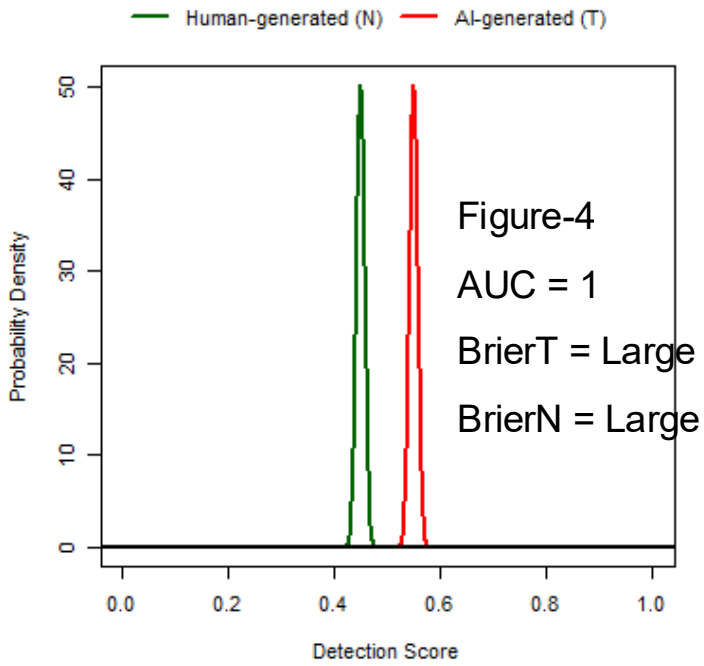
- Generated AI content is expected to be indistinguishable from Human-generated content.
- A sensible goal for a generator **A** is to generate content that is likely to result in **AUC around 0.5** based on a test-set containing AI content generated by **A** as well as Human-generated content.
- But generators provide only ‘part’ of the test-set. The other part comes from human-generated content. Hence the AUC will also depend on the particular collection of human-generated content used.
- So, a generator may be trained in a way that the generated content gets a ‘low detection score’ from any detector. This will make the user decide the content is ‘human-generated’ (for instance, using a cutoff of 0.5). This will make the BrierT score large (closer to 1).

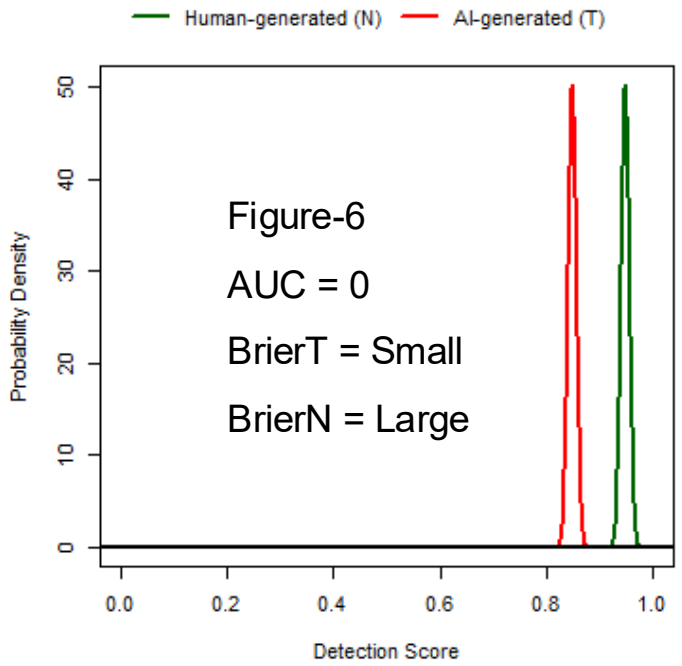

Human-generated (N)
AI-generated (T)
Probability Density
Detection Score
Figure-1
AUC = 1
BrierT = Small
BrierN = Small

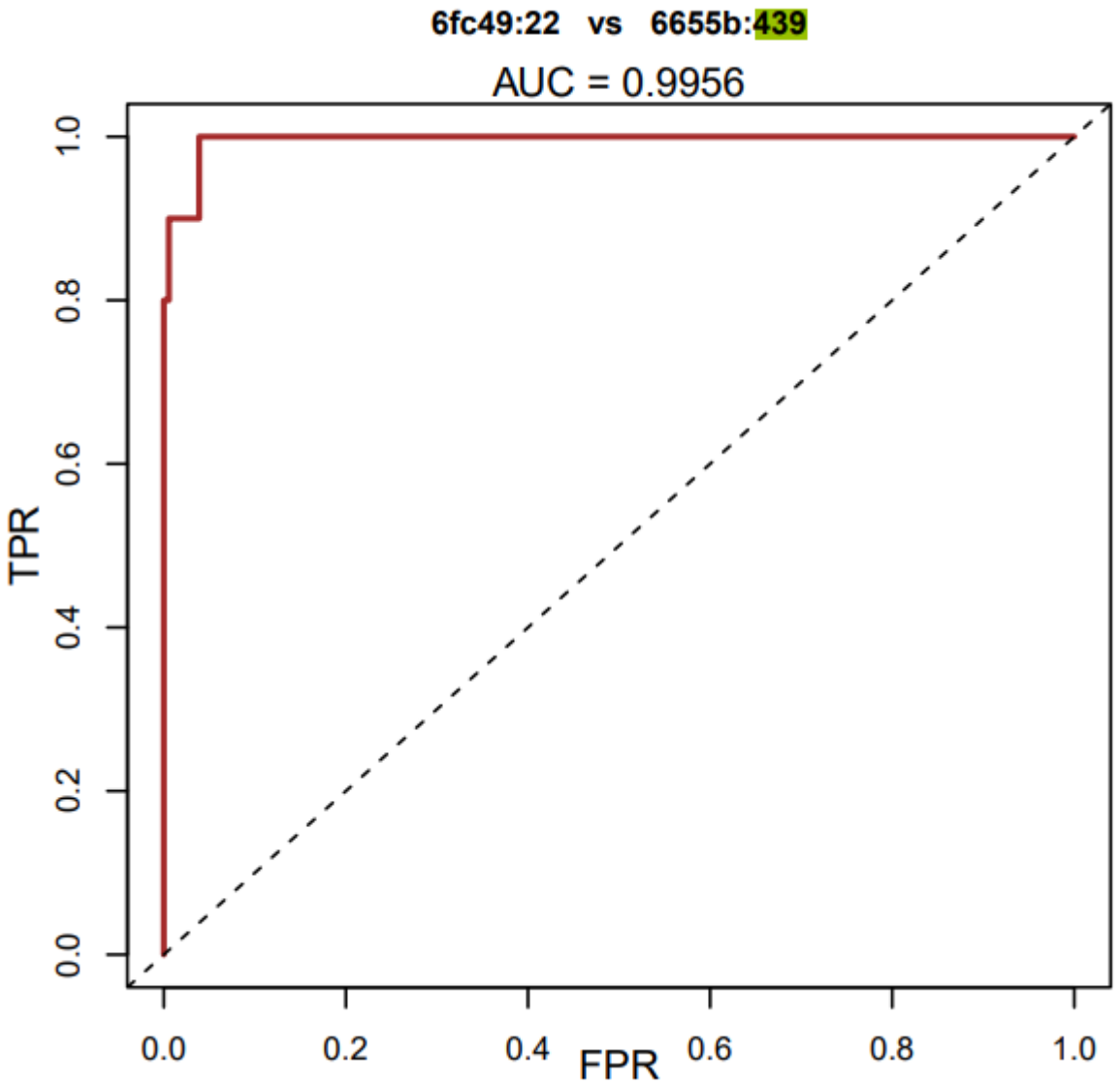

Human-generated (N)
AI-generated (T)
Probability Density
Detection Score
Figure-2
AUC = 1
BrierT = Large
BrierN = Small

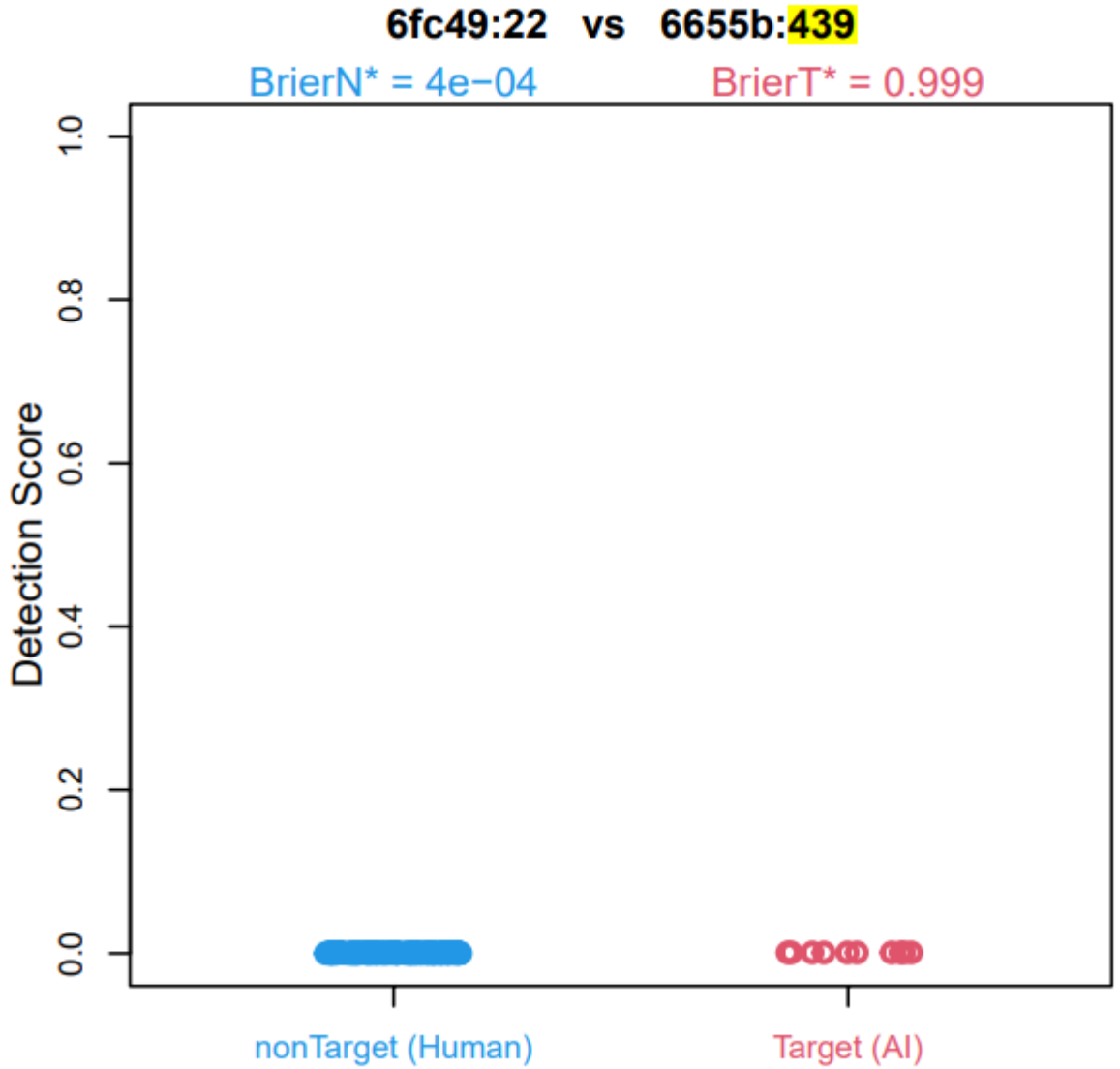

Human-generated (N)
AI-generated (T)
Probability Density
Detection Score
Figure-5
AUC = 0
BrierT = Large
BrierN = Small

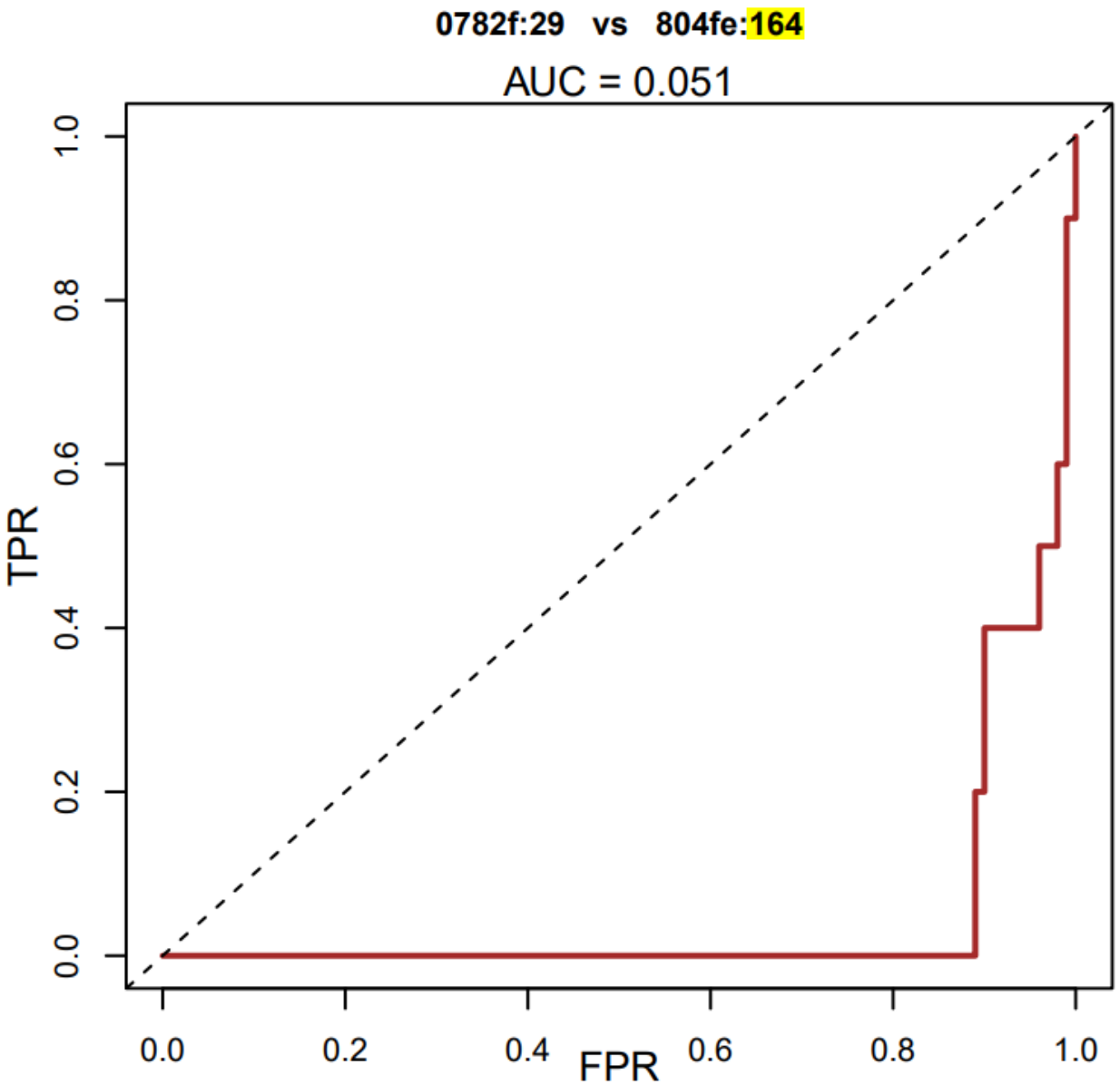

Human-generated (N)
AI-generated (T)
Probability Density
Detection Score
Figure-3
AUC = 1
BrierT = Small
BrierN = Large

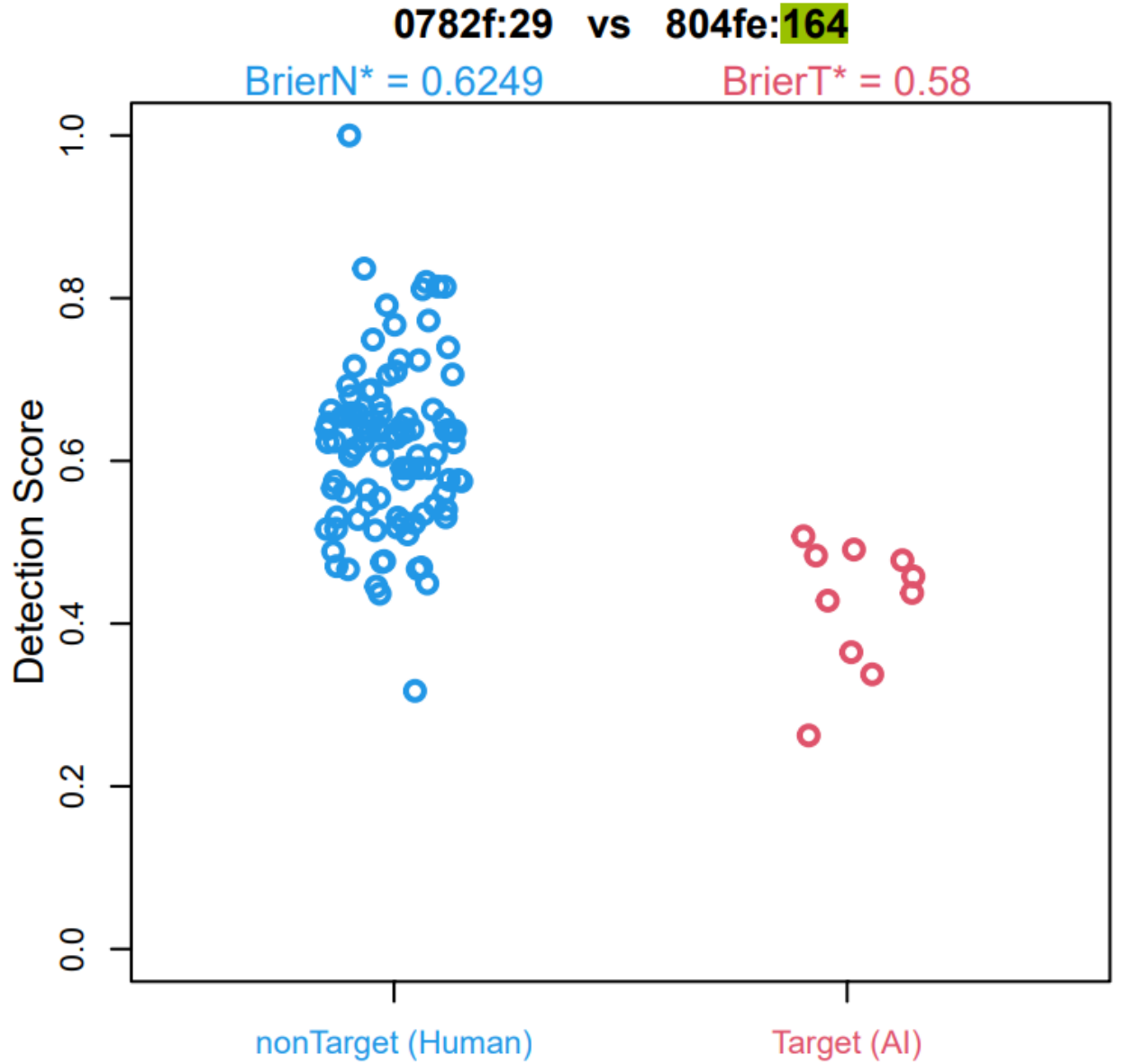

Human-generated (N)
AI-generated (T)
Probability Density
Detection Score
Figure-4
AUC = 1
BrierT = Large
BrierN = Large

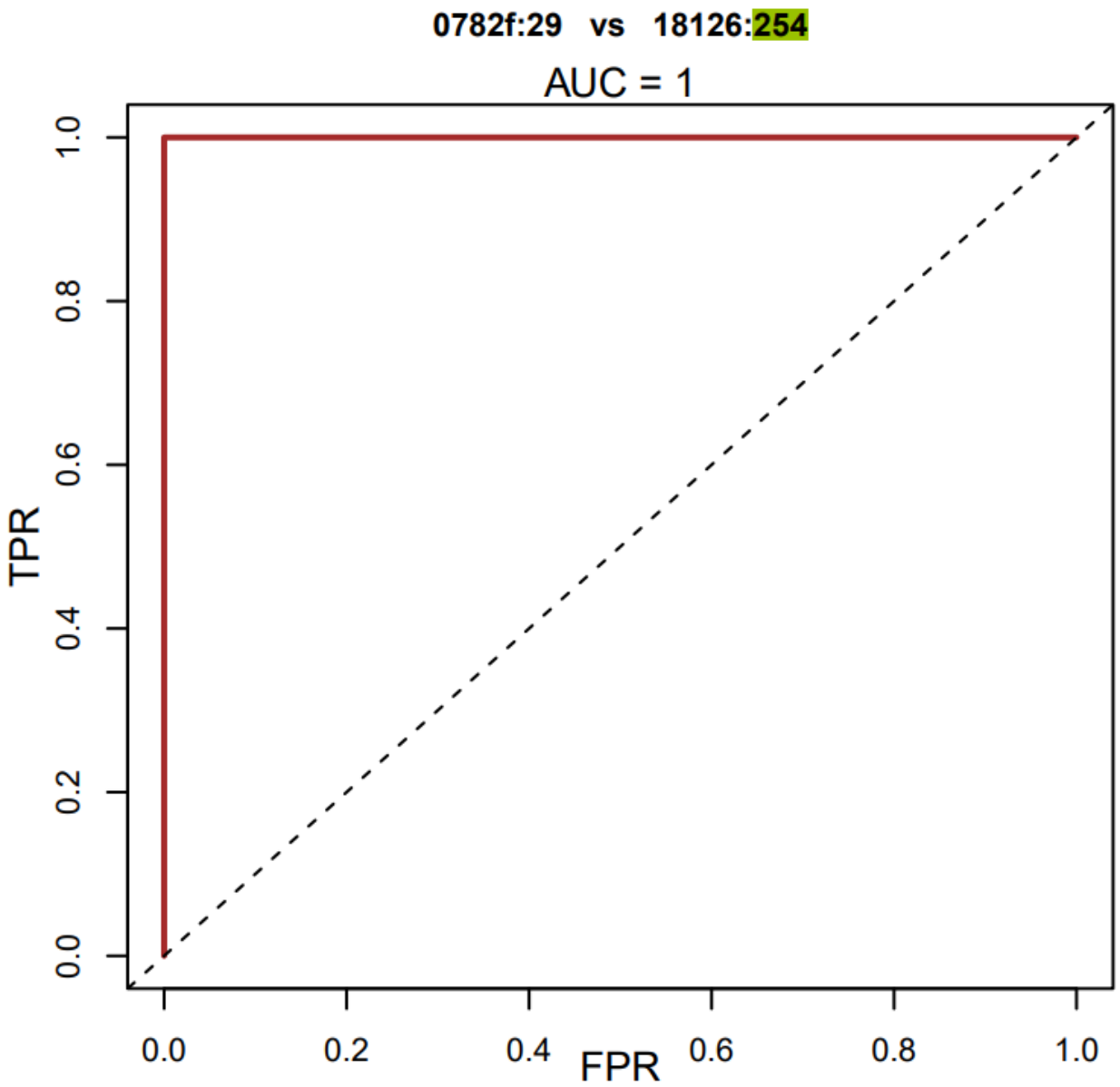

Human-generated (N)
AI-generated (T)
Probability Density
Detection Score
Figure-6
AUC = 0
BrierT = Small
BrierN = Large

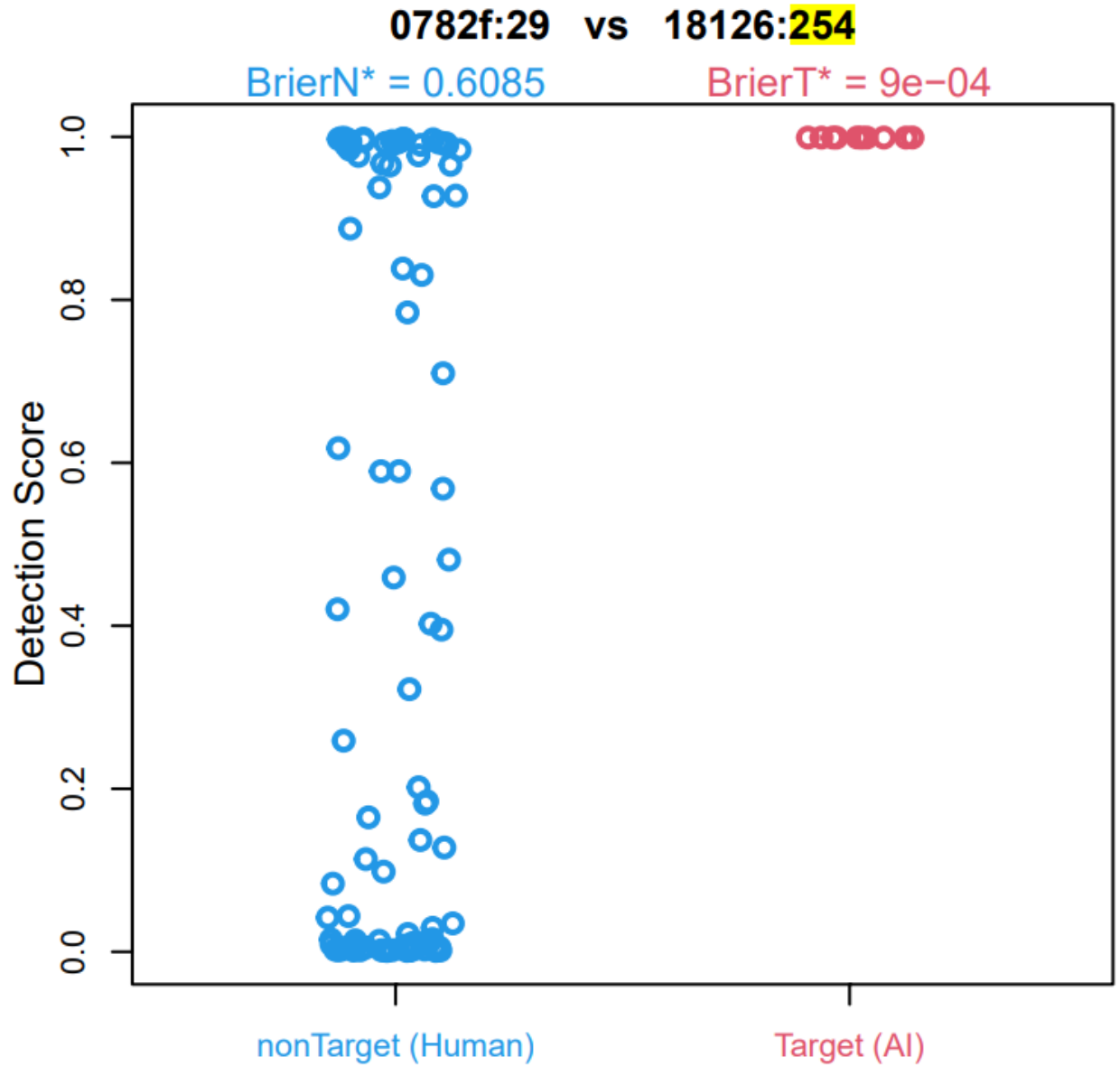

6fc49:22 vs 6655b:439
AUC = 0.9956
TPR
FPR
0.0
0.2
0.4
0.6
0.8
1.0

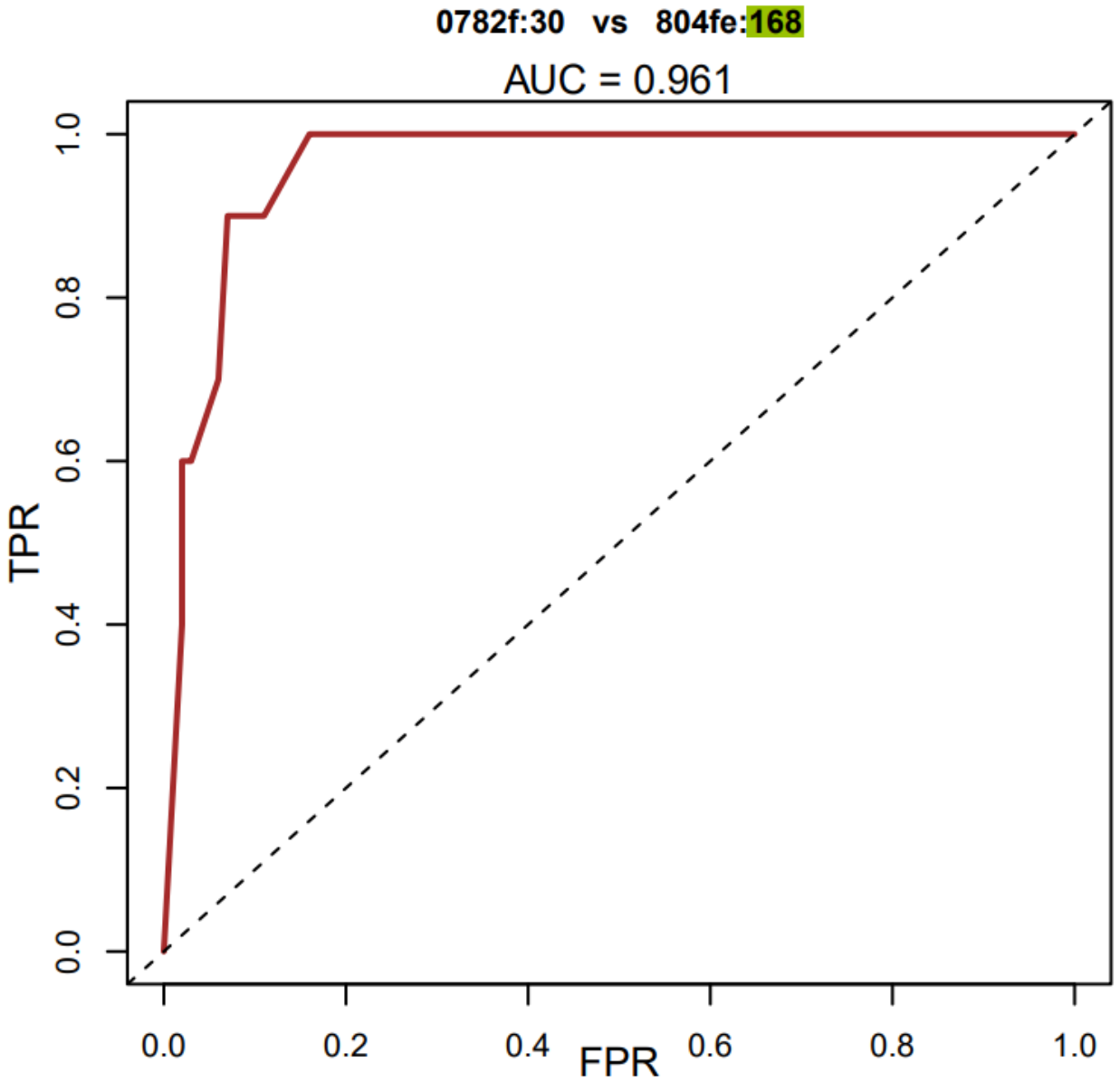

6fc49:22 vs 6655b:439
BrierN* = 4e−04
BrierT* = 0.999
Detection Score
0.0
0.2
0.4
0.6
0.8
1.0
nonTarget (Human)
Target (AI)

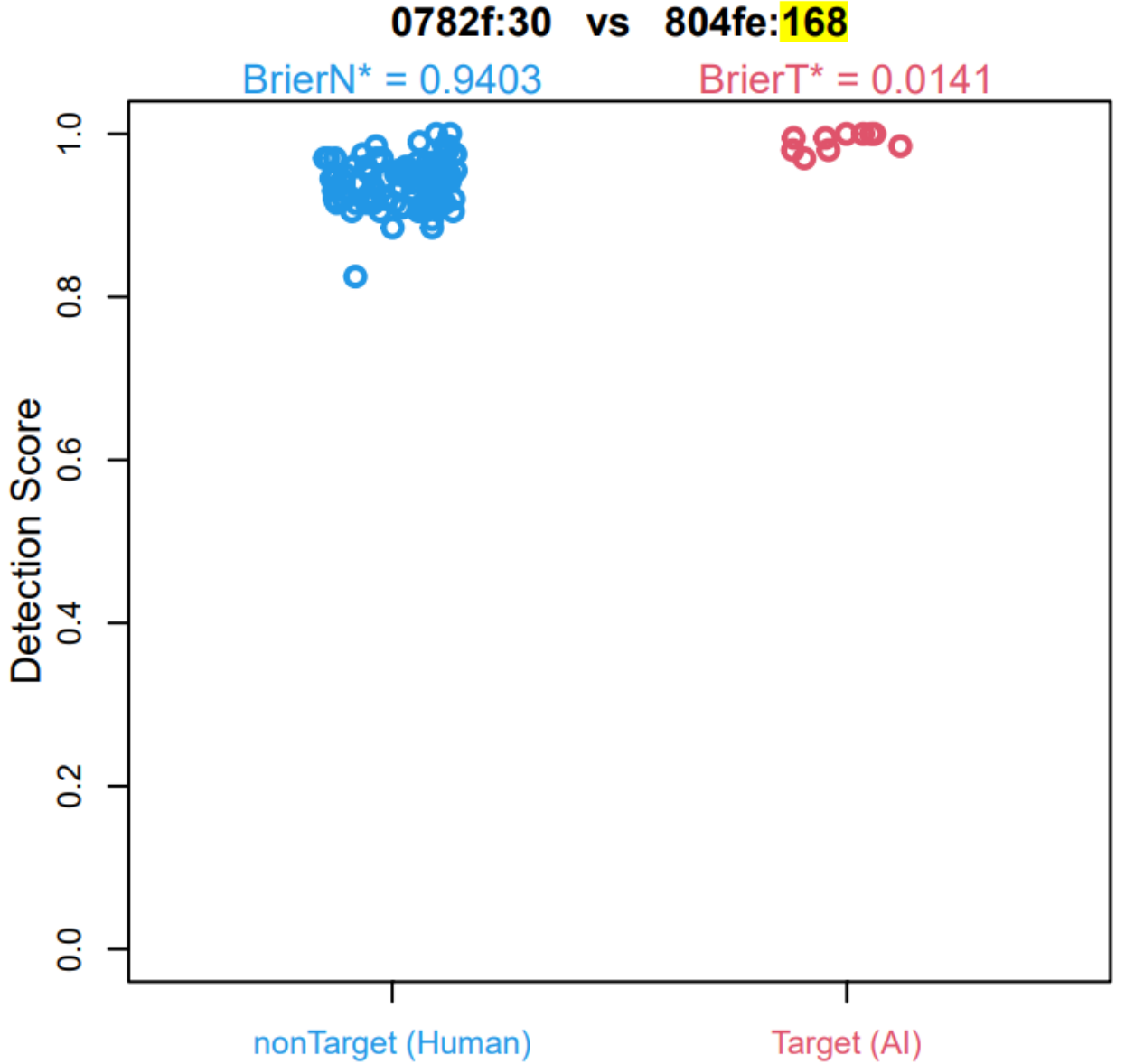

0782f:29 vs 804fe:164
AUC = 0.051
TPR
FPR
0.0
0.2
0.4
0.6
0.8
1.0

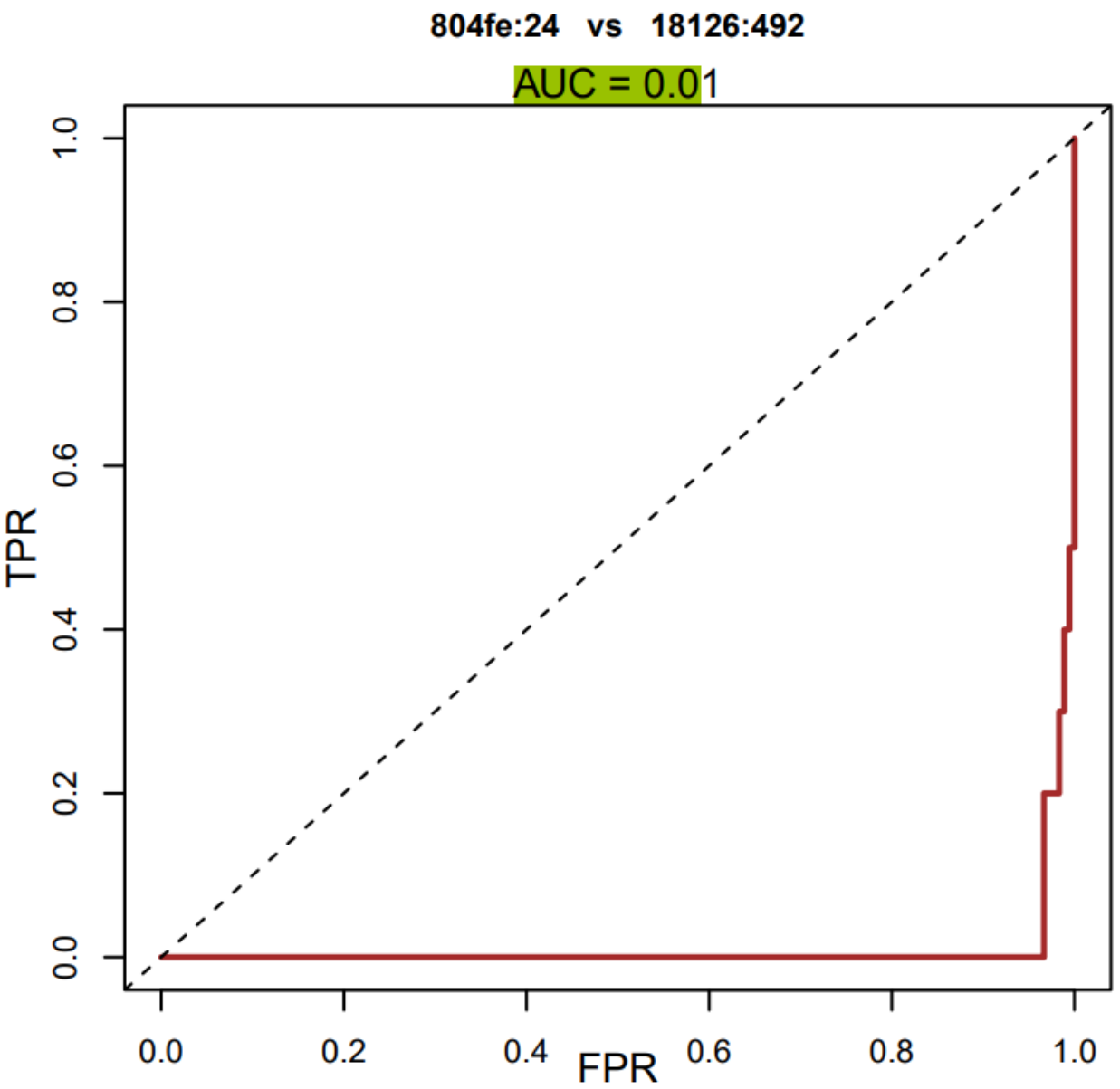

0782f:29 vs 804fe:164
BrierN* = 0.6249
BrierT* = 0.58
Detection Score
0.0
0.2
0.4
0.6
0.8
1.0
nonTarget (Human)
Target (AI)

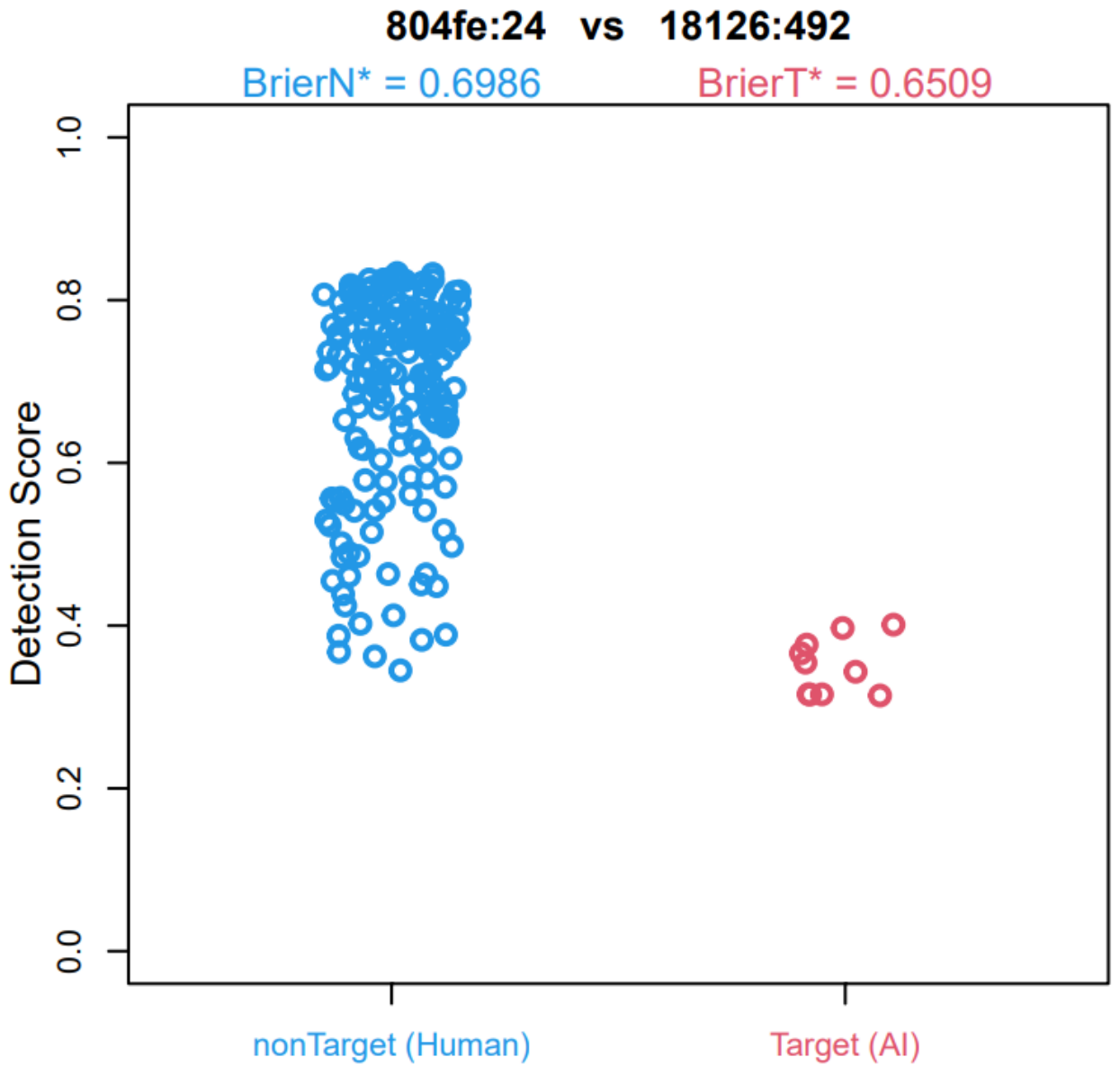

0782f:29 vs 18126:254
AUC = 1
TPR
FPR
0.0
0.2
0.4
0.6
0.8
1.0

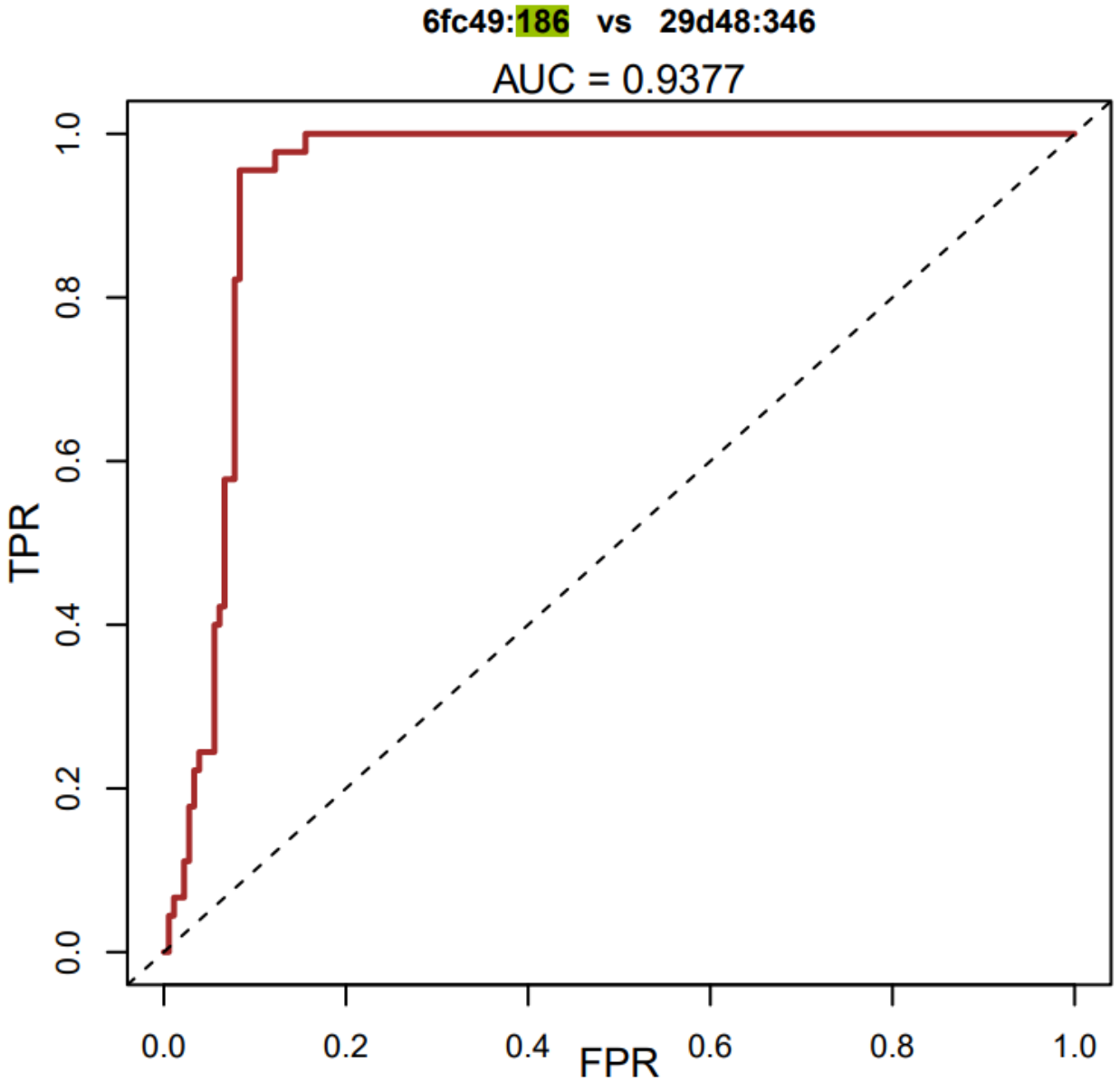

0782f:29 vs 18126:254
BrierN* = 0.6085
BrierT* = 9e−04
Detection Score
0.0
0.2
0.4
0.6
0.8
1.0
nonTarget (Human)
Target (AI)

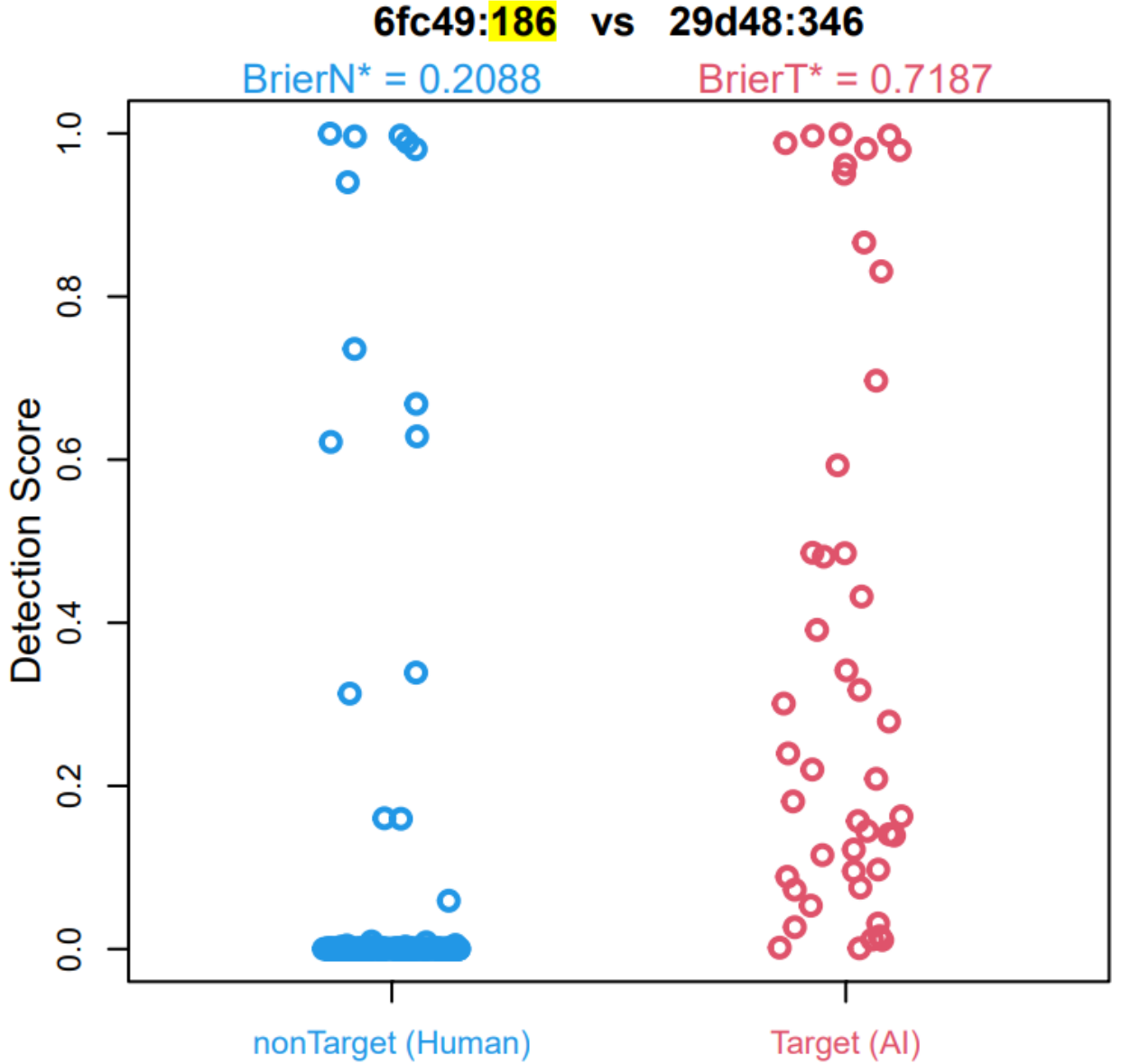

0782f:30 vs 804fe:168
AUC = 0.961
TPR
FPR
0.0
0.2
0.4
0.6
0.8
1.0


0782f:30 vs 804fe:168
BrierN* = 0.9403
BrierT* = 0.0141
Detection Score
0.0
0.2
0.4
0.6
0.8
1.0
nonTarget (Human)
Target (AI)


804fe:24 vs 18126:492
AUC = 0.01
TPR
FPR
0.0
0.2
0.4
0.6
0.8
1.0


804fe:24 vs 18126:492
BrierN* = 0.6986
BrierT* = 0.6509
Detection Score
0.0
0.2
0.4
0.6
0.8
1.0
nonTarget (Human)
Target (AI)


6fc49:186 vs 29d48:346
AUC = 0.9377
TPR
FPR
0.0
0.2
0.4
0.6
0.8
1.0

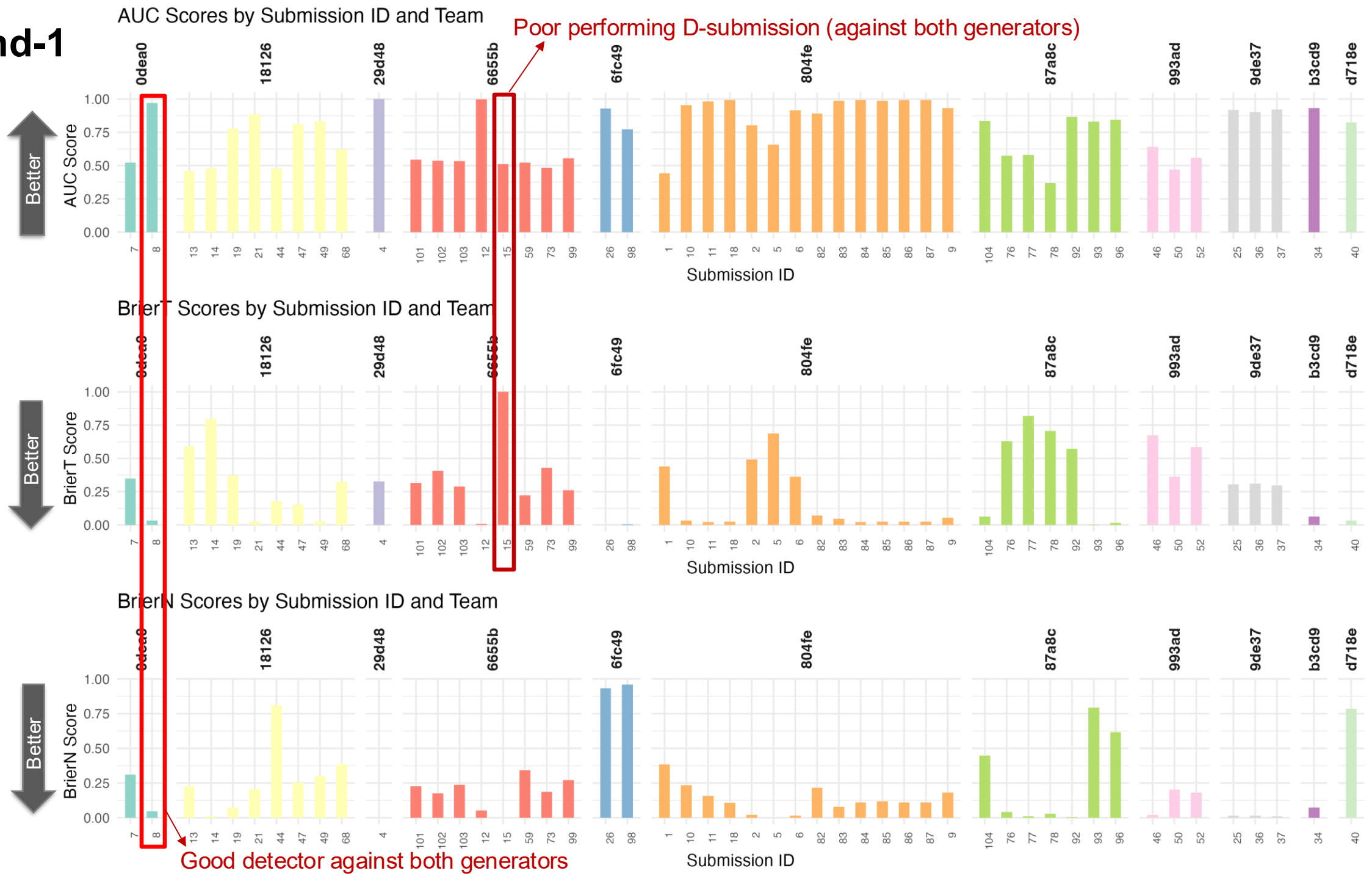

6fc49:186 vs 29d48:346
BrierN* = 0.2088
BrierT* = 0.7187
Detection Score
0.0
0.2
0.4
0.6
0.8
1.0
nonTarget (Human)
Target (AI)

System Performance Overview

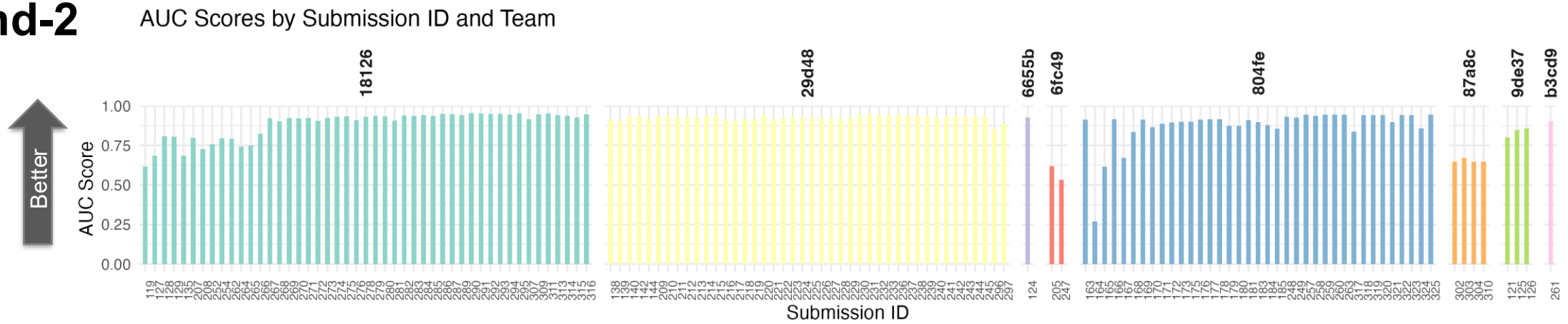
Caution

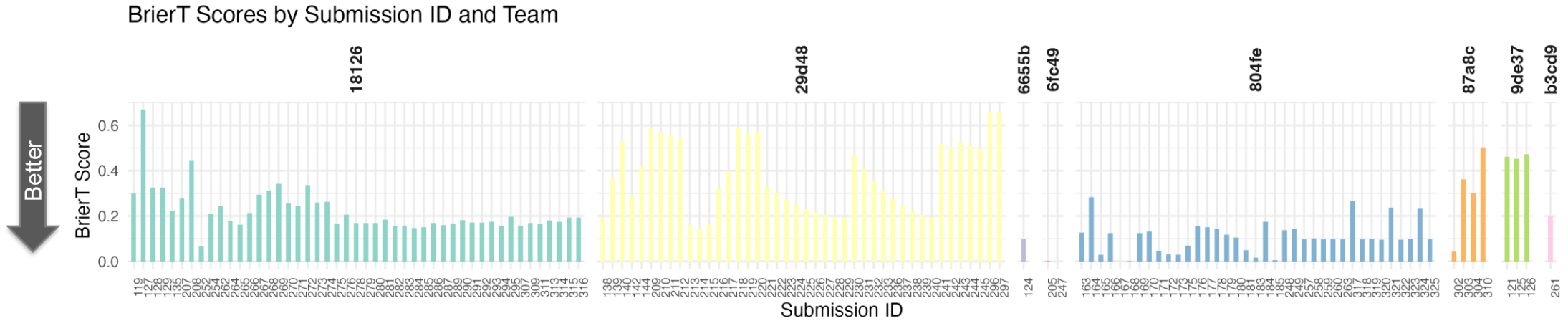
- Performance evaluations from our pilot study are not automatically generalizable to actual application scenarios.
- The extent to which such generalizations can be made is a function of how representative the test-sets used in the pilot study are of actual application scenarios.
- Serious inferences from data require attention to “uncertainties”. Here we are only presenting Descriptive Statistics.

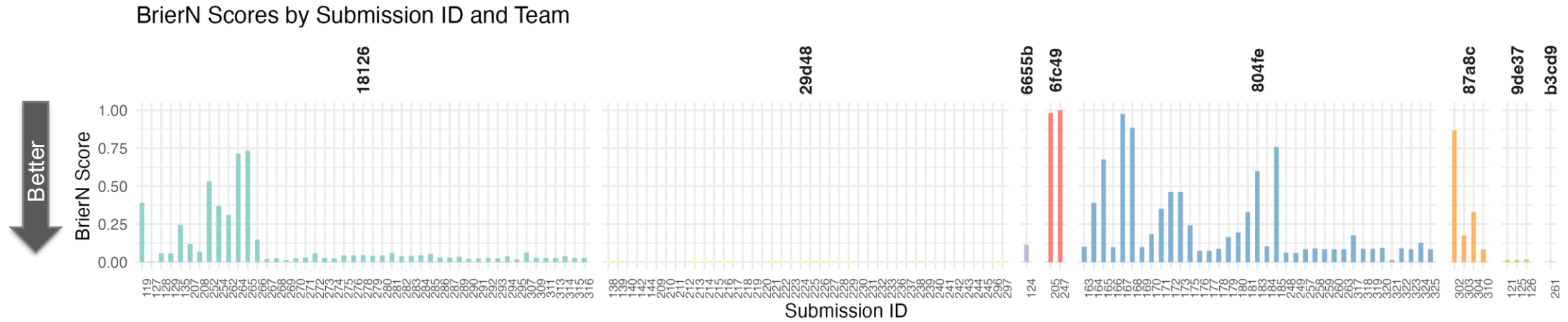
Round-1


AUC Scores by Submission ID and Team
Poor performing D-submission (against both generators)
Better
AUC Score
0.00
0.25
0.50
0.75
1.00
0dea0
18126
29d48
6655b
6fc49
804fe
87a8c
993ad
9de37
b3cd9
d718e
7 8
13 14 19 21 44 47 49 68
4
101 102 103 12 15 59 73 99
26 98
1 10 11 18 2 5 6 82 83 84 85 86 87 9
104 76 77 78 92 93 96
46 50 52
25 36 37
34
40
Submission ID
BrierT Scores by Submission ID and Team
Better
BrierT Score
Submission ID
BrierN Scores by Submission ID and Team
Better
BrierN Score
Submission ID
Good detector against both generators

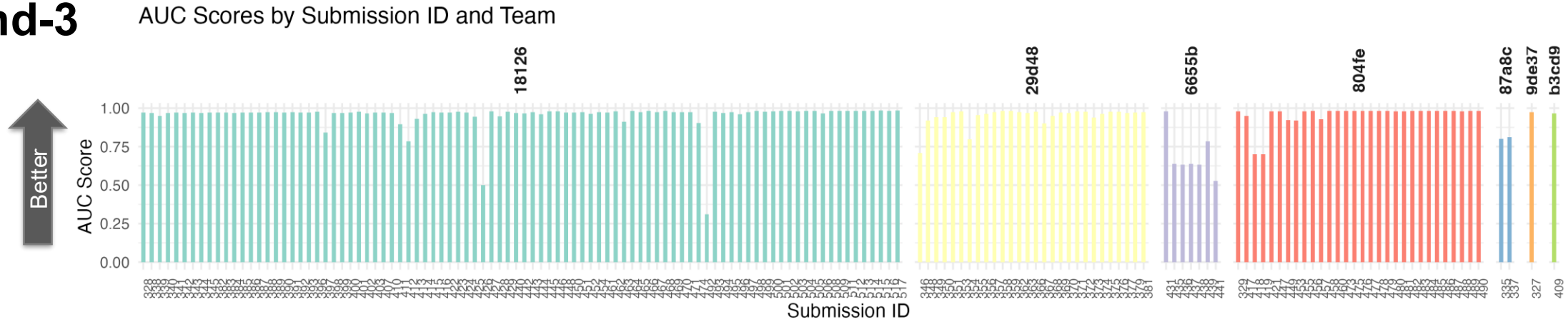
Round-2

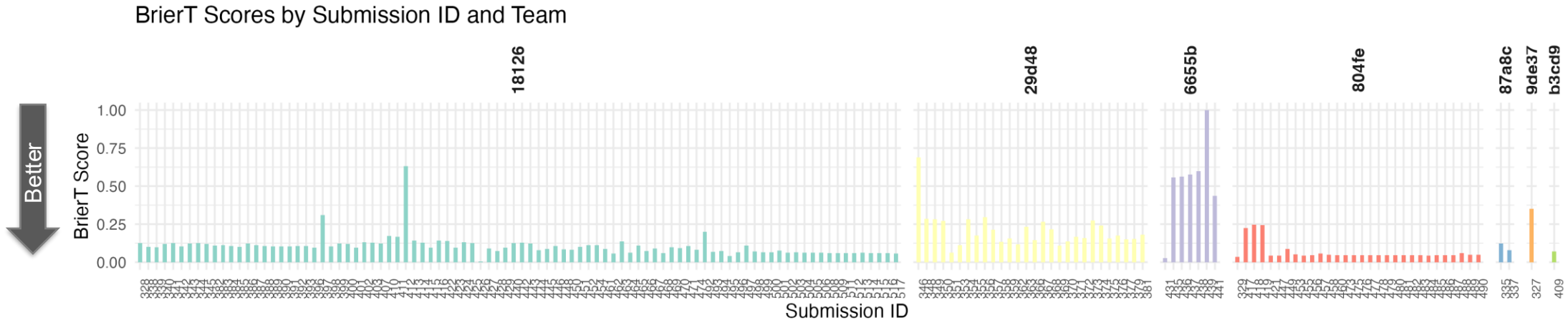

AUC Scores by Submission ID and Team
Better
AUC Score
1.00
0.75
0.50
0.25
0.00
18126
29d48
6655b
6fc49
804fe
87a8c
9de37
b3cd9
Submission ID

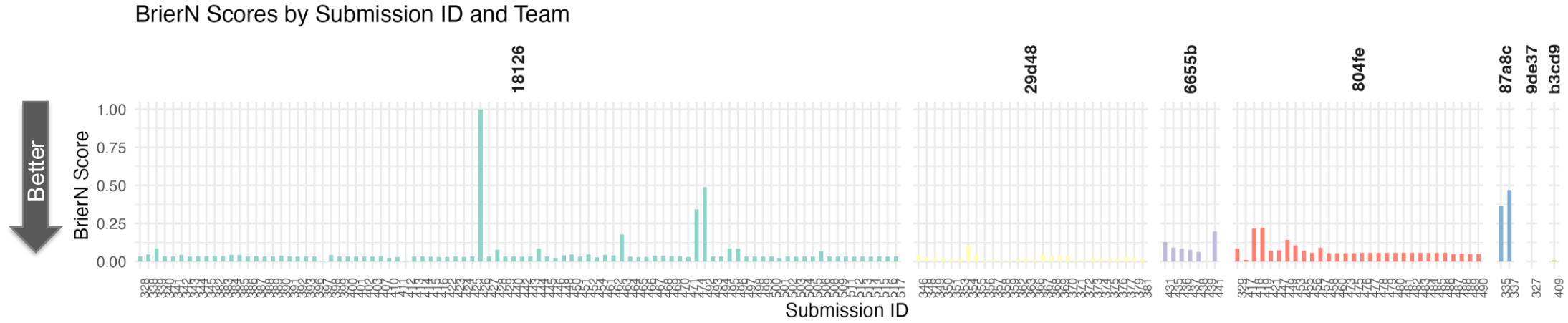

BrierT Scores by Submission ID and Team
Better
BrierT Score
0.6
0.4
0.2
0.0
18126
29d48
6655b
6fc49
804fe
87a8c
9de37
b3cd9
Submission ID


BrierN Scores by Submission ID and Team
Better
BrierN Score
1.00
0.75
0.50
0.25
0.00
18126
29d48
6655b
6fc49
804fe
87a8c
9de37
b3cd9
Submission ID

Round-3

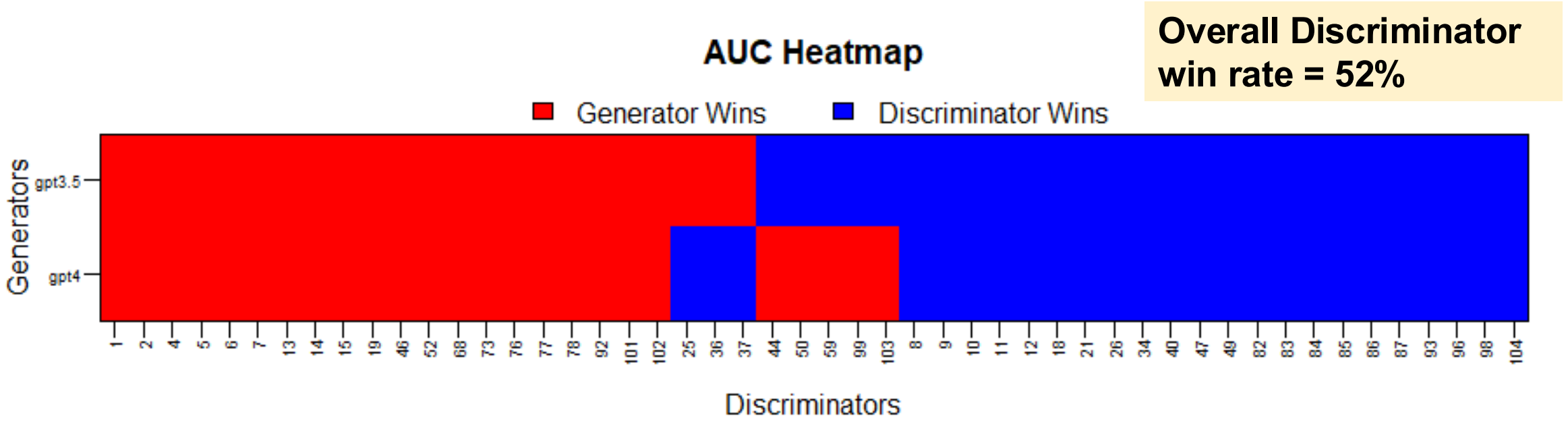

AUC Scores by Submission ID and Team
Better
AUC Score
1.00
0.75
0.50
0.25
0.00
18126
29d48
6655b
804fe
87a8c
9de37
b3cd9
Submission ID


BrierT Scores by Submission ID and Team
Better
BrierT Score
1.00
0.75
0.50
0.25
0.00
18126
29d48
6655b
804fe
87a8c
9de37
b3cd9
Submission ID


BrierN Scores by Submission ID and Team
Better
BrierN Score
1.00
0.75
0.50
0.25
0.00
18126
29d48
6655b
804fe
87a8c
9de37
b3cd9
Submission ID


LEAGUE OF GenAI
Detectors
Generators
6:03 / 14:42

TEST SET-1 (D-Round 1)



Overall Discriminator win rate = 52%

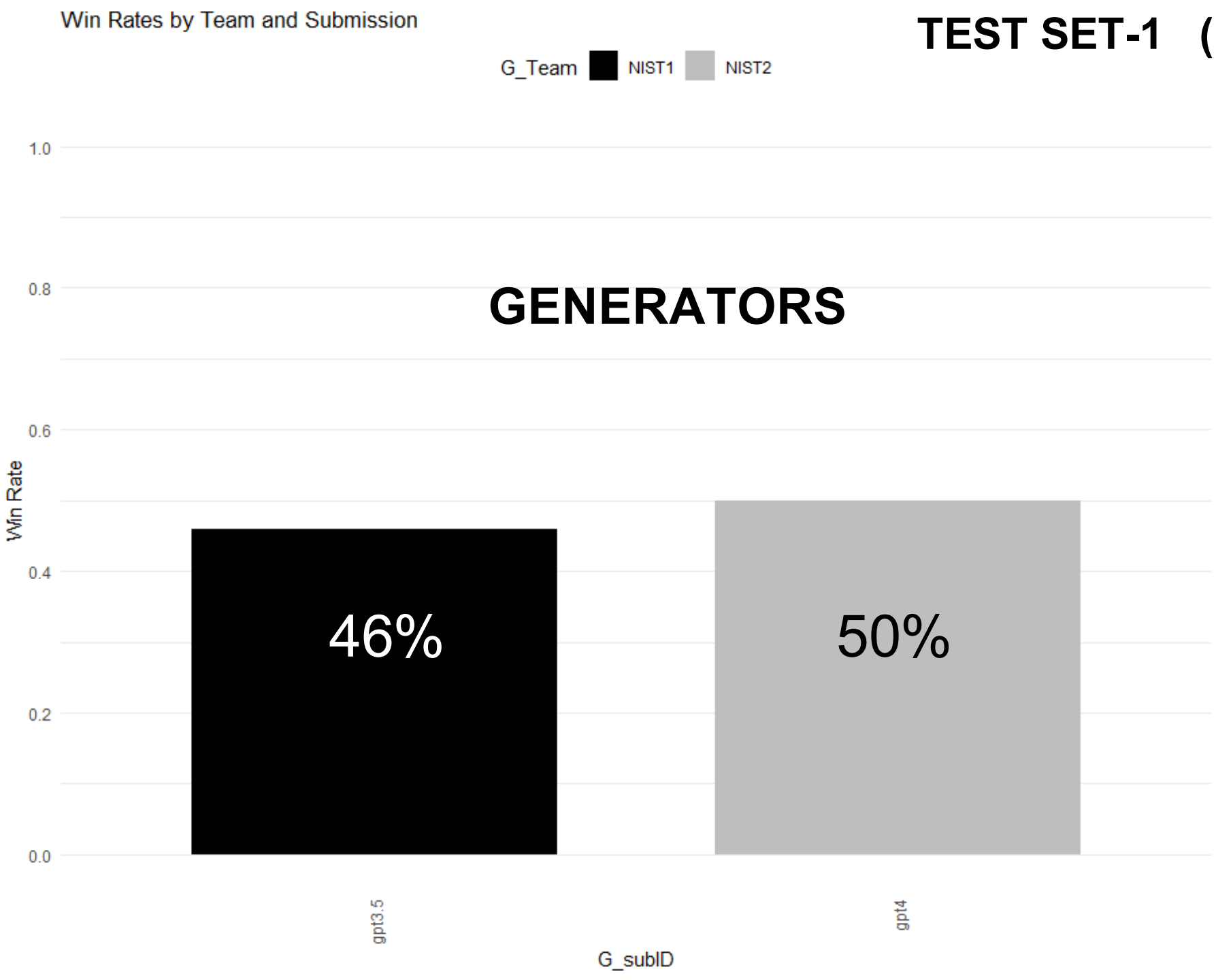
Detector submission **A** wins against Generator submission **B** if

(a) Detector AUC for the test set is greater than 0.5

AND (b) BrierT score for the test set is less than 0.25

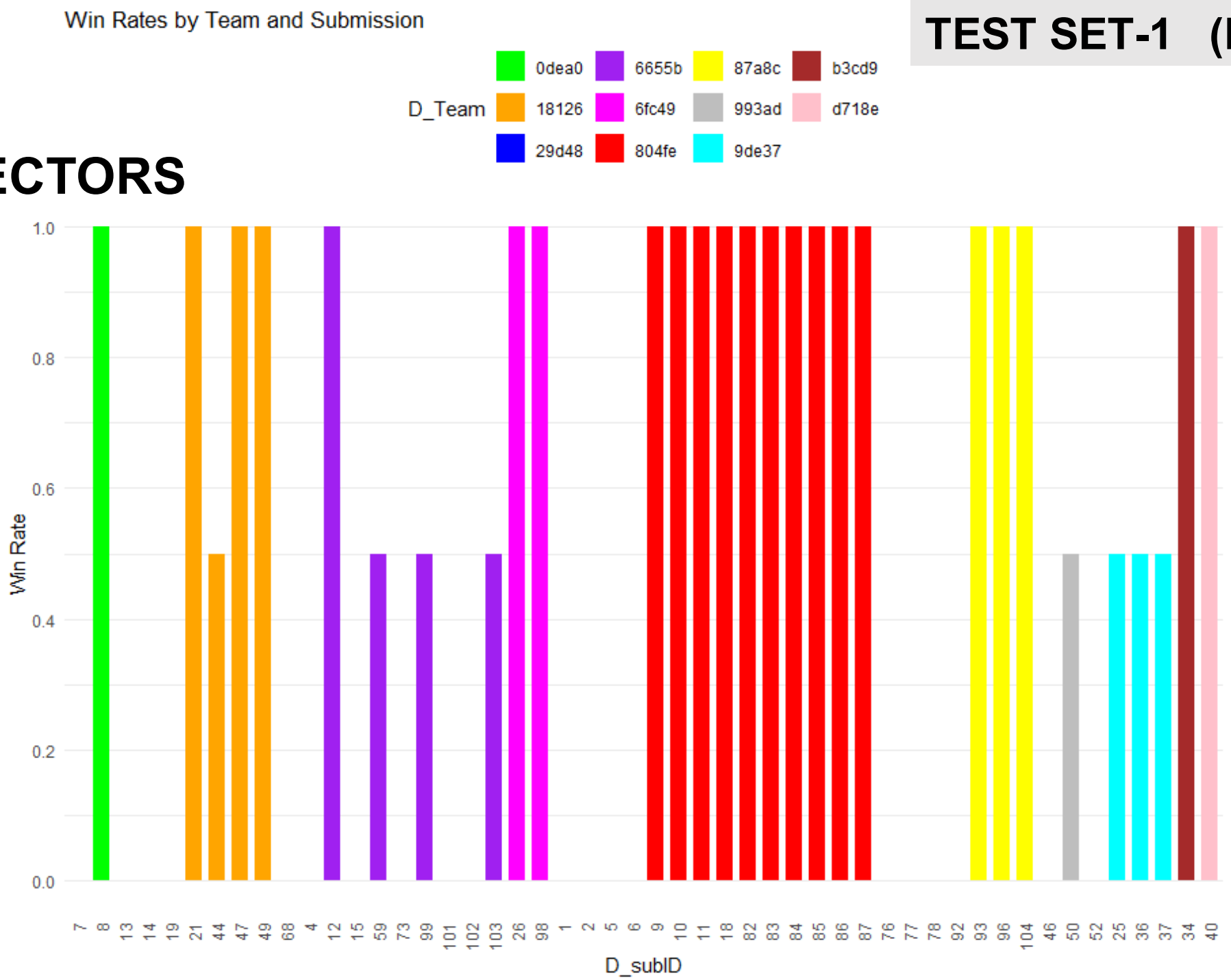
(sqrt BrierT is less than 0.5)

TEST SET-1 (D-Round 1)

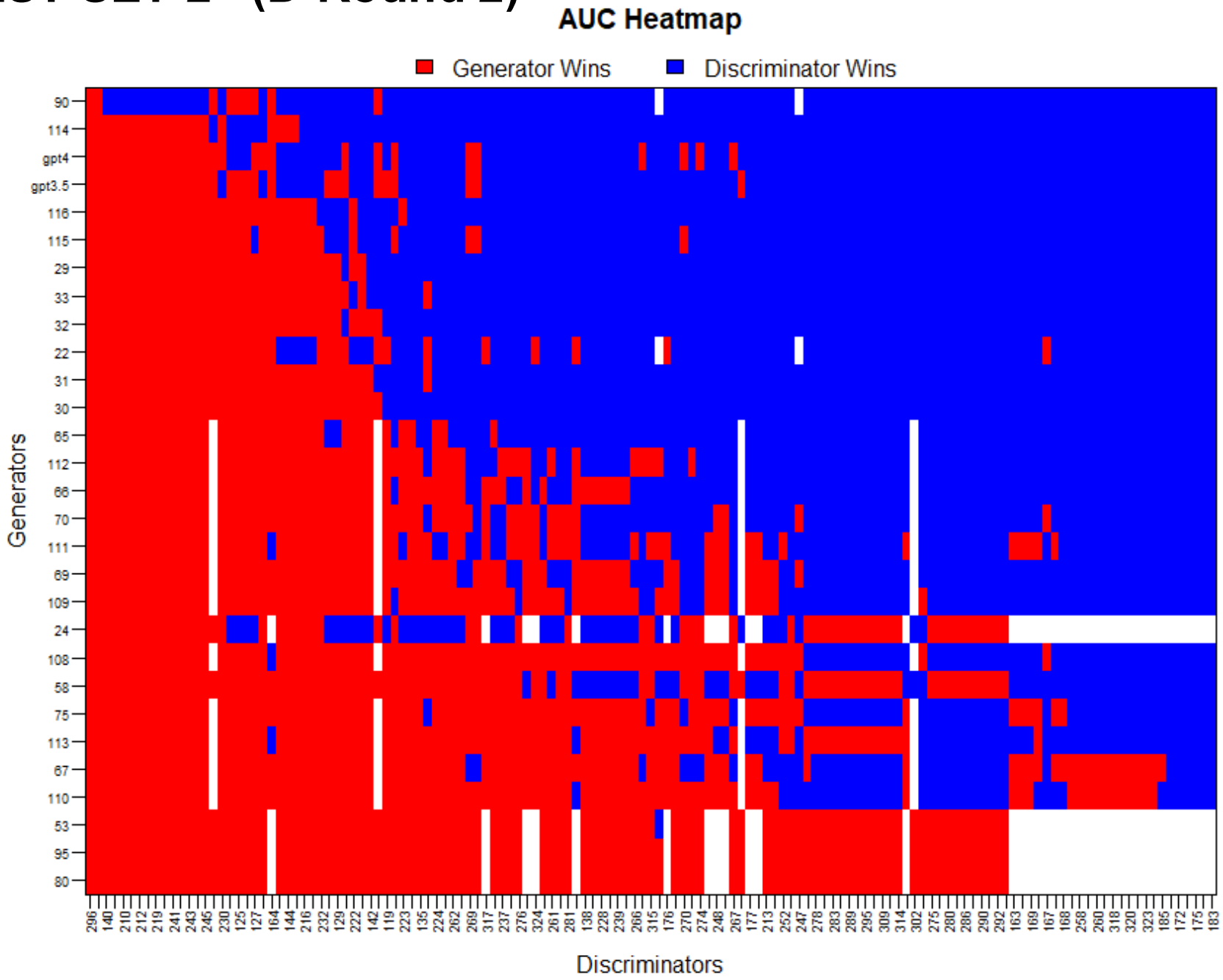


TEST SET-1 (D-Round 1)

DETECTORS

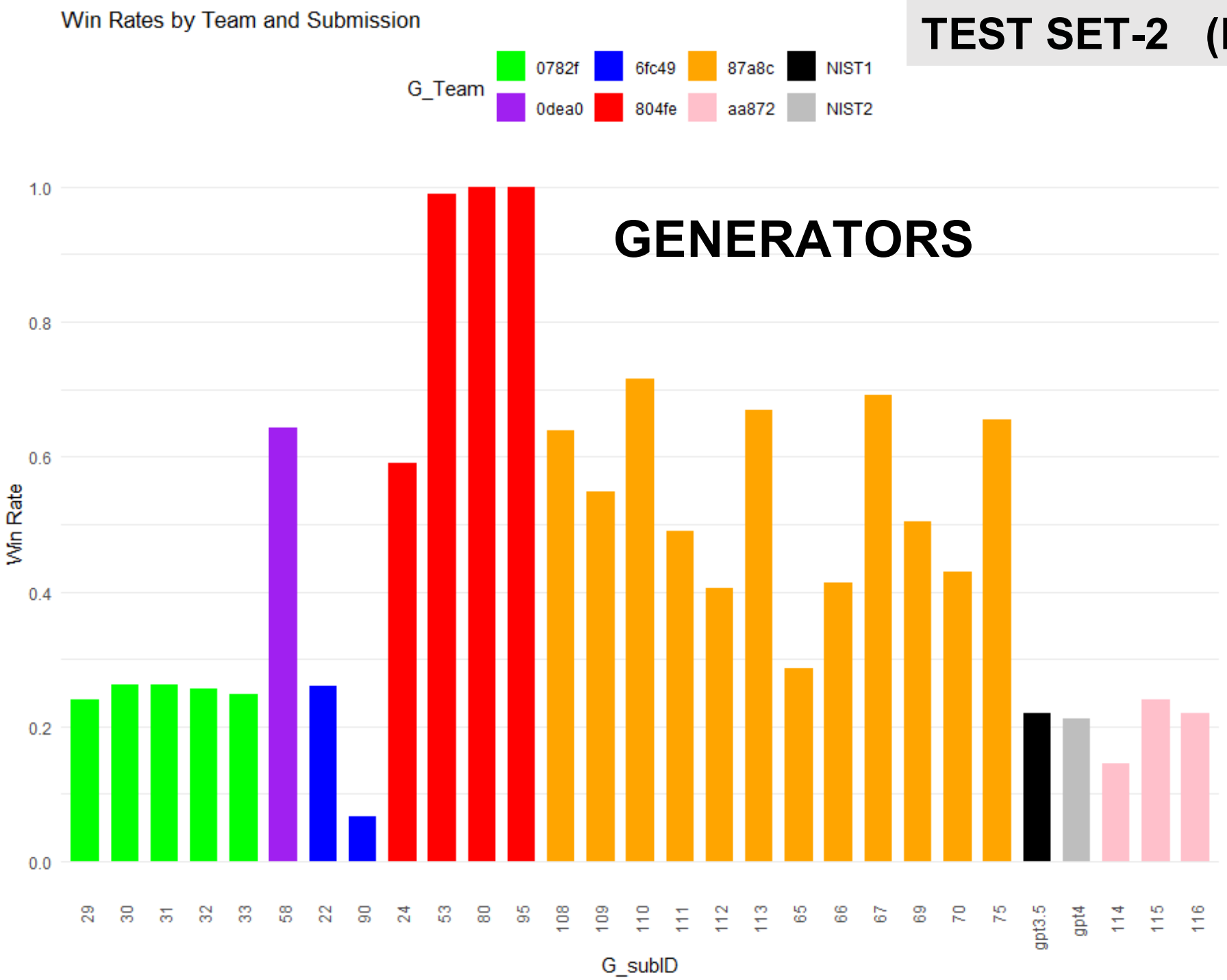


TEST SET-2 (D-Round 2)

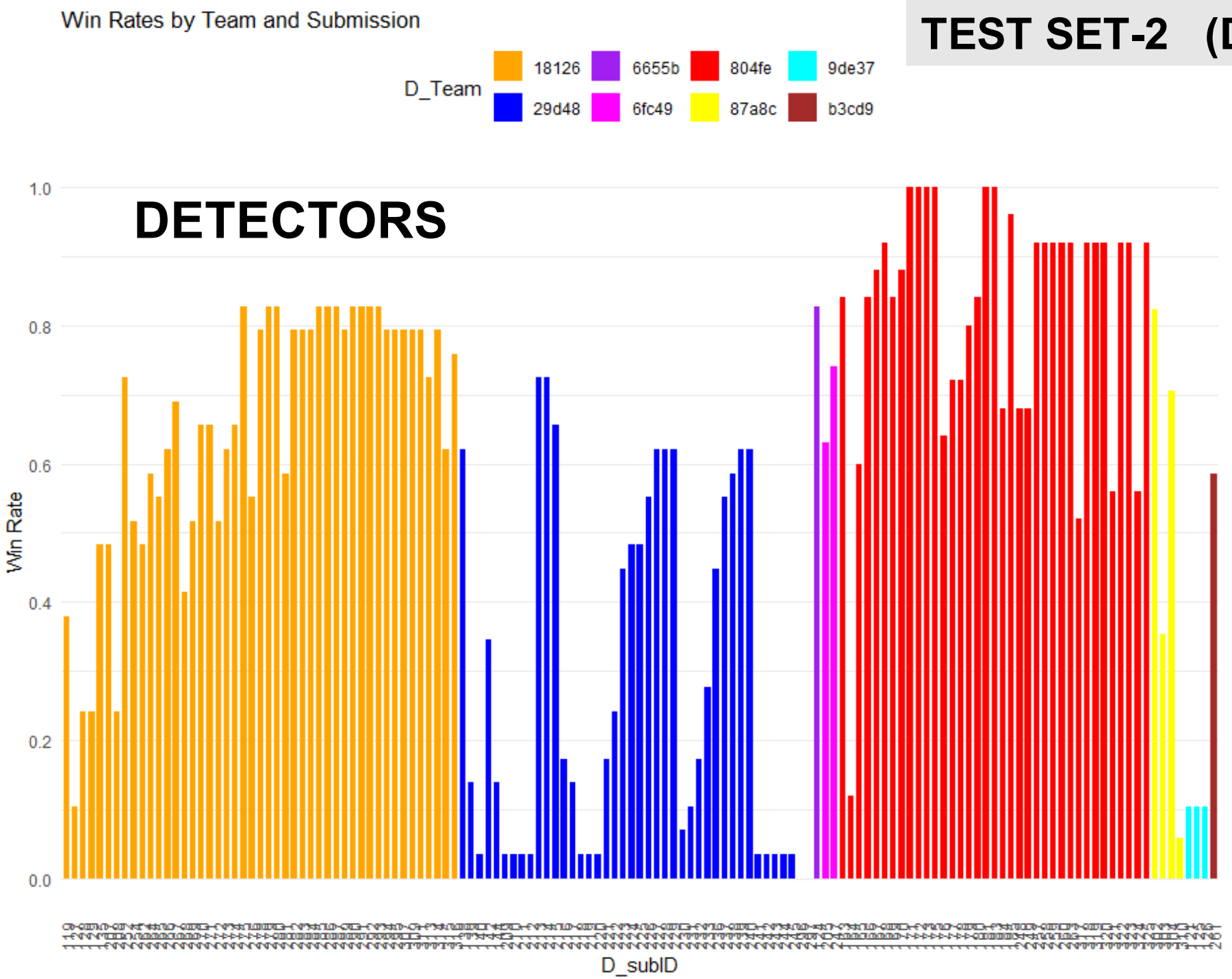


Overall Detector Win Rate = 56%

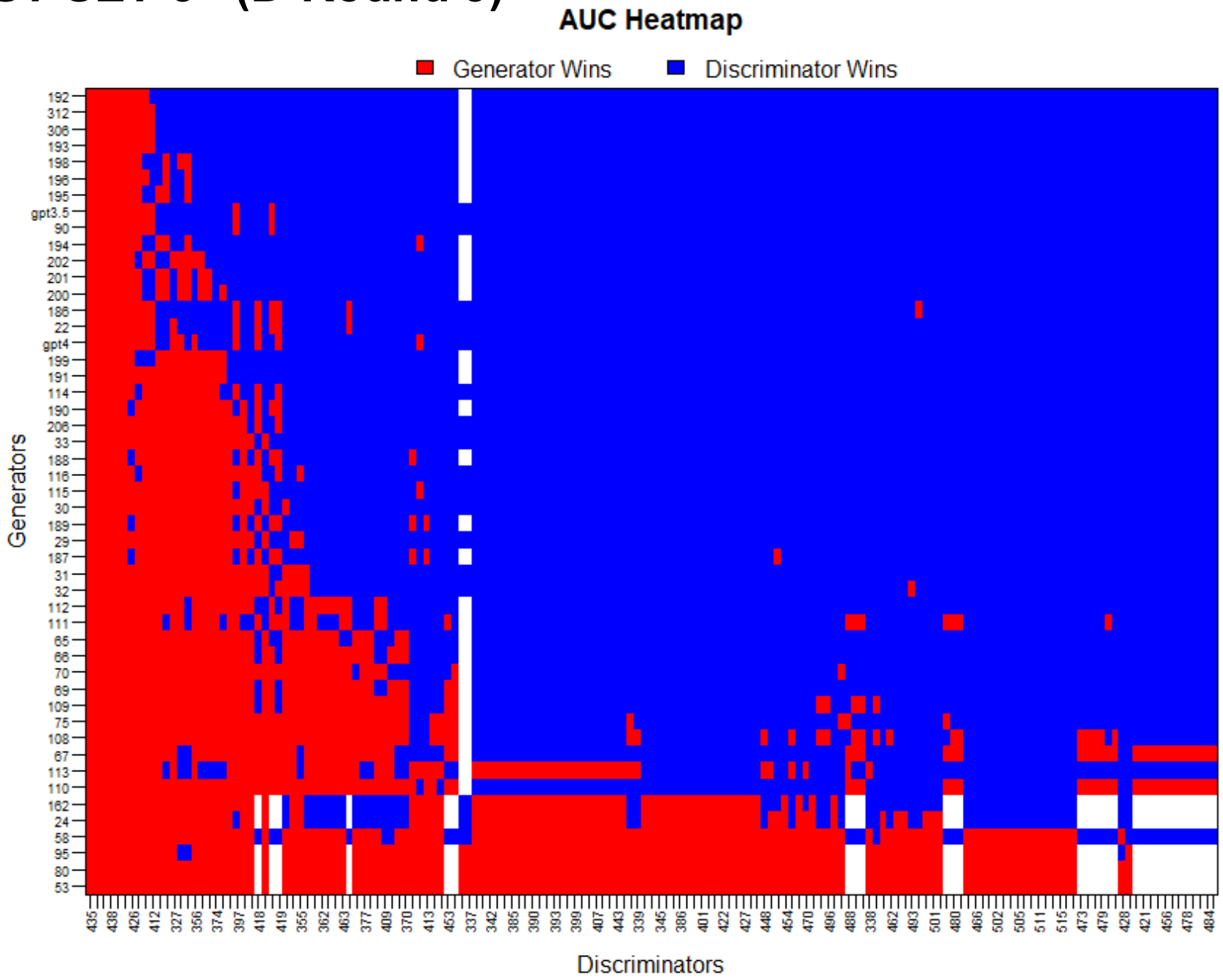
TEST SET-2 (D-Round 2)



TEST SET-2 (D-Round 2)

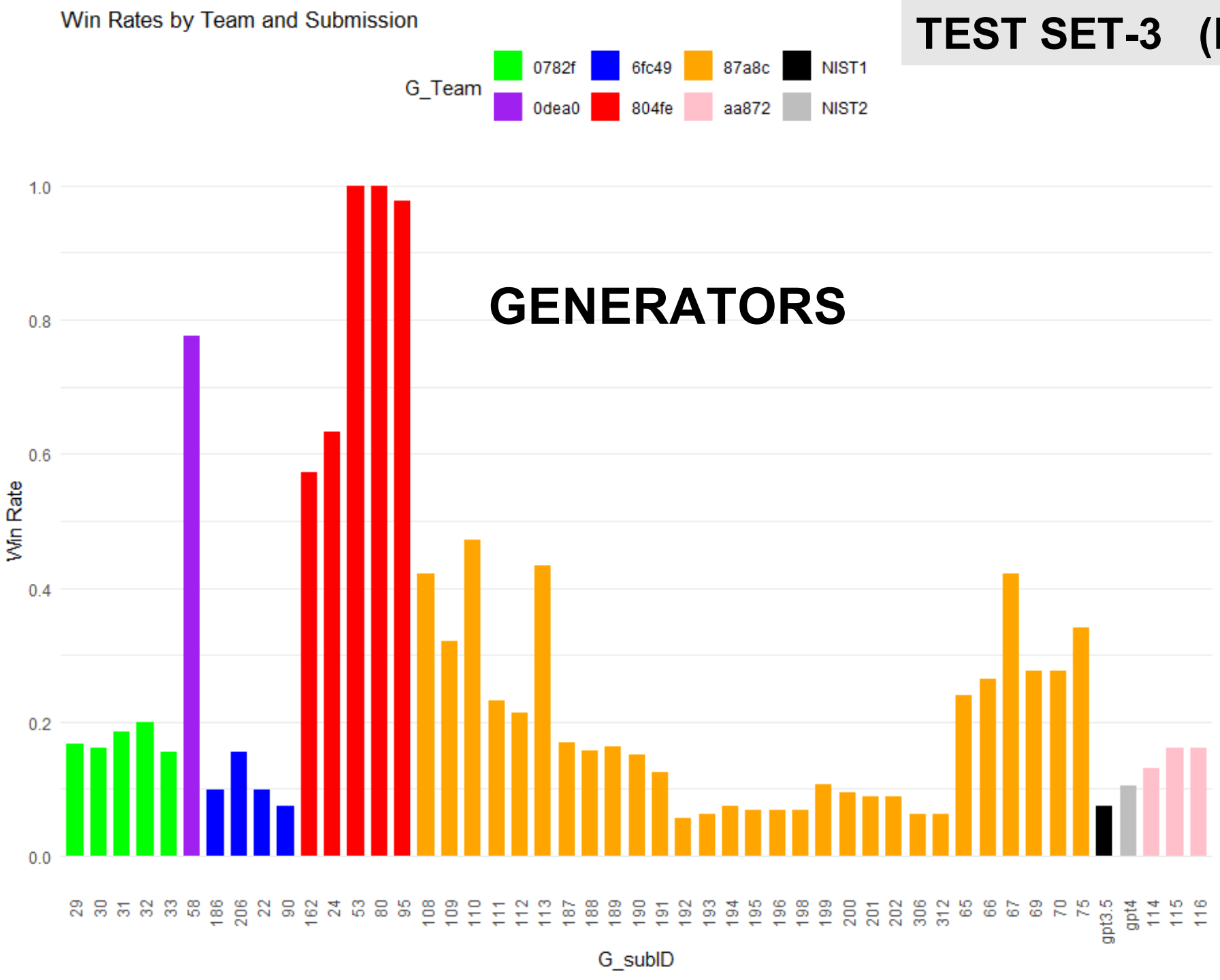


TEST SET-3 (D-Round 3)



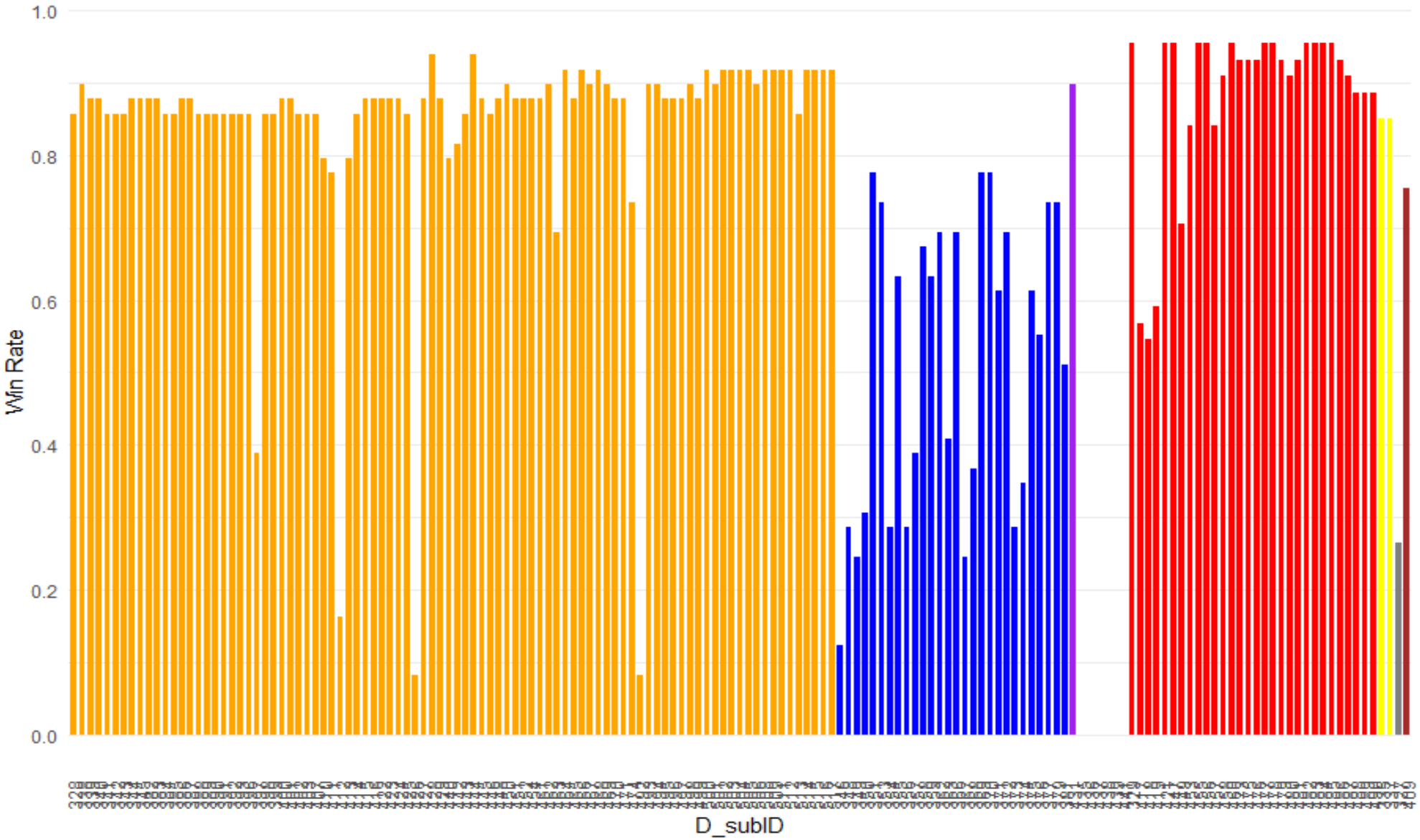
Overall Detector Win Rate = 76%

TEST SET-3 (D-Round 3)



TEST SET-3 (D-Round 3)

DETECTORS



Summary

- Except for a handful of generator submissions, detectors appear to do well against generators.
- The win rate for detectors showed improvement over the 3 rounds.
- We identified a few issues that need to be addressed before launching the main large-scale study.
  - For instance, some of the detector submissions appeared to be not “real” submissions. One of the submissions gave a detection score of 1 to every example. Such submissions will have to be removed before making inferences.

Strengths of the Study

- **Adversarial Testing Framework** – Models iteratively adapt, ensuring realistic challenges.
- **Robust Evaluation** – Multi-round testing with AUC, BrierT, BrierN, ROC.
- **Diverse Data** – Multiple topics & human summaries enhance evaluation.
- **Benchmarking** – Standardized, reproducible AI assessment framework.

Limitations of the Study

- **Narrow Scope** – Focuses on summarization, missing other text types.
- **AI-Human Overlap** – Increasing AI sophistication and prompt engineering blurs classification.

Challenges

- **AI Advancements** – Generators and Detectors evolve rapidly
- **Difficulty of Acquiring Representative Data** – Privacy hurdles, policy hurdles, copyright hurdles.
- **Participation of Frontier Modelers** – It would be beneficial to have more frontier (SOTA) modelers and models.

Future Work

- **Expand Text Domains**: Move beyond summarization to evaluate AI-generated content in creative writing, legal texts, and conversational AI.
- **Hybrid Systems**: Future research should account for hybrid human-AI content and prompt engineering advances.
- **Support Continuous Benchmarking**: Establish regular evaluations and competitions to track progress in AI generation and detection.

Q & A

Contact: genai-poc@nist.gov

https://ai-challenges.nist.gov/genai