Workshop – 31st August 2024

ASVspoof
Automatic Speaker Verification and Spoofing Countermeasures Challenge

A satellite event of

INTERSPEECH 2024

ASVspoof Workshop 2024

ORGANISERS

Héctor Delgado, Microsoft, Spain
Nicholas Evans, EURECOM, France
Jee-weon Jung, Carnegie Mellon University, USA
Tomi Kinnunen, University of Eastern Finland, Finland
Ivan Kukanov, KLASS Engineering and Solutions, Singapore
Kong Aik Lee, The Hong Kong Polytechnic University, Hong Kong
Xuechen Liu, National Institute of Informatics, Japan
Hye-jin Shim, Carnegie Mellon University, USA
Md Sahidullah, TCG CREST, India
Hemlata Tak, Pindrop, USA
Massimiliano Todisco, EURECOM, France
Xin Wang, National Institute of Informatics, Japan
Junichi Yamagishi, National Institute of Informatics, Japan

WORKSHOP SPONSORS

Pindrop

KLASS

SPONSORS

ACADEMY OF FINLAND

JST Japan Science and Technology Agency

Agency for Science, Technology and Research SINGAPORE

FRANCE 2030 PROGRAMME DE RECHERCHE CYBERSÉCURITÉ

anr agence nationale de la recherche

ASVspoof

Small, purpose collected datasets; low attack variability

1999

2013

First special session @ INTERSPEECH 2013

First common datasets, metrics and protocols; modest attack variability

ASVspoof initiative launched

2014

Large, standard datasets adapted to study spoofing; low attack variability

2015

ASVspoof 2015 LA

Improved methodology; broader attack variability

2017

ASVspoof 2017 PA

Tandem assessment

2019

ASVspoof 2019 LA + PA

Neural & acoustic waveform models; controlled replay scenario

2021

ASVspoof 2021 LA + PA + DF

2024

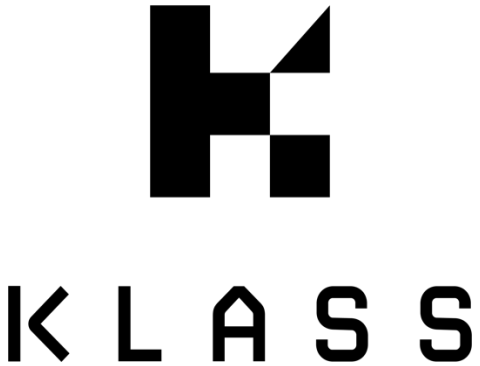
- LA: telephony encoding and transmission
- PA: real physical environments, variety of recording and replay devices
- DF: speech deepfake detection task

- a new dataset
- adversarial attacks, neural codec
- open condition
- metrics beyond EER

Workshop Acknowledgement

We would like thank to:

Pindrop (USA) and **KLASS Engineering and Solutions** (Singapore) for sponsoring the ASVspoof Workshop 2024.

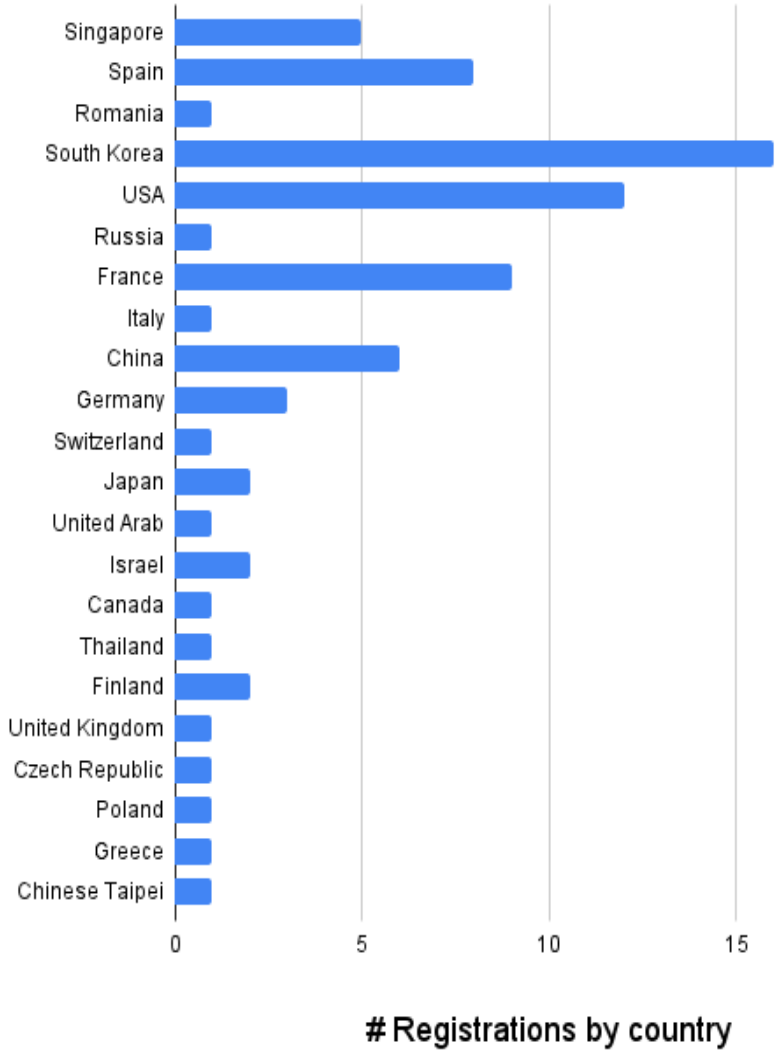


Workshop info

- Papers
  - Submitted: 37
  - Accepted: 29

Challenge summary: 1
ASVspoof 5 challenge: 25
Regular research papers: 3

- Workshop registered attendees: 67
- Venue: Kipriotis Hotels & Conference Center, Kos, Greece



Program
https://www.asvspoof.org/workshop2024_program

Schedule

From	To	Session
9:45	10:00	Opening, Welcome Message and Logistics
10:00	10:45	Challenge Summary
10:45	12:25	ASVspoof 5 Site Presentations Session 1
12:25	13:35	Lunch (1 hour 10 minutes)
13:35	15:10	ASVspoof 5 and beyond
15:10	15:40	Coffee break (30 minutes)
15:40	17:00	ASVspoof 5 Site Presentations Session 2
17:00	18:00	ASVspoof 5 Forum

Workshop – 31st August 2024

ASVspoof
Automatic Speaker Verification and Spoofing Countermeasures Challenge

A satellite event of

INTERSPEECH 2024

ASVspoof 5

Crowdsourced Speech Data, Deepfakes, and Adversarial Attacks at Scale

Xin Wang, National Institute of Informatics, Japan
Héctor Delgado, Microsoft, Spain
Hemlata Tak, Pindrop, USA
Jee-weon Jung, Carnegie Mellon University, USA
Hye-jin Shim, Carnegie Mellon University, USA
Massimiliano Todisco, EURECOM, France
Ivan Kukanov, KLASS Engineering and Solutions, Singapore
Xuechen Liu, National Institute of Informatics, Japan
Md Sahidullah, TCG CREST, India
Tomi Kinnunen, University of Eastern Finland, Finland
Nicholas Evans, EURECOM, France
Kong Aik Lee, The Hong Kong Polytechnic University, Hong Kong
Junichi Yamagishi, National Institute of Informatics, Japan

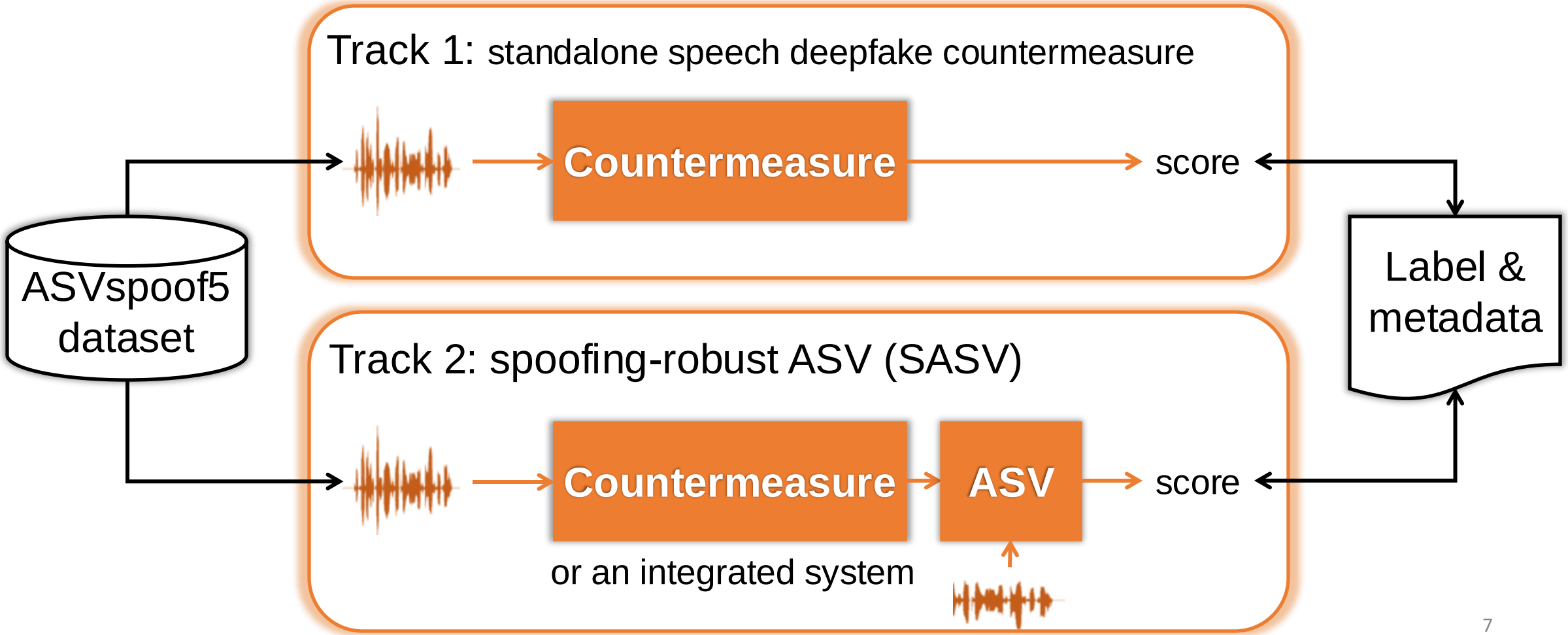
ACADEMY OF FINLAND
JST Japan Science and Technology Agency
Agency for Science, Technology and Research SINGAPORE
FRANCE 2030 PROGRAMME DE RECHERCHE CYBERSÉCURITÉ
anr agence nationale de la recherche

ASVspoof 5

Organisers: dataset creation

Participants: system building & scoring

Organisers: evaluation



ASVspoof 5

Organisers: dataset creation

Participants: system building & scoring

Organisers: evaluation

Track 1: standalone speech deepfake countermeasure

A new dataset

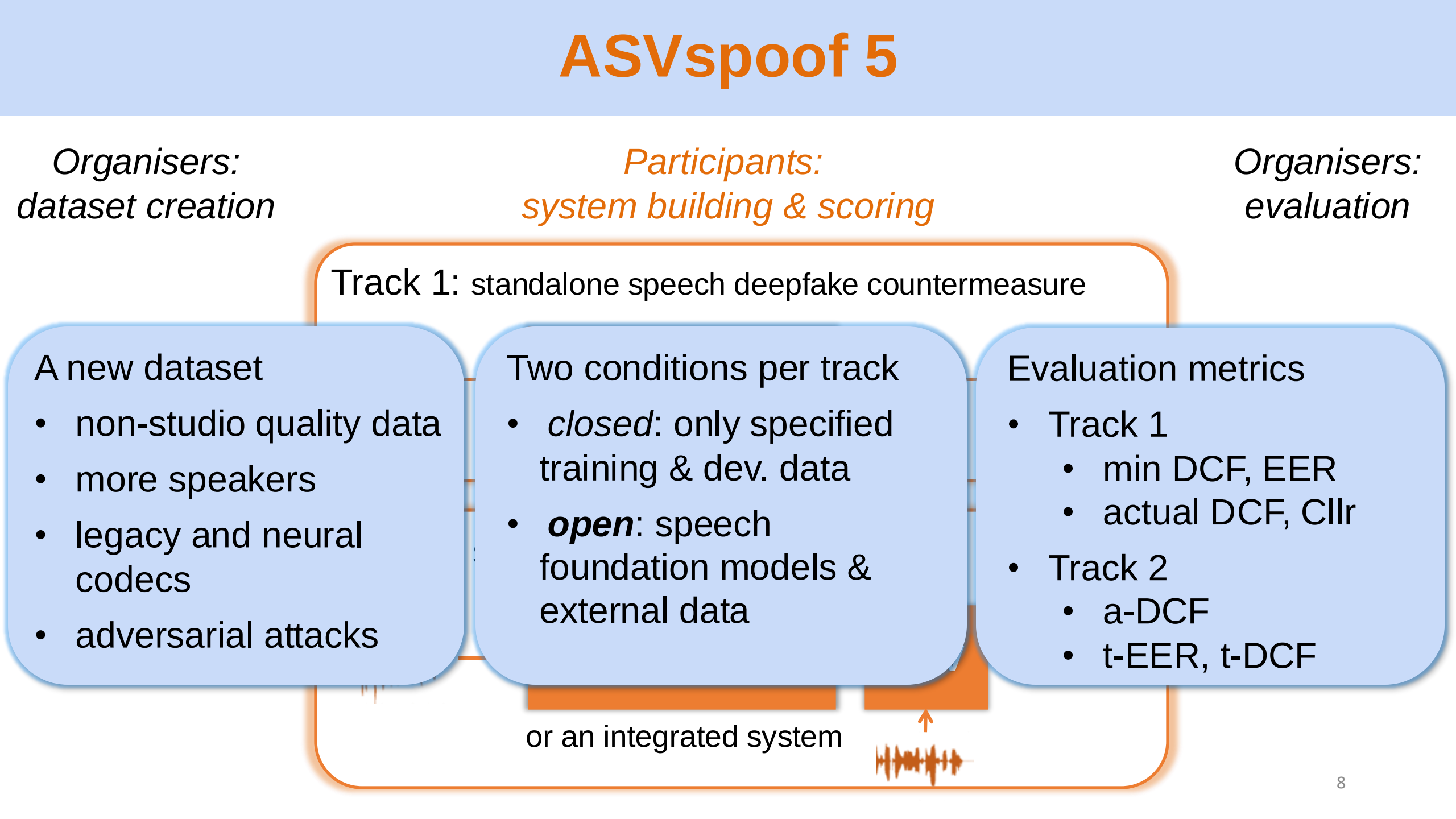
- non-studio quality data
- more speakers
- legacy and neural codecs
- adversarial attacks

Two conditions per track

- *closed*: only specified training & dev. data
- ***open***: speech foundation models & external data

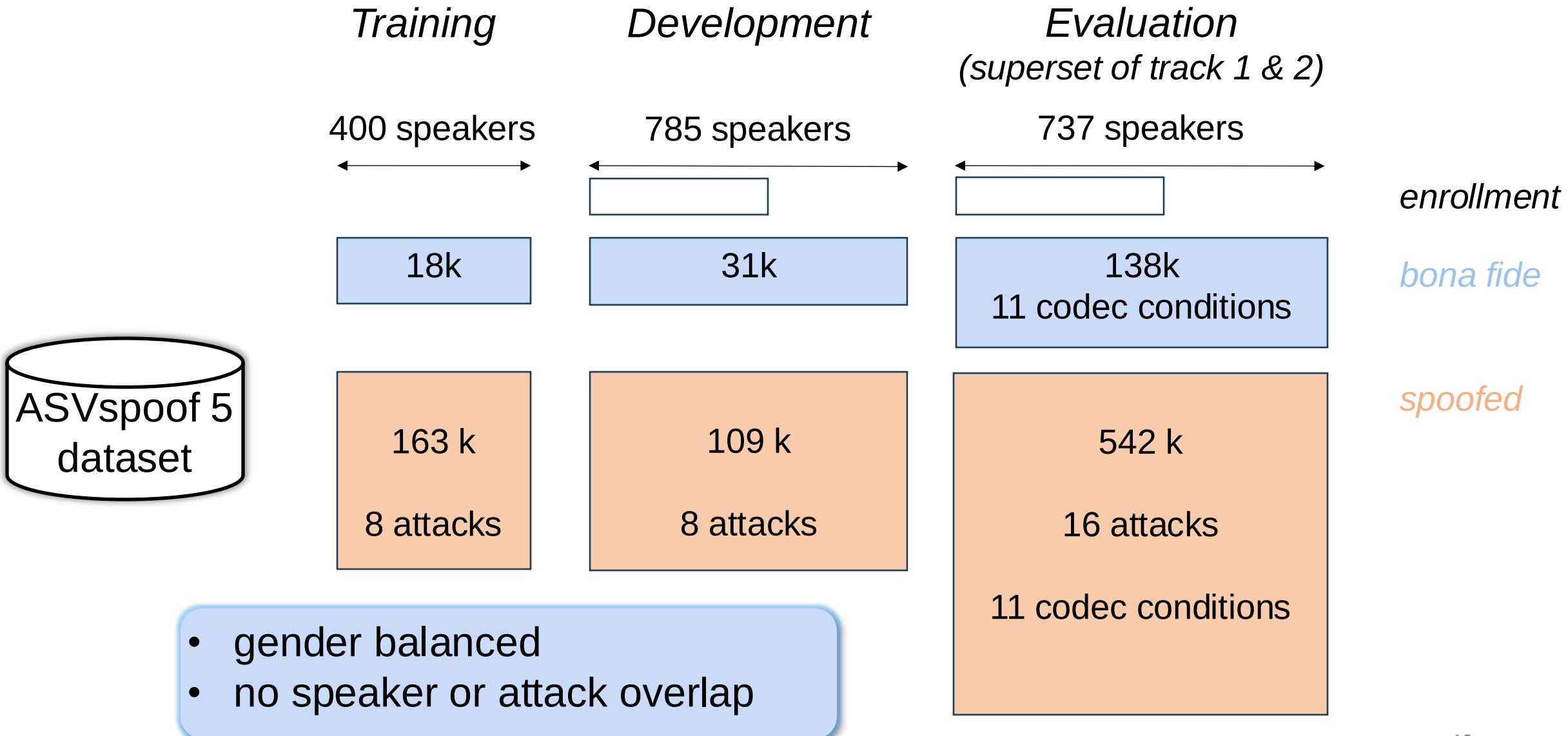
Evaluation metrics

- Track 1
  - min DCF, EER
  - actual DCF, Cllr
- Track 2
  - a-DCF
  - t-EER, t-DCF



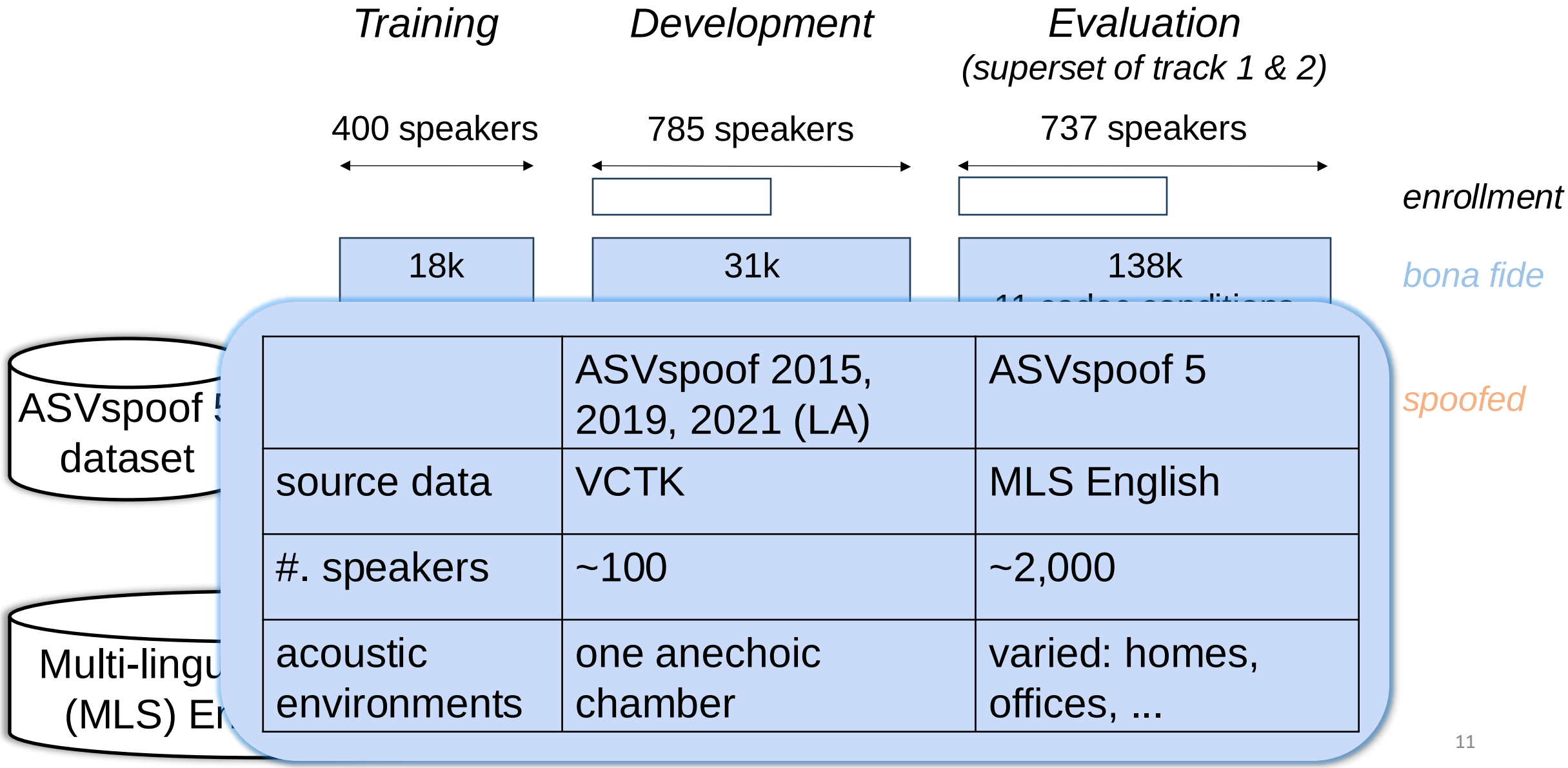
ASVspoof 5 dataset

ASVspoof 5 dataset



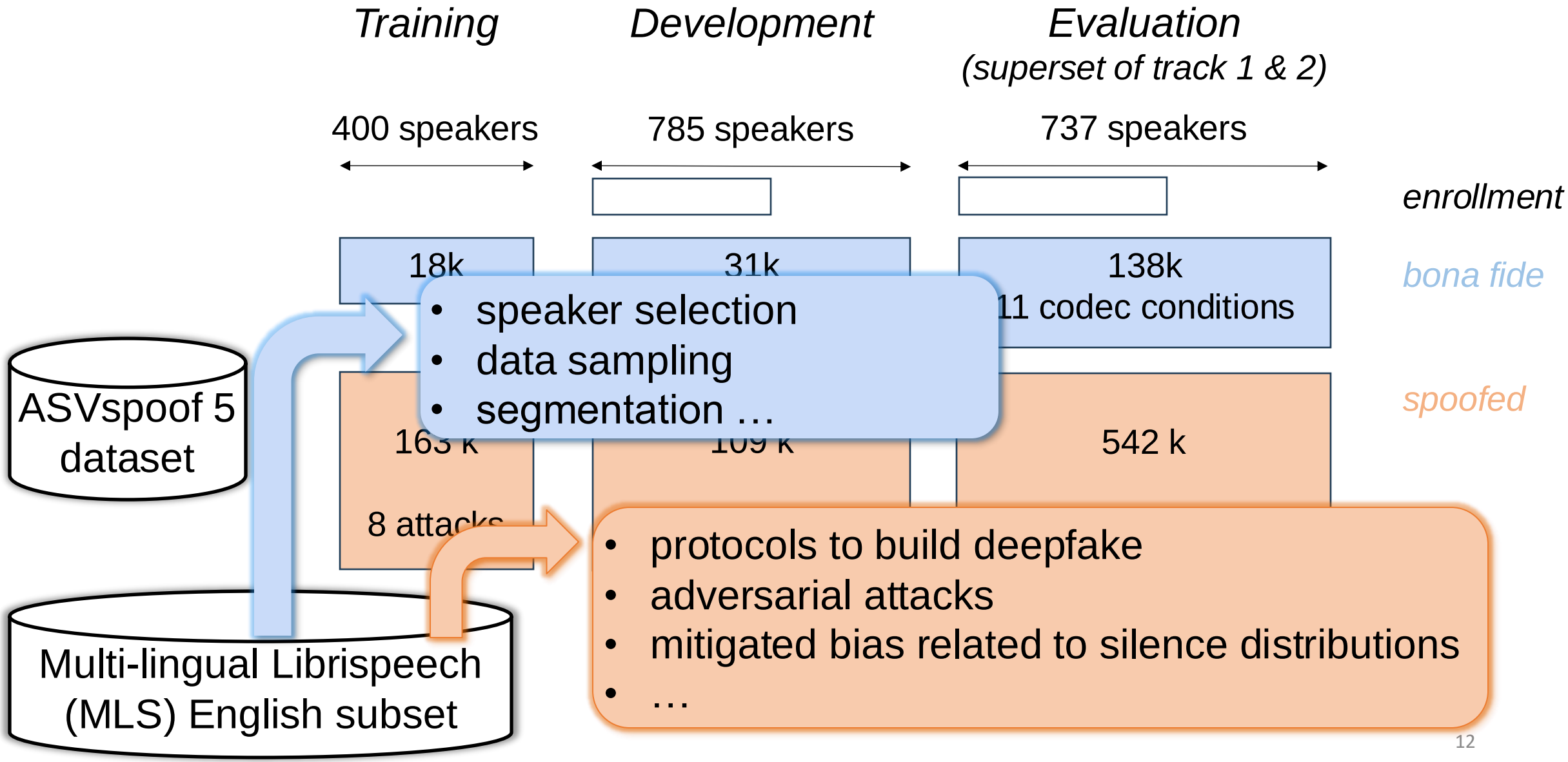
- gender balanced
- no speaker or attack overlap

ASVspoof 5 dataset



	ASVspoof 2015, 2019, 2021 (LA)	ASVspoof 5
source data	VCTK	MLS English
#. speakers	~100	~2,000
acoustic environments	one anechoic chamber	varied: homes, offices, ...

ASVspoof 5 dataset



ASVspoof 5 dataset: spoofed data

Attack	Input	Input processor	Acoustic model	Speaker embedding	Acoustic feature	Waveform model	Post-processing
A01	text	DNN-encoder	Glow	x-vector	latent	HifiGAN	-
A02	text	DNN-encoder	Glow	y-vector	latent	HifiGAN	-
A03	text	DNN-encoder	Glow	ECAPA	latent	HifiGAN	-
A04	text	DNN-encoder	Diffusion	x-vector	Mel-spec	WaveGrad	-
A05	text	DNN-encoder	Diffusion	y-vector	Mel-spec	WaveGrad	-
A06	text	DNN-encoder	Diffusion	ECAPA	Mel-spec	WaveGrad	-
A07	text	DNN-encoder	FastPitch	x-vector	Mel-spec	HifiGAN	-
A08	text	DNN-encoder	VITS	x-vector	latent	HifiGAN	-
A09	text	NLP	FS-like	GST	log-spec	HifiGAN	-
A10	text	NLP	FS-like	GST	log-spec	HifiGANv2	-
A11	text	DNN-encoder	Tacotron2	G2E	Mel-sepc	WaveGrad	-
A12	text	NLP	-	-	-	unit.	-
A13	speech	DNN-encoder	StarGANv2	style-encoder	PWG	-	
A14	text	DNN-encoder	YourTTS	-	latent	HifiGAN	-
A15	speech	VAE	GAN	-	Mel-spec	WaveNet	-
A16	speech	ASR	DNN	CAM++	Mel-spec	HifiGAN	-
A21	text	NLP	FS-based	GST	log-spec	BigVGAN	-
A22	text	NLP	FS+ProsodyTransfer	GST	log-spec	BigVGAN	-
A28							
A26							
A17							
A19							
A24							
A25							
A29							
A23							
A20							
A18							
A27							
A31							
A32							
A30	A18	-	-	-	-	-	Malahide+Malacopula

popular text-to-speech & voice conversion algorithms

- GAN, diffusions …
- VITS, glow …
- x-vector, ECAPA

Training

Development

Evaluation

ASVspoof 5 dataset: spoofed data (eval. set)

Attack	Input	Input processor	Acoustic model	Speaker embedding	Acoustic feature	Waveform model	Post-processing
A21	text	NLP	FS-based	GST	log-spec	BigVGAN	-
A22	text	NLP	FS+ProsodyTransfer	GST	log-spec	BigVGAN	-
A28	text	DNN-encoder	YourTTS(pre.)	H/ASP	latent	HifiGAN	-
A26	speech	ASR	DNN+F0 est.	CAM++	Mel-spec	HifiGAN	original genuine noise
A17	text	BERT-based	Transformer-based	x-vector	DNN-latent	HifiGAN	-
A19	text	NLP	MaryTTS	-	-	unit-selec	-
A24	speech	PPG	DNN	x-vector	LSP	HifiGAN	-
A25	speech	DNN-encoder	DiffVC	latent	Mel-spec	HifiGAN	-
A29	text	DNN-encoder	XTTS(pre.)	ECAPA2	latent	HifiGAN	-
A23	A09	-	-	-	-	-	Malafide
A20	A12	-	-	-	-	-	Malafide
A18	A17	-	-	-	-	-	Malafide
A27	A26	-	-	-	-	-	Malacopula
A31	A22	-	-	-	-	-	Malacopula
A32	A25	-	-	-	-	-	Malacopula
A30	A18	-	-	-	-	-	Malafide+Malacopula

FS: FastSpeech
NLP: natural-language-process-based front-end
GST: global style token

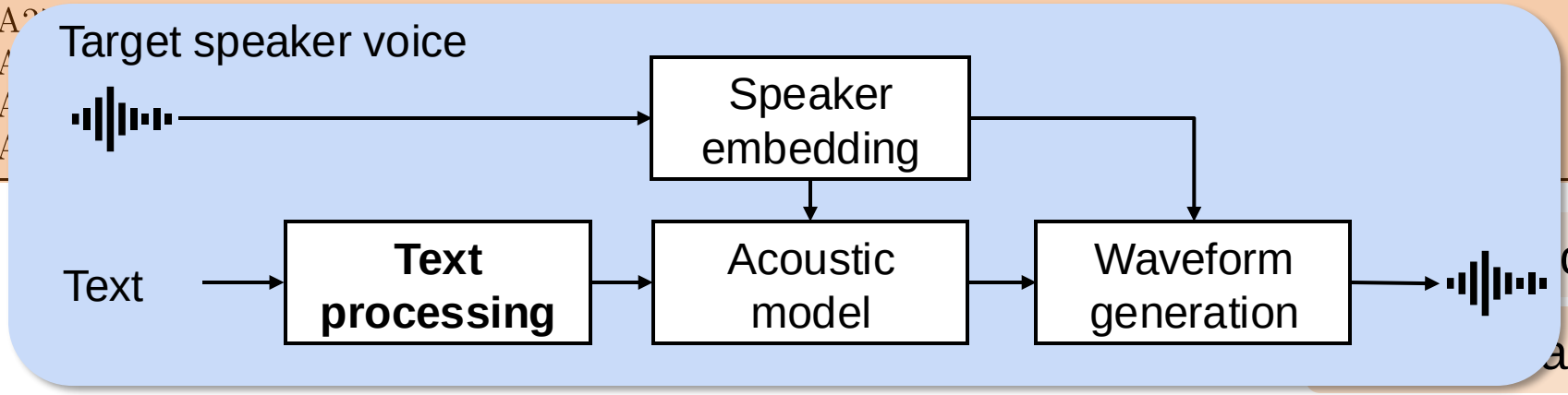
varied attacks of dev, train

unknown attacks

adversarial attacks

ASVspoof 5 dataset: spoofed data (eval. set)

Attack	Input	Input processor	Acoustic model	Speaker embedding	Acoustic feature	Waveform model	Post-processing
A21	text	NLP	FS-based	GST	log-spec	BigVGAN	-
A22	text	NLP	FS+ProsodyTransfer	GST	log-spec	BigVGAN	-
A28	text	DNN-encoder	YourTTS(pre.)	H/ASP	latent	HifiGAN	-
A26	speech	ASR	DNN+F0 est.	CAM++	Mel-spec	HifiGAN	original genuine noise
A17	text	BERT-based	Transformer-based	x-vector	DNN-latent	HifiGAN	-
A19	text	NLP	MaryTTS	-	-	unit-selec	-
A24	speech	PPG	DNN	x-vector	LSP	HifiGAN	-
A25	speech	DNN-encoder	DiffVC	latent	Mel-spec	HifiGAN	-
A29	text	DNN-encoder	XTTS(pre.)	ECAPA2	latent	HifiGAN	-
A23	A09	-	-	-	-	-	Malafide
A20	A12	-	-	-	-	-	Malafide
A18	A17	-	-	-	-	-	Malafide
A						-	Malacopula
A						-	Malacopula
A						-	Malacopula
A						-	Malafide+Malacopula



zero-shot text-to-speech synthesis

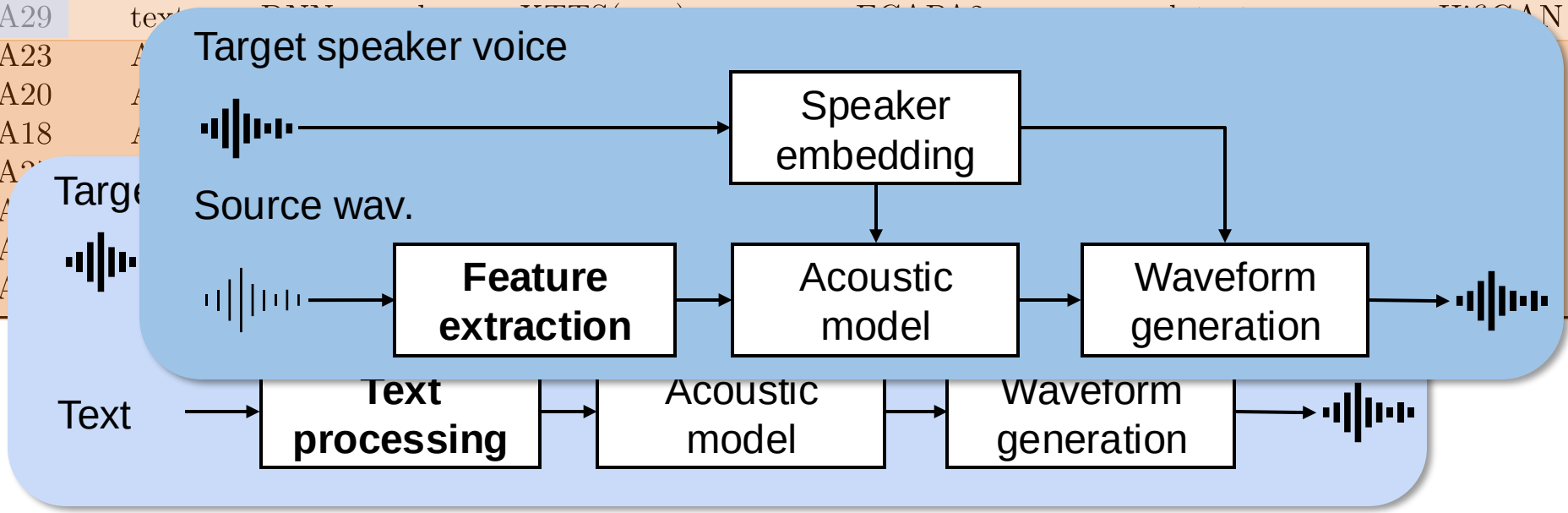
cks of dev, train

attacks

adversarial attacks

ASVspoof 5 dataset: spoofed data (eval. set)

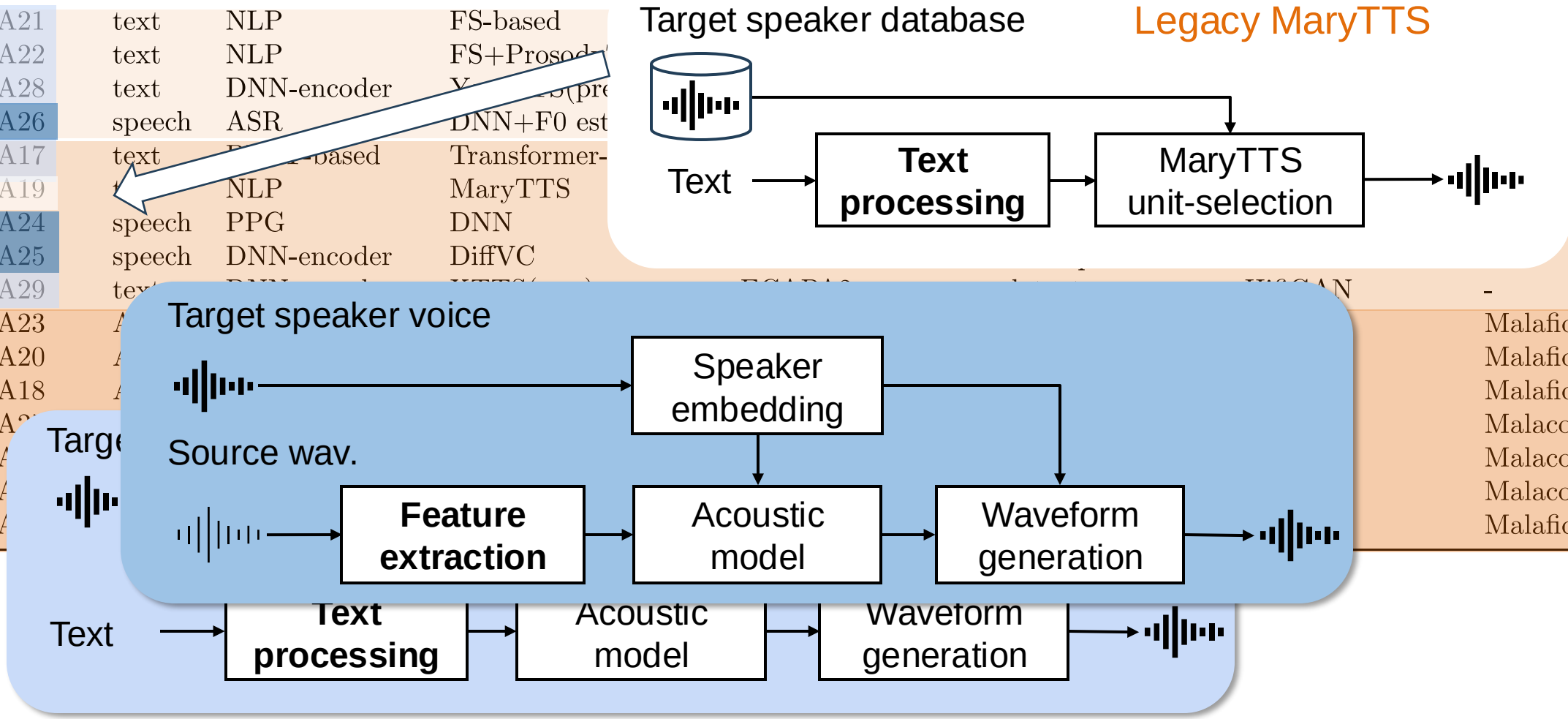
Attack	Input	Input processor	Acoustic model	Speaker embedding	Acoustic feature	Waveform model	Post-processing
A21	text	NLP	FS-based	GST	log-spec	BigVGAN	-
A22	text	NLP	FS+ProsodyTransfer	GST	log-spec	BigVGAN	-
A28	text	DNN-encoder	YourTTS(pre.)	H/ASP	latent	HifiGAN	-
A26	speech	ASR	DNN+F0 est.	CAM++	Mel-spec	HifiGAN	original genuine noise
A17	text	BERT-based	Transformer-based	x-vector	DNN-latent	HifiGAN	-
A19	text	NLP	MaryTTS	-	-	unit-selec	-
A24	speech	PPG	DNN	x-vector	LSP	HifiGAN	-
A25	speech	DNN-encoder	DiffVC	latent	Mel-spec	HifiGAN	-
A29	tex						-
A23							Malafide
A20							Malafide
A18							Malafide
							Malacopula
							Malacopula
							Malacopula
							Malafide+Malacopula



zero-shot (any-to-any) voice conversion

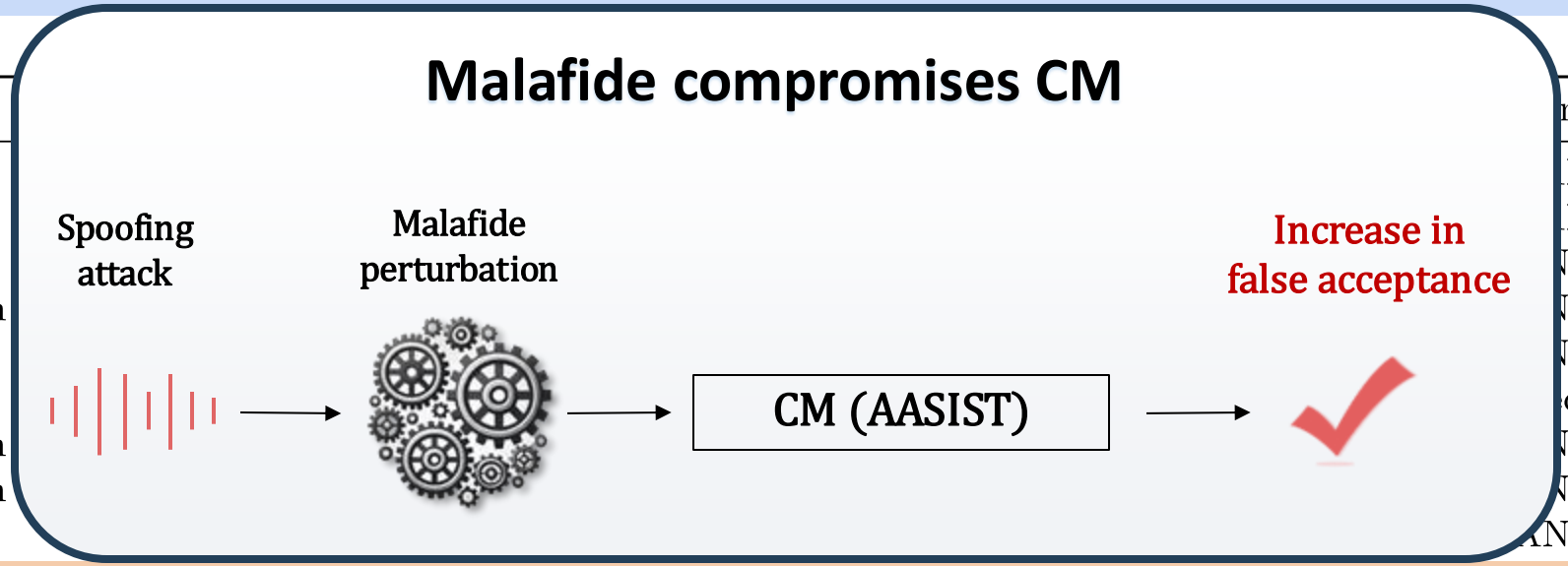
ASVspoof 5 dataset: spoofed data (eval. set)

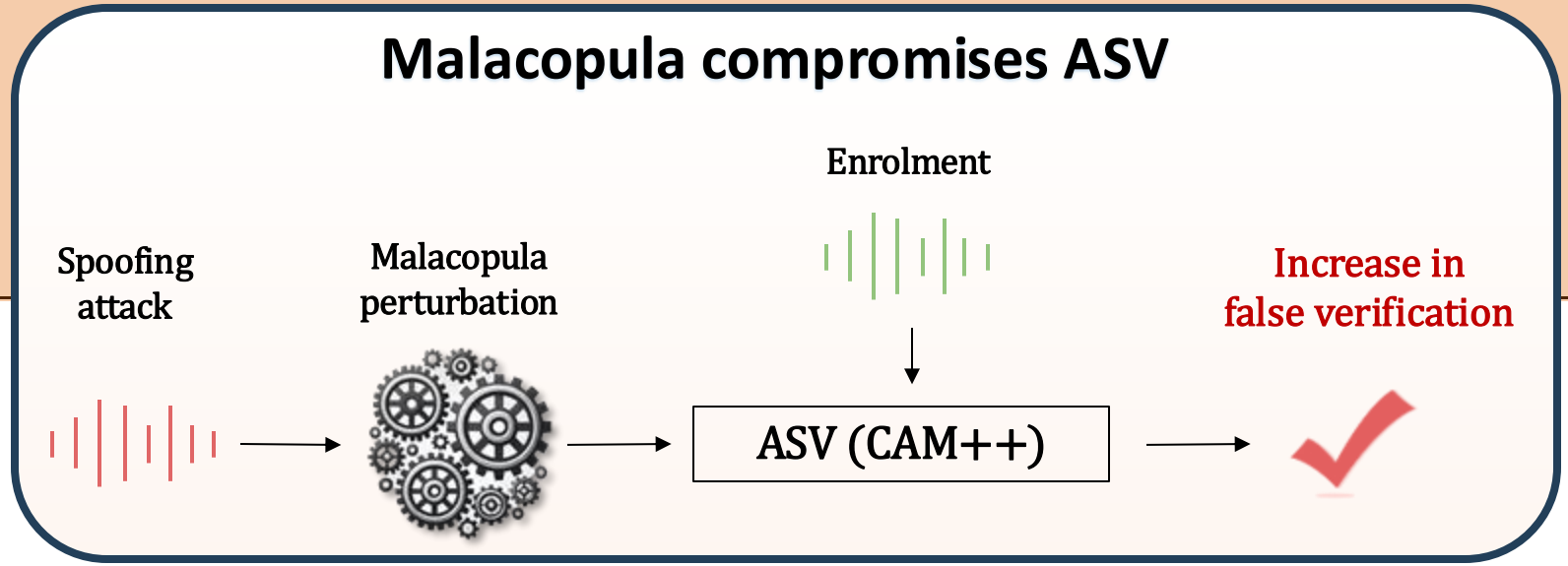
Attack	Input	Input processor	Acoustic model	Speaker embedding	Acoustic feature	Waveform model	Post-processing
A21	text	NLP	FS-based				
A22	text	NLP	FS+Prosod				
A28	text	DNN-encoder					
A26	speech	ASR	DNN+F0 est				l genuine noise
A17	text		Transformer-				
A19		NLP	MaryTTS				
A24	speech	PPG	DNN				
A25	speech	DNN-encoder	DiffVC				
A29	tex					HiFiGAN	-
A23							Malafide
A20							Malafide
A18							Malafide
							Malacopula
							Malacopula
							Malacopula
							Malafide+Malacopula



ASVspoof 5 dataset: spoofed data (eval. set)

Attack	Input					m model	Post-processing
A21	text					N	-
A22	text					N	-
A28	text						-
A26	speech						original genuine noise
A17	text						-
A19	text					c	-
A24	speech						-
A25	speech						-
A29	text					N	-
A23	A09	-	-	-	-	-	Malafide
A20	A12						Malafide
A18	A17						Malafide
A27	A26						Malacopula
A31	A22						Malacopula
A32	A25						Malacopula
A30	A18						Malafide+Malacopula





ASVspoof 5 dataset: spoofed data (eval. set)

Attack	Input	Input processor	Acoustic model	Speaker embedding	Acoustic feature	Waveform model	Post-processing
A21	text	NLP	FS-based	GST	log-spec	BigVGAN	-
A22	text	NLP	FS+ProsodyTransfer	GST	log-spec	BigVGAN	-
A28	text	DNN-encoder	YourTTS(pre.)	H/ASP	latent	HifiGAN	-
A26	speech	ASR	DNN+F0 est.	CAM++	Mel-spec	HifiGAN	original genuine noise
A17	text	BERT-based	Transformer-based	x-vector	DNN-latent	HifiGAN	-
A19	text	NLP	MaryTTS	-	-	unit-selec	-
A24	speech	PPG	DNN	x-vector	LSP	HifiGAN	-
A25	speech	DNN-encoder	DiffVC	latent	Mel-spec	HifiGAN	-
A29	text	DNN-encoder	XTTS(pre.)	ECAPA2	latent	HifiGAN	-
A23	A09	-	-	-	-	-	Malafide
A20	A12	-	-	-	-	-	Malafide
A18	A17	-	-	-	-	-	Malafide
A27	A26	-	-	-	-	-	Malacopula
A31	A22	-	-	-	-	-	Malacopula
A32	A25	-	-	-	-	-	Malacopula
A30	A18	-	-	-	-	-	Malafide+Malacopula

Bona fide A17 A18 A30

ASVspoof 5 dataset: spoofed data (eval. set)

Attack	Input	Input processor	Acoustic model	Speaker embedding	Acoustic feature	Waveform model	Post-processing
A21	text	NLP	FS-based	GST	log-spec	BigVGAN	-
A22	text	NLP	FS+ProsodyTransfer	GST	log-spec	BigVGAN	-
A28	text	DNN-encoder	YourTTS(pre.)	H/ASP	latent	HifiGAN	-
A26	speech	ASR	DNN+F0 est.	CAM++	Mel-spec	HifiGAN	original genuine noise
A17	text	BERT-based	Transformer-based	x-vector	DNN-latent	HifiGAN	-
A19	text	NLP	MaryTTS	-	-	unit-selec	-
A24	speech	PPG	DNN	x-vector	LSP	HifiGAN	-
A25	speech	DNN-encoder	DiffVC	latent	Mel-spec	HifiGAN	-
A29	text	DNN-encoder	XTTS(pre.)	ECAPA2	latent	HifiGAN	-
A23	A09	-	-	-	-	-	Malafide
A20	A12	-	-	-	-	-	Malafide
A18	A17	-	-	-	-	-	Malafide
A27	A26	-	-	-	-	-	Malacopula
A31	A22	-	-	-	-	-	Malacopula
A32	A25	-	-	-	-	-	Malacopula
A30	A18	-	-	-	-	-	Malafide+Malacopula

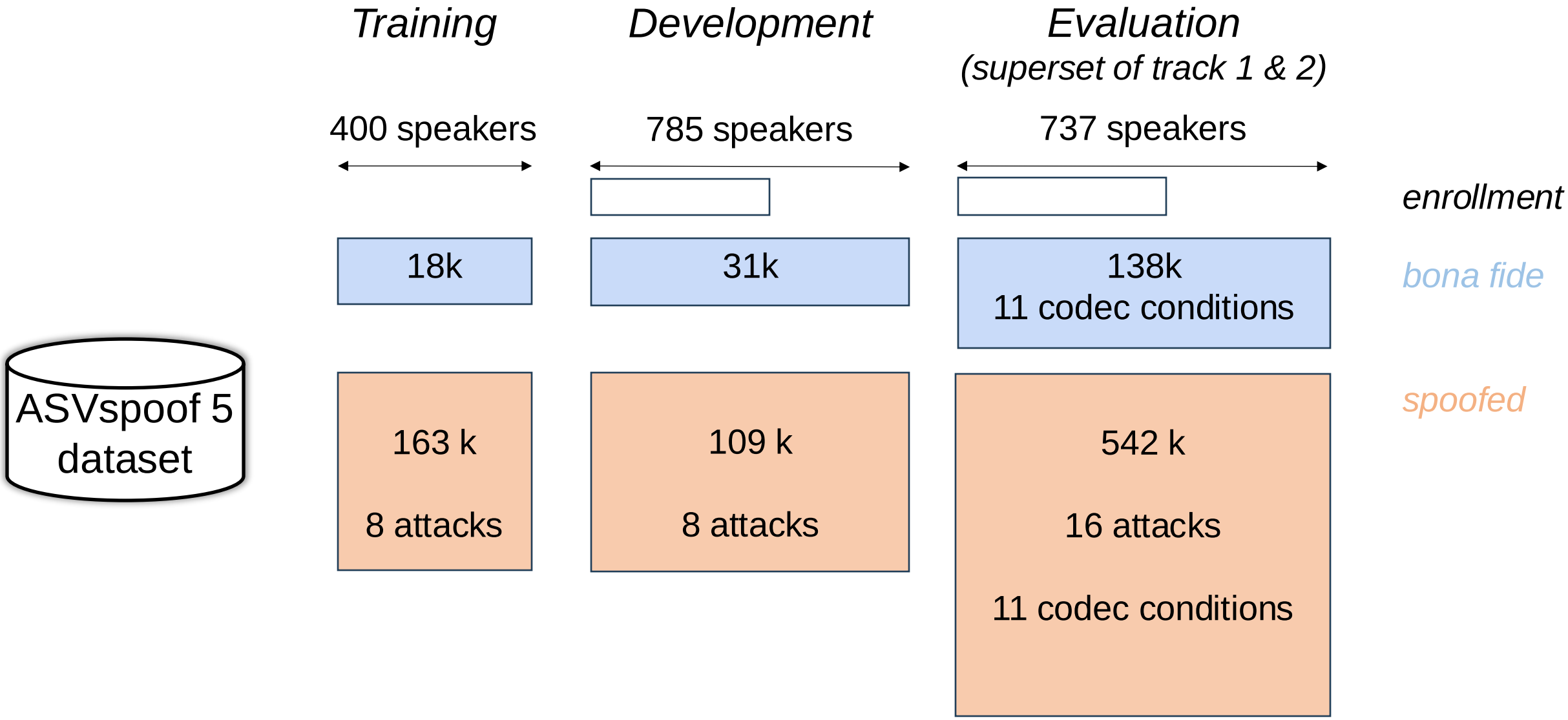
Bona fide A21 A22 A28 A26

ASVspoof 5 dataset: spoofed data (eval. set)

Attack	Input	Input processor	Acoustic model	Speaker embedding	Acoustic feature	Waveform model	Post-processing
A21	text	NLP	FS-based	GST	log-spec	BigVGAN	-
A22	text	NLP	FS+ProsodyTransfer	GST	log-spec	BigVGAN	-
A28	text	DNN-encoder	YourTTS(pre.)	H/ASP	latent	HifiGAN	-
A26	speech	ASR	DNN+F0 est.	CAM++	Mel-spec	HifiGAN	original genuine noise
A17	text	BERT-based	Transformer-based	x-vector	DNN-latent	HifiGAN	-
A19	text	NLP	MaryTTS	-	-	unit-selec	-
A24	speech	PPG	DNN	x-vector	LSP	HifiGAN	-
A25	speech	DNN-encoder	DiffVC	latent	Mel-spec	HifiGAN	-
A29	text	DNN-encoder	XTTS(pre.)	ECAPA2	latent	HifiGAN	-
A23	A09	-	-	-	-	-	Malafide
A20	A12	-	-	-	-	-	Malafide
A18	A17	-	-	-	-	-	Malafide
A27	A26	-	-	-	-	-	Malacopula
A31	A22	-	-	-	-	-	Malacopula
A32	A25	-	-	-	-	-	Malacopula
A30	A18	-	-	-	-	-	Malafide+Malacopula

Bona fide A19 A24 A25 A29

ASVspoof 5 dataset: spoofed data (eval. set)



ASVspoof 5 dataset: codec (eval. set)

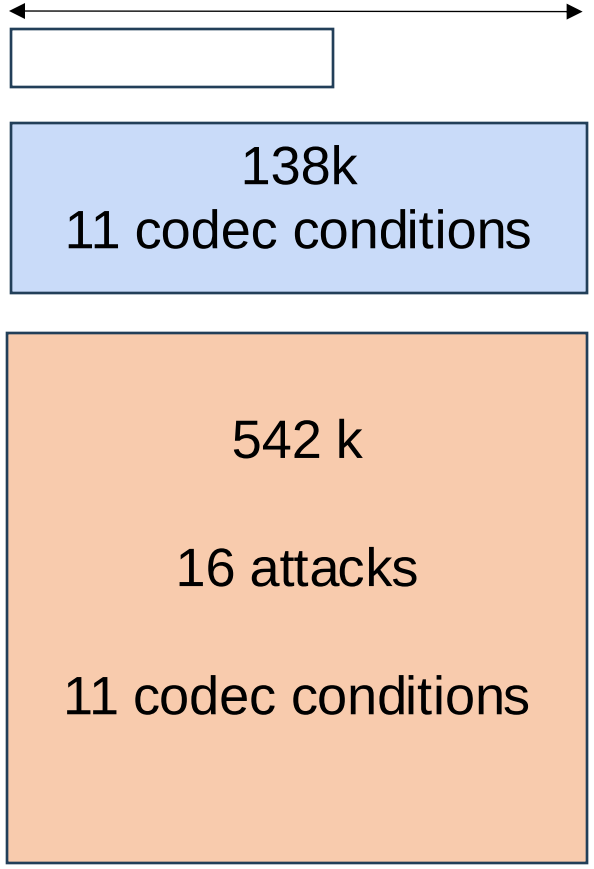
	Codec	Bandwidth	Bitrate range
C00	-	16 kHz	-
C01	opus	16 kHz	6.0 - 30.0
C02	amr	16 kHz	6.6 - 23.05
C03	speex	16 kHz	5.75 - 34.20
C04	Encodec	16 kHz	1.5 - 24.0
C05	mp3	16 kHz	45 - 256
C06	m4a	16 kHz	16 - 128
C07	mp3+Encodec	16 kHz	varied
C08	opus	8 kHz	4.0 - 20.0
C09	arm	8 kHz	4.75 - 12.20
C10	speex	8 kHz	3.95 - 24.60
C11	varied	8 kHz	varied

DNN codec

simulation of real applications (appendix)

Evaluation
(superset of track 1 & 2)

737 speakers

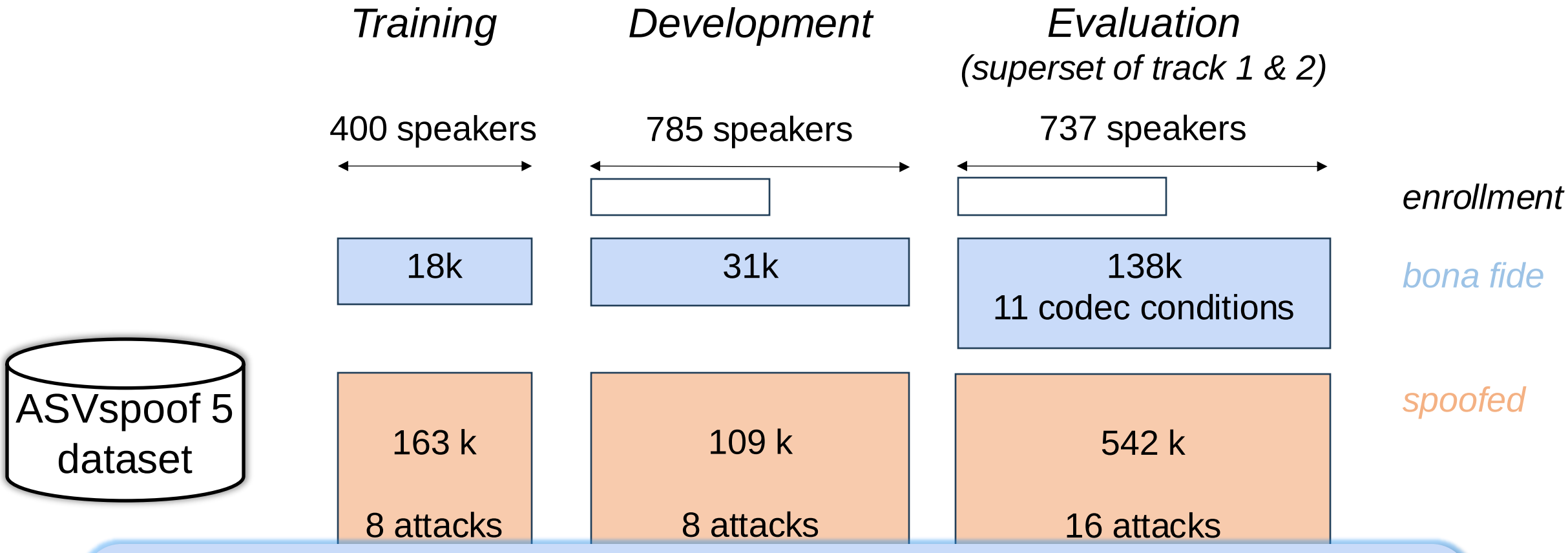


enrollment

bona fide

spoofed

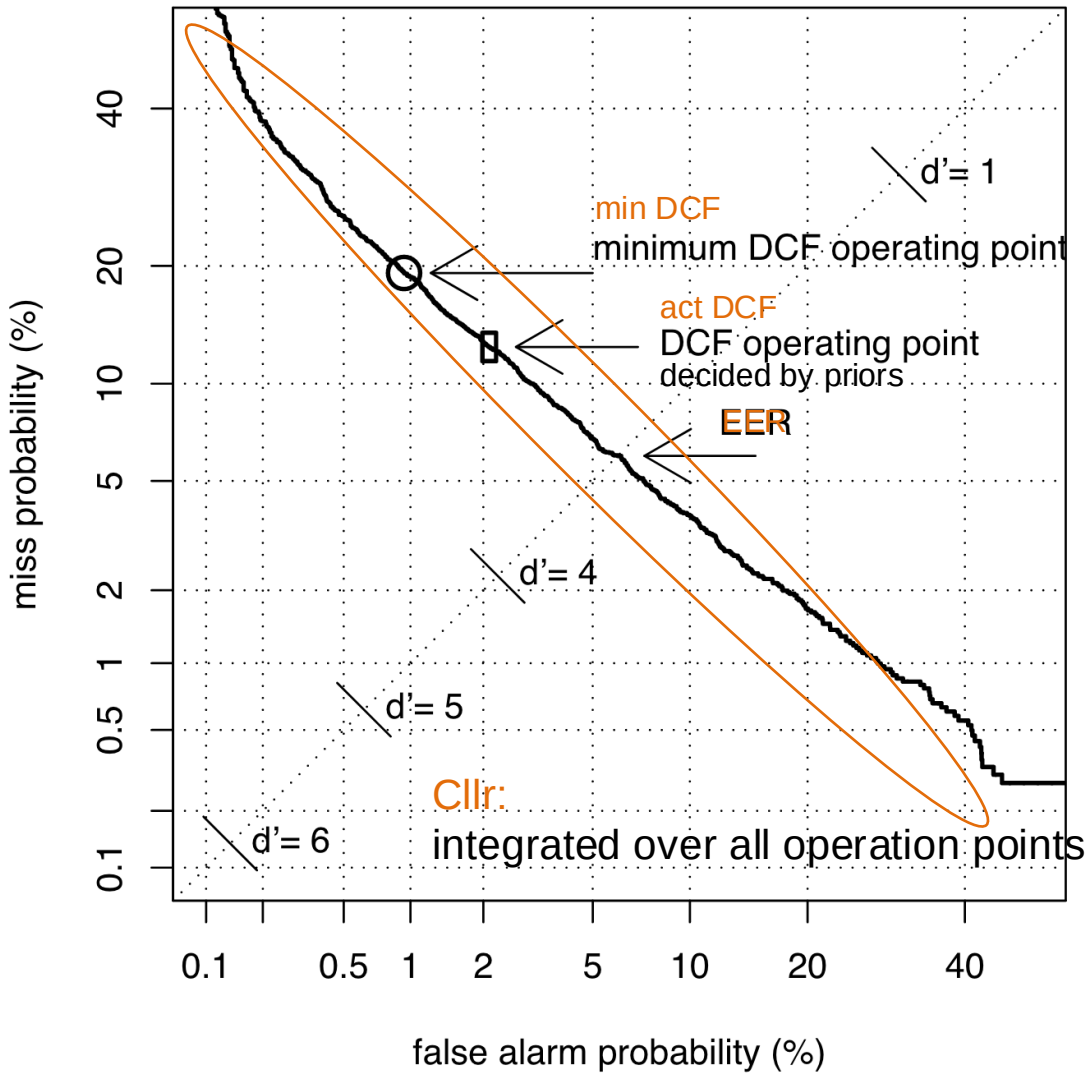
ASVspoof 5 dataset



Contributed by more than 10 research groups
(see acknowledgement)
A paper on the dataset is under planning

Evaluation metrics

Track 1 evaluation metrics



Countermeasure → CM scores

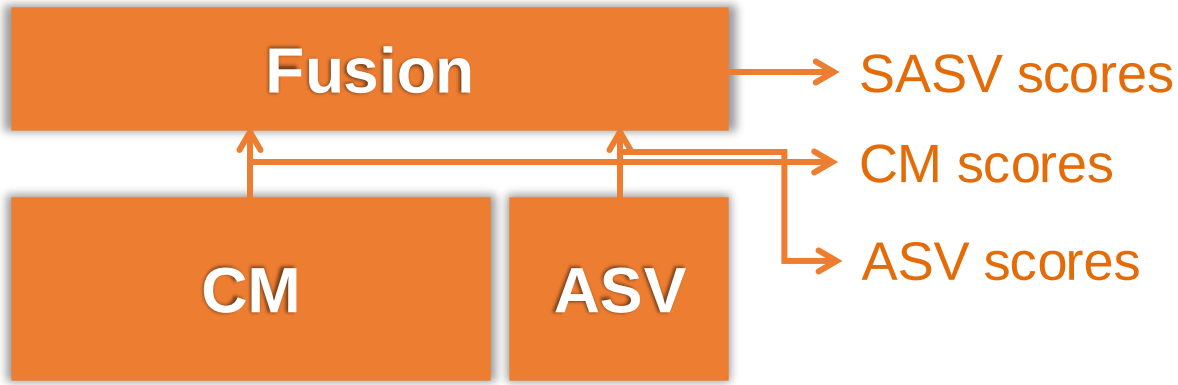
	explicit detection cost	calibration aware?
min DCF	✓	*x*
act DCF	✓	✓
Cllr	*x*	✓
EER	*x*	*x*

Figure adopted from D. A. Van Leeuwen and N. Brümmer, "An introduction to application-independent evaluation of speaker recognition systems," in Speaker classification I, Springer, 2007, pp. 330–353.

Track 2 evaluation metrics

Type 1 solution

Type 2 solution



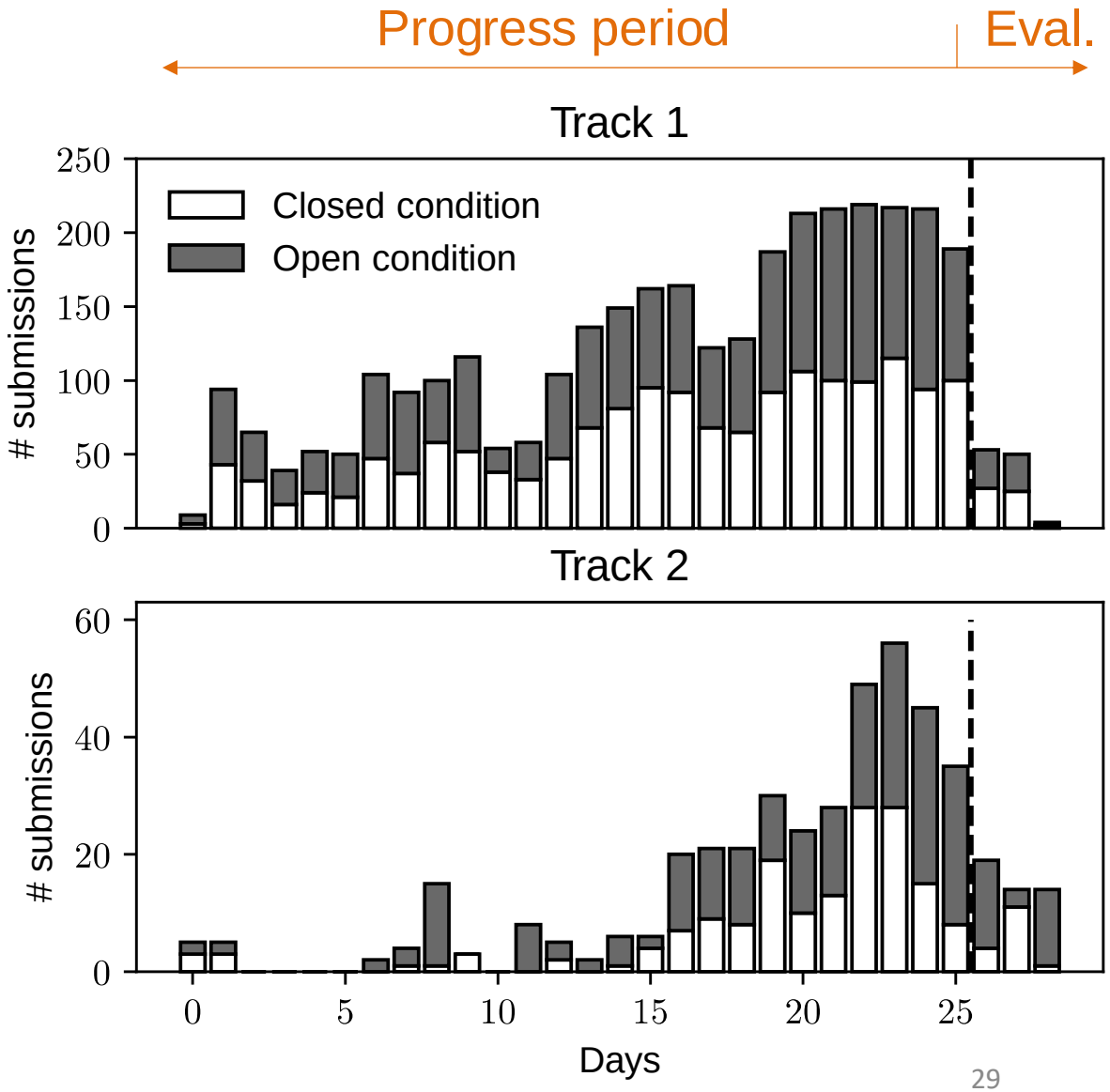
	explicit detection cost	scores required	applicable to type 1 solution?	applicable to type 2 solution?
a-DCF	✓	**SASV**	✓	✓
t-DCF	✓	ASV (by organizer), CM	*x*	✓
t-EER	*x*	ASV, CM	*x*	✓

H. Shim, et al, “a-DCF: an architecture agnostic metric with application to spoofing-robust speaker verification,” in Proc. Odyssey, 2024,
T. Kinnunen, et al, “t-EER: Parameter-Free Tandem Evaluation of Countermeasures and Biometric Comparators,” IEEE Trans. Pattern Anal. Mach. Intell., pp. 1–16, 2023
T. Kinnunen et al., “t-DCF: a detection cost function for the tandem assessment of spoofing countermeasures and automatic speaker verification,” in Proc. Odyssey, 2018

Evaluation platform & participation

Evaluation platform

- Codalab
- Progress period (06/12 – 07/21)
  - ~1 month
  - subset of evaluation data
  - 4 submissions per day
- Evaluation period (07/21 – 07/24)
  - 3 days
  - one submission only

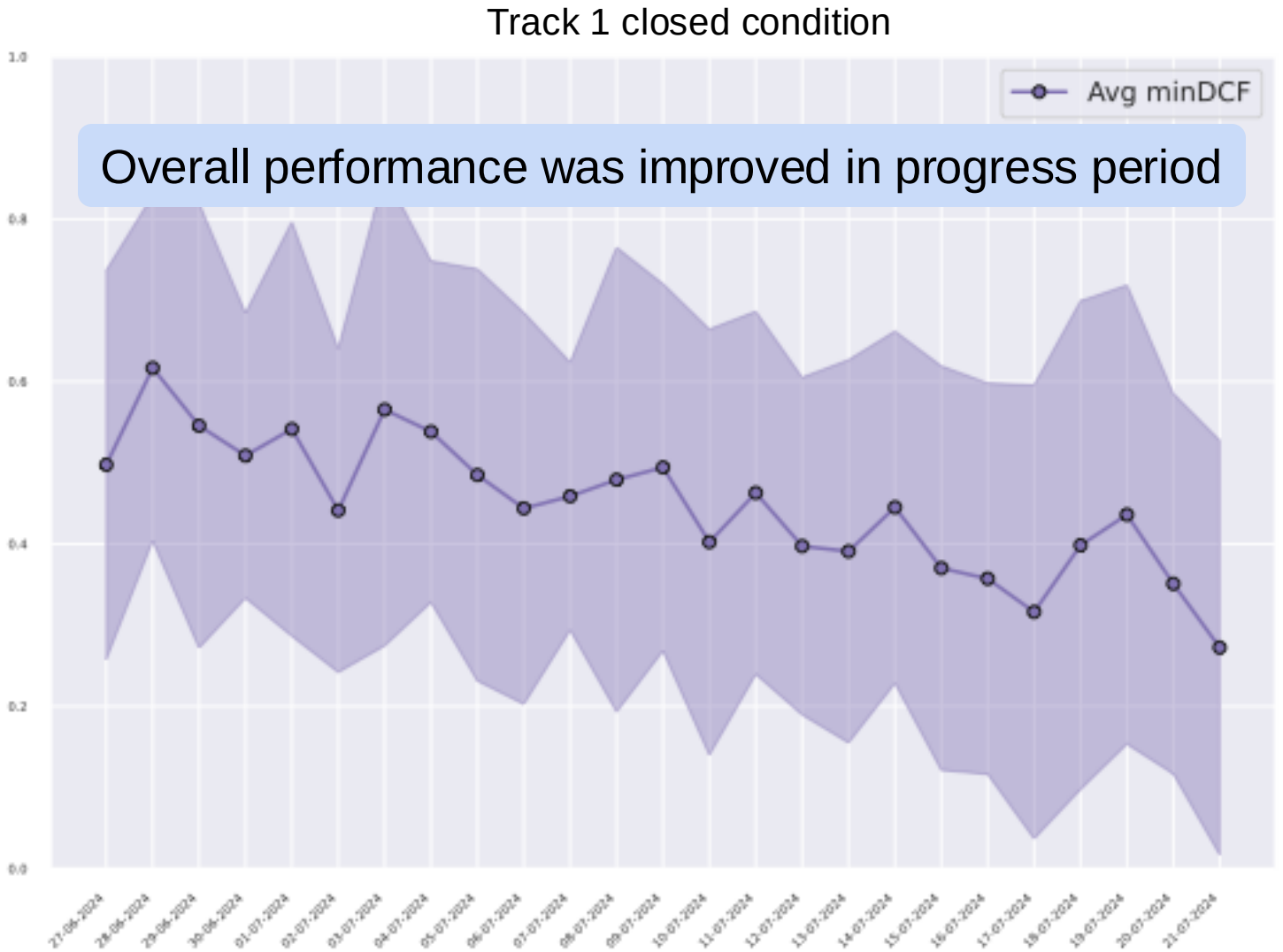


Progress phase

- Codalab

- Progress period
  - ~1 month
  - subset of evaluation data
  - 4 submissions per day

- Evaluation period
  - 3 days
  - one submission only

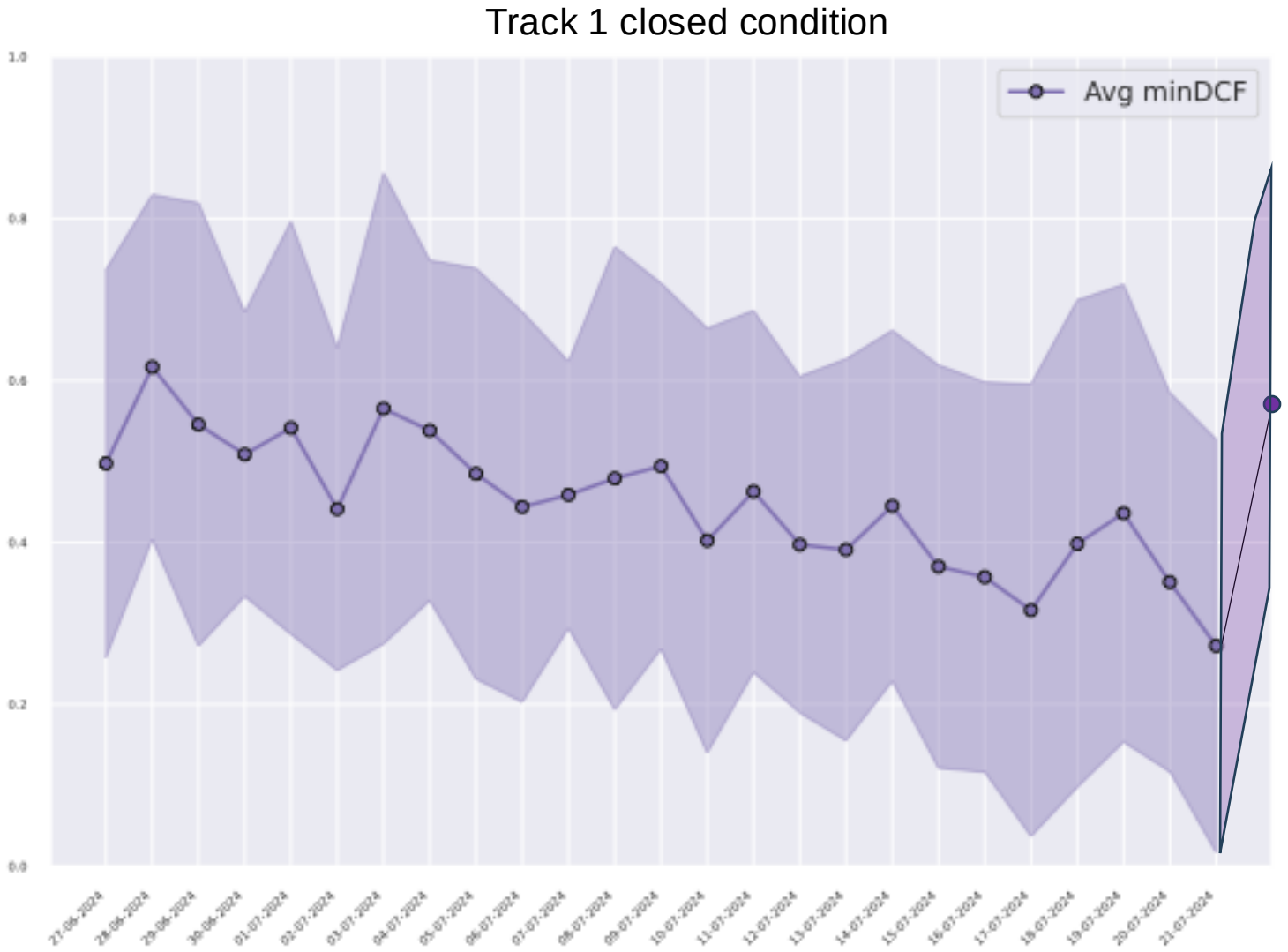


Evaluation phase

- Codalab

- Progress period
  - ~1 month
  - subset of evaluation data
  - 4 submissions per day

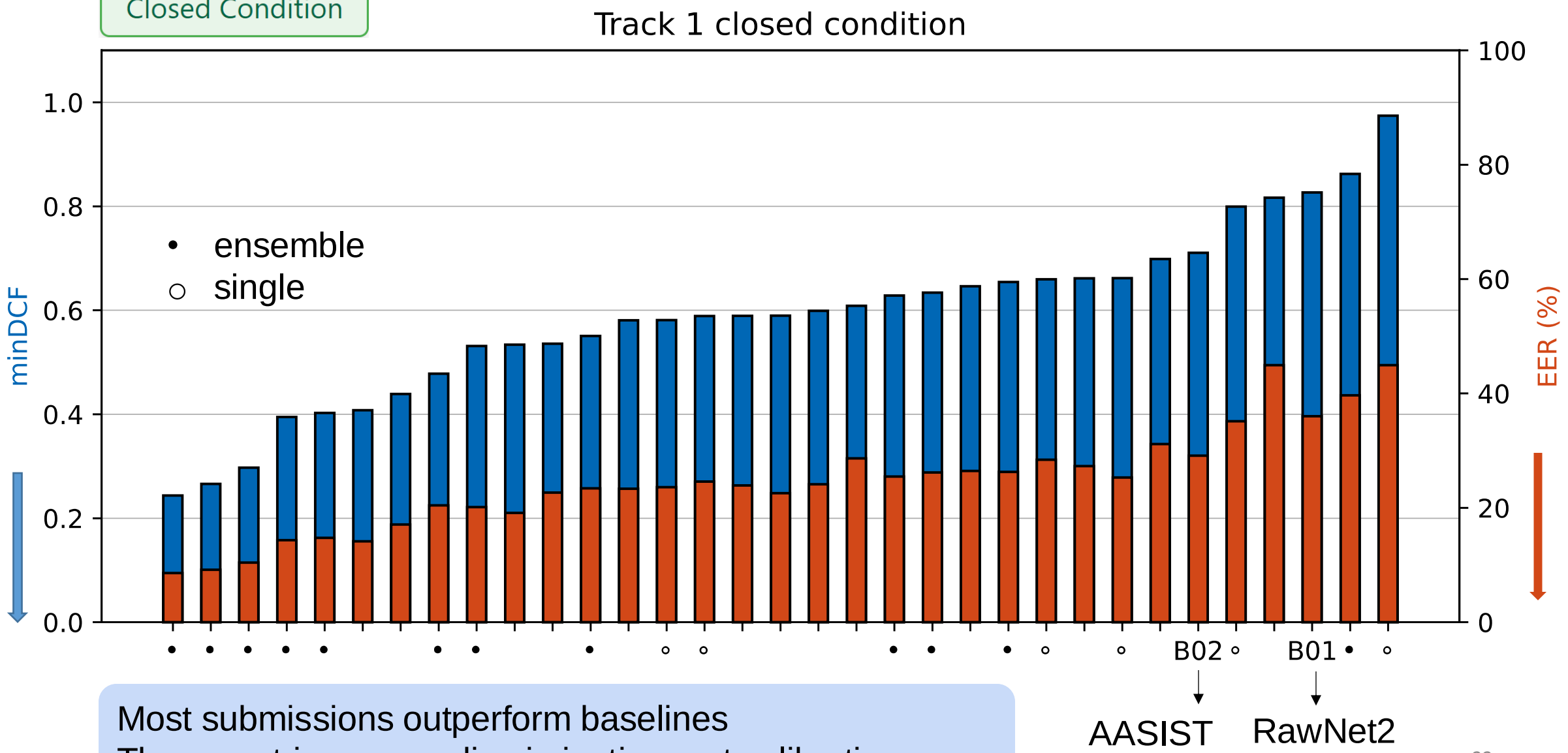
- Evaluation period
  - 3 days
  - one submission only



Overall results

Track 1 - overall results

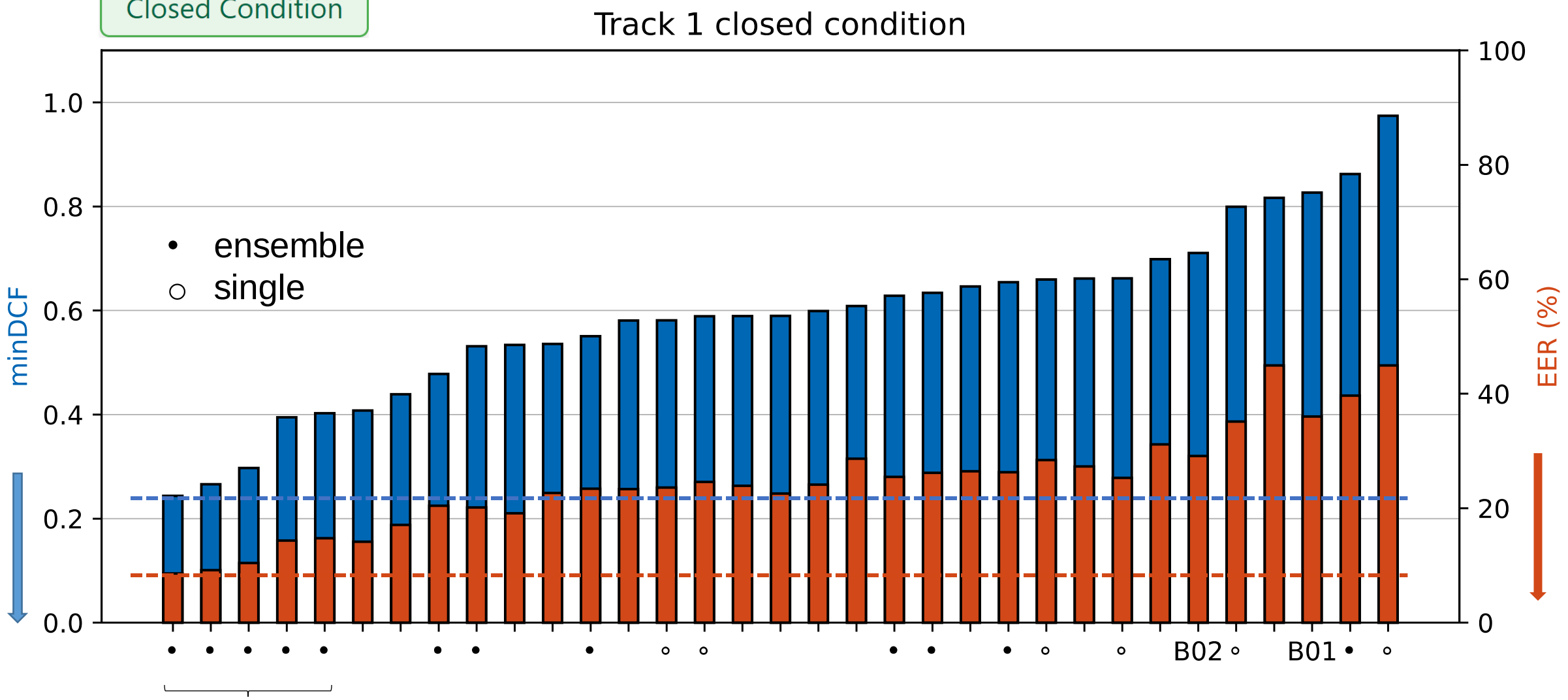
Closed Condition



Most submissions outperform baselines
These metrics gauge discrimination, not calibration

Track 1 - overall results

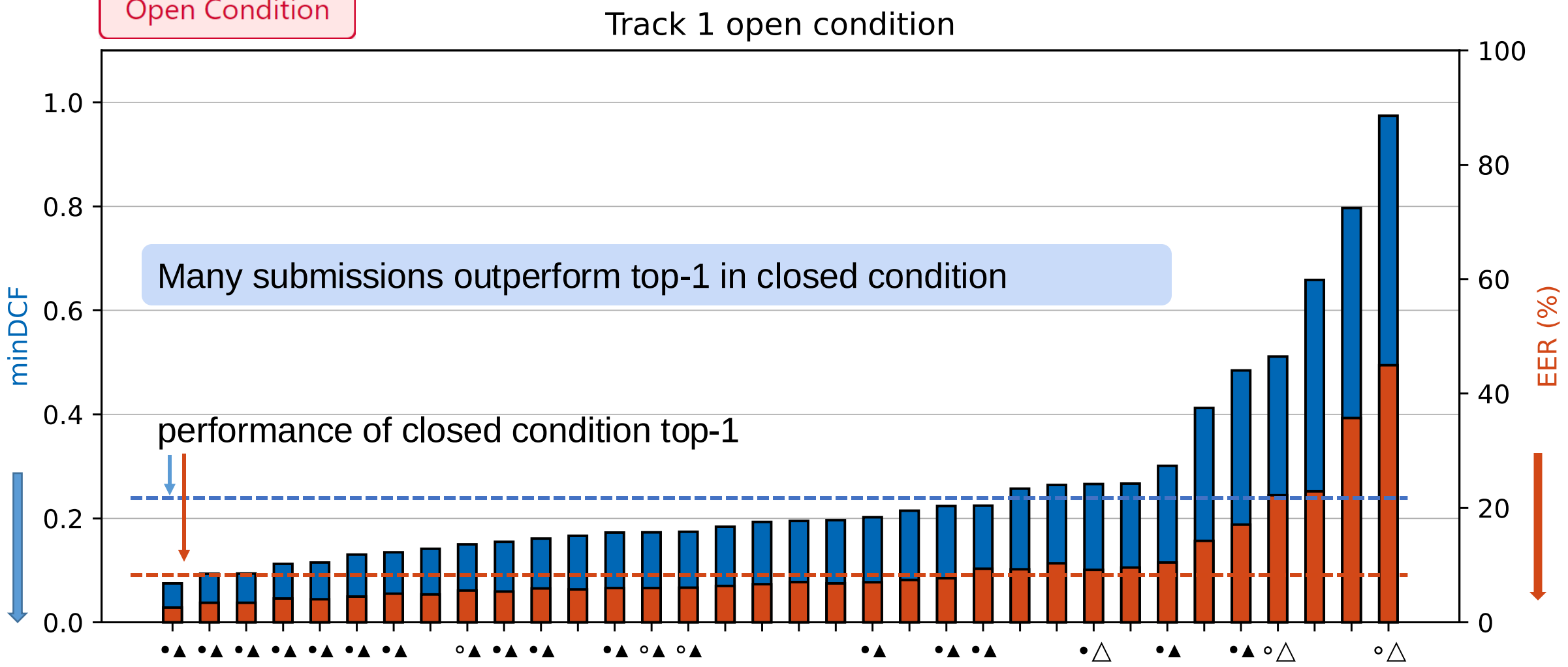
Closed Condition



Top systems are ensembles

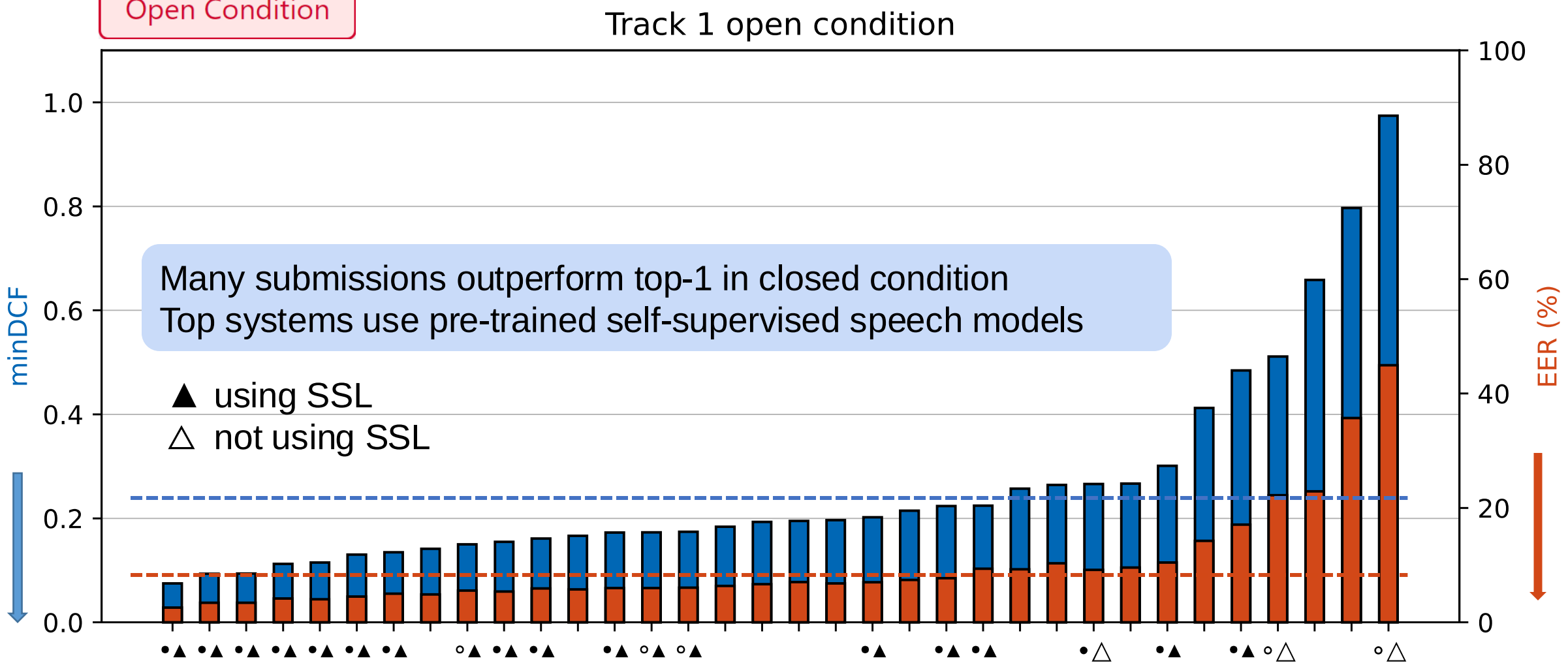
Track 1 - overall results

Open Condition



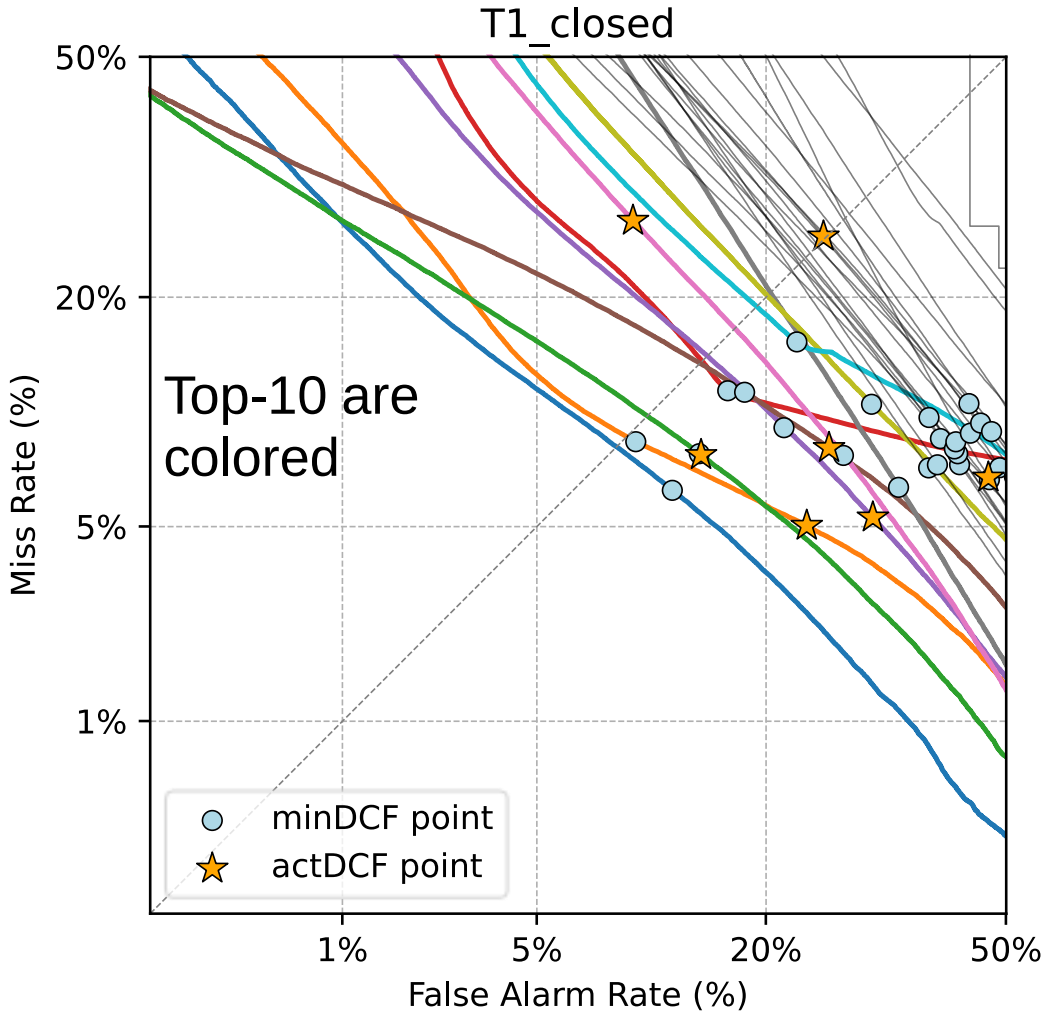
Track 1 - overall results

Open Condition

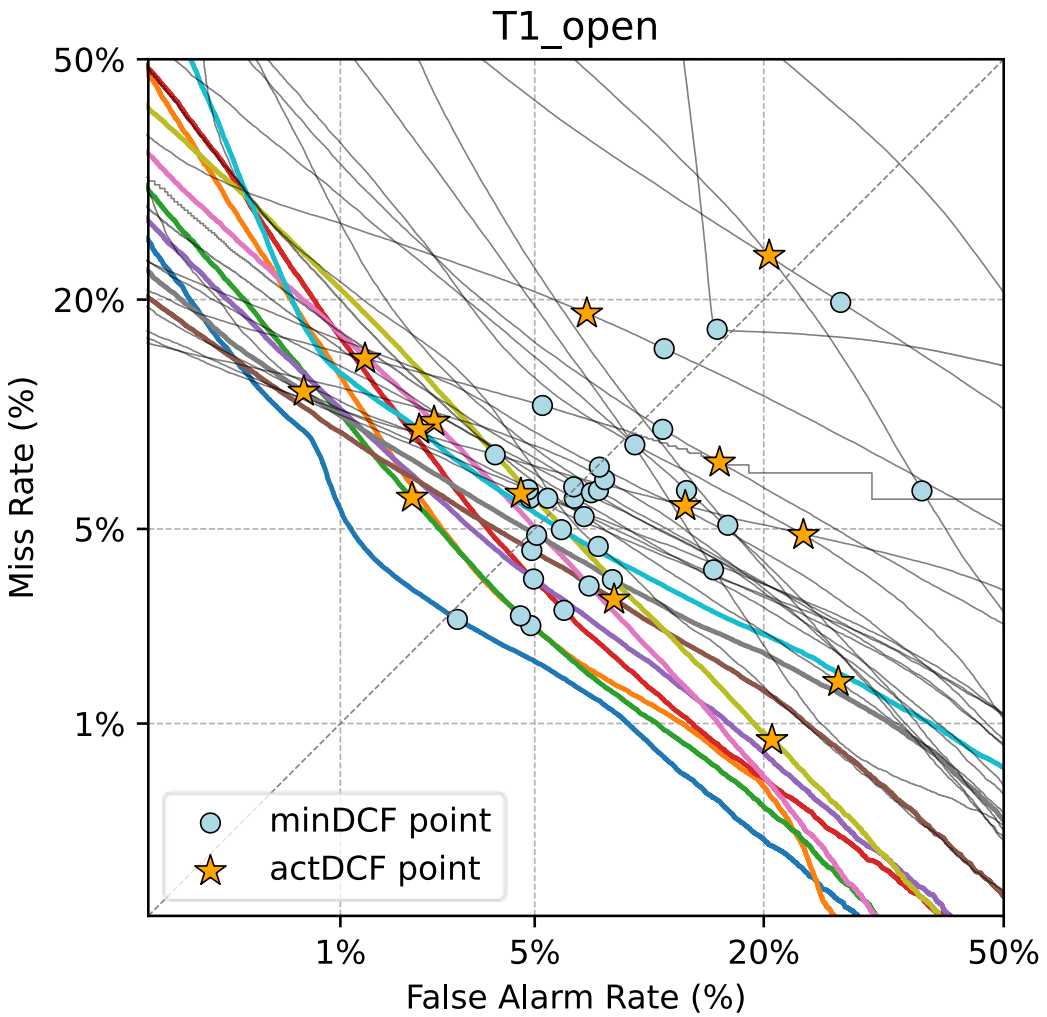


Track 1 - overall results

Closed Condition

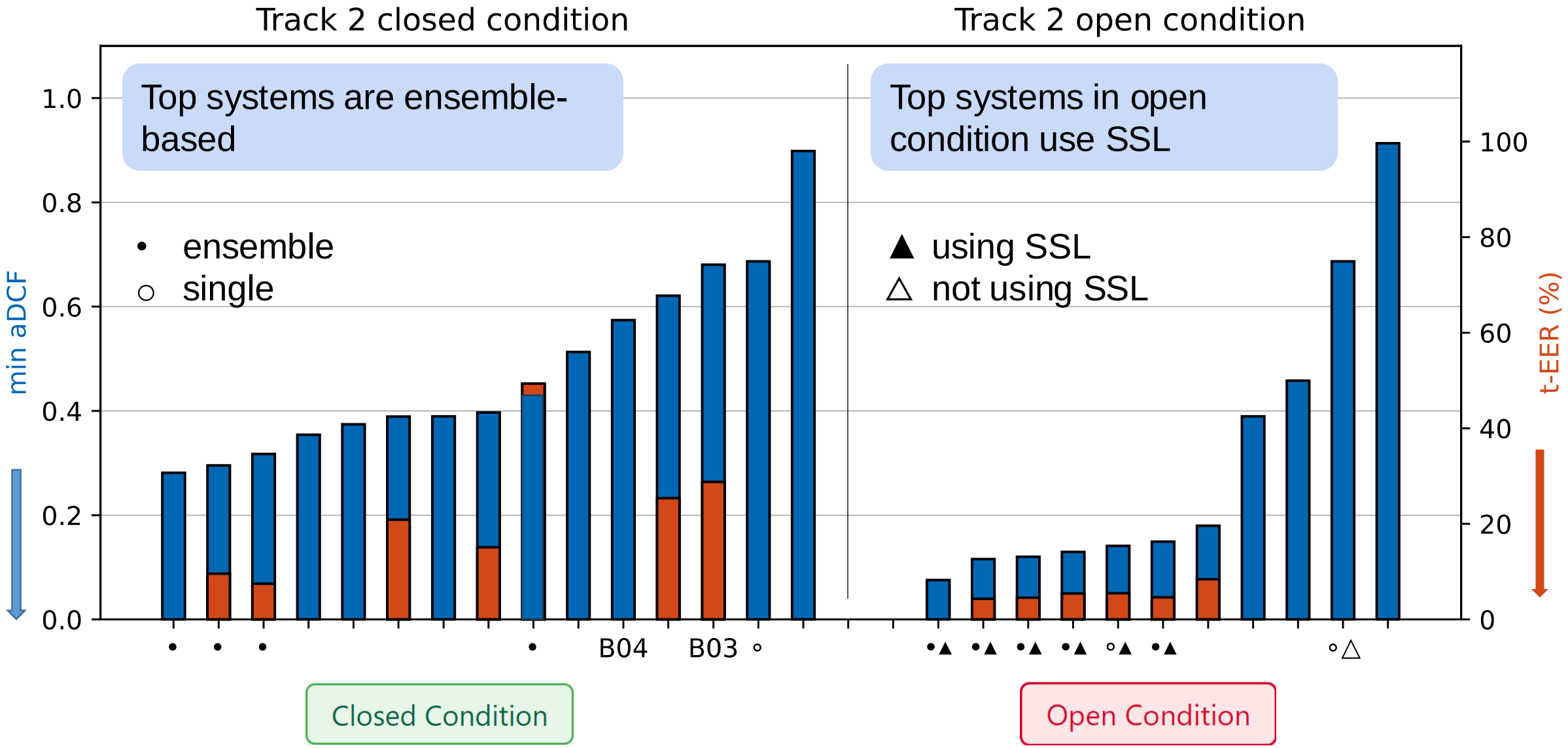


Open Condition

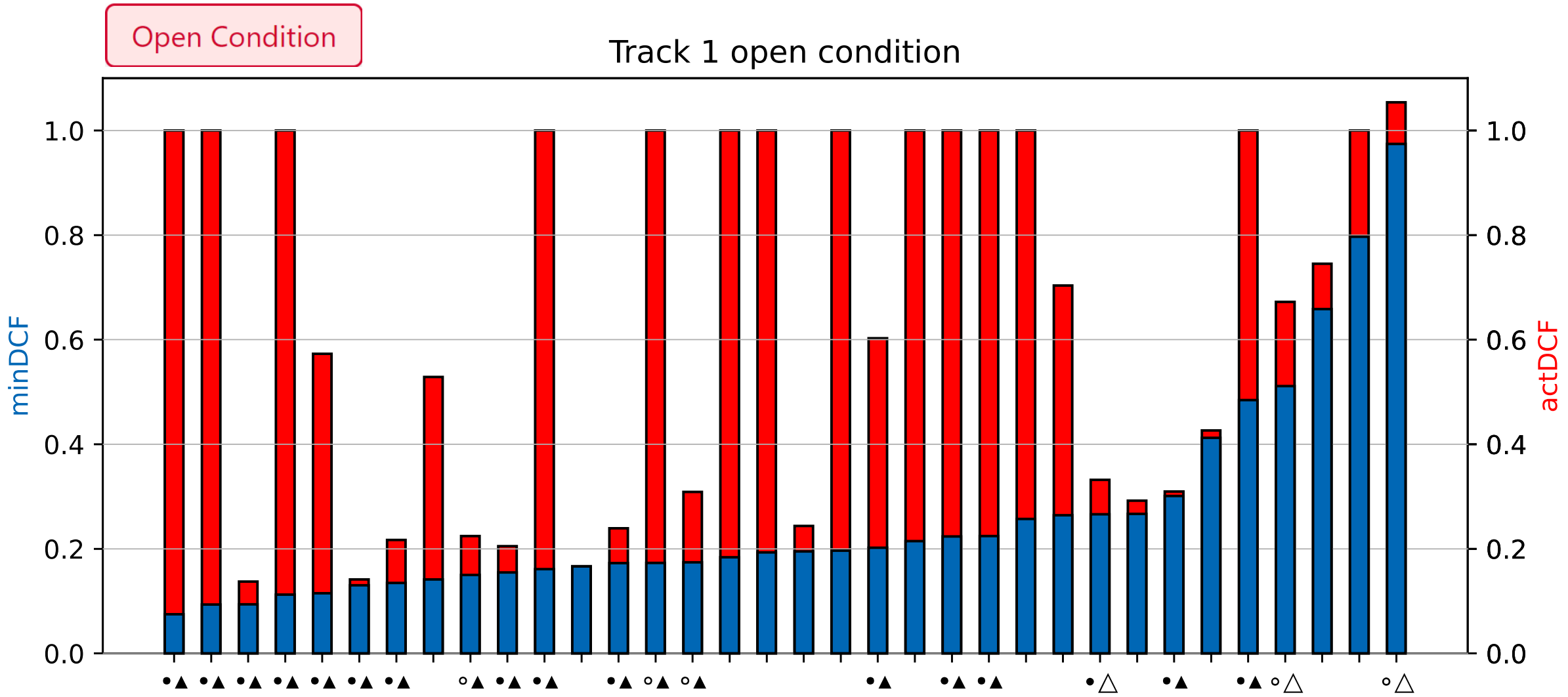


Some systems don't work properly at *actDCF* operation point – see analysis of calibration

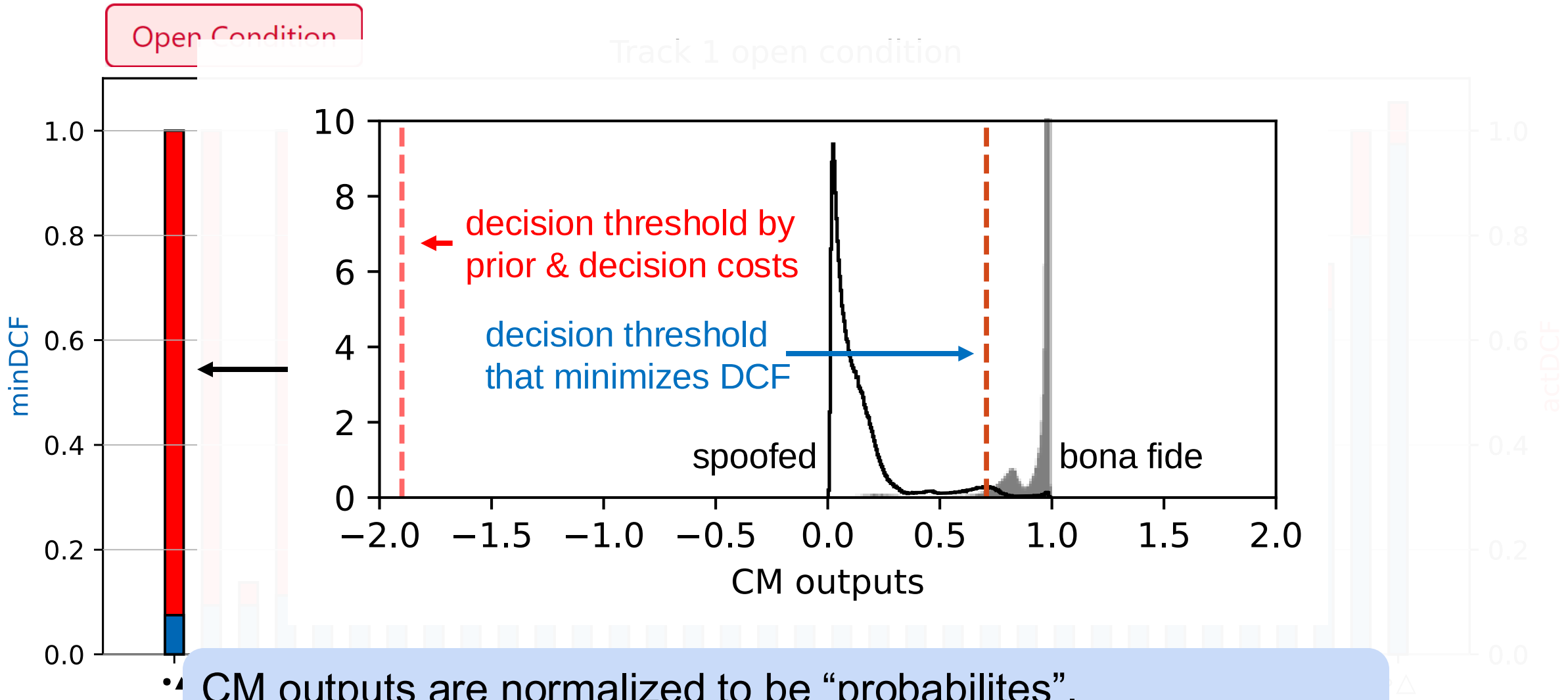
Track 2 - overall results



Analysis – score calibration

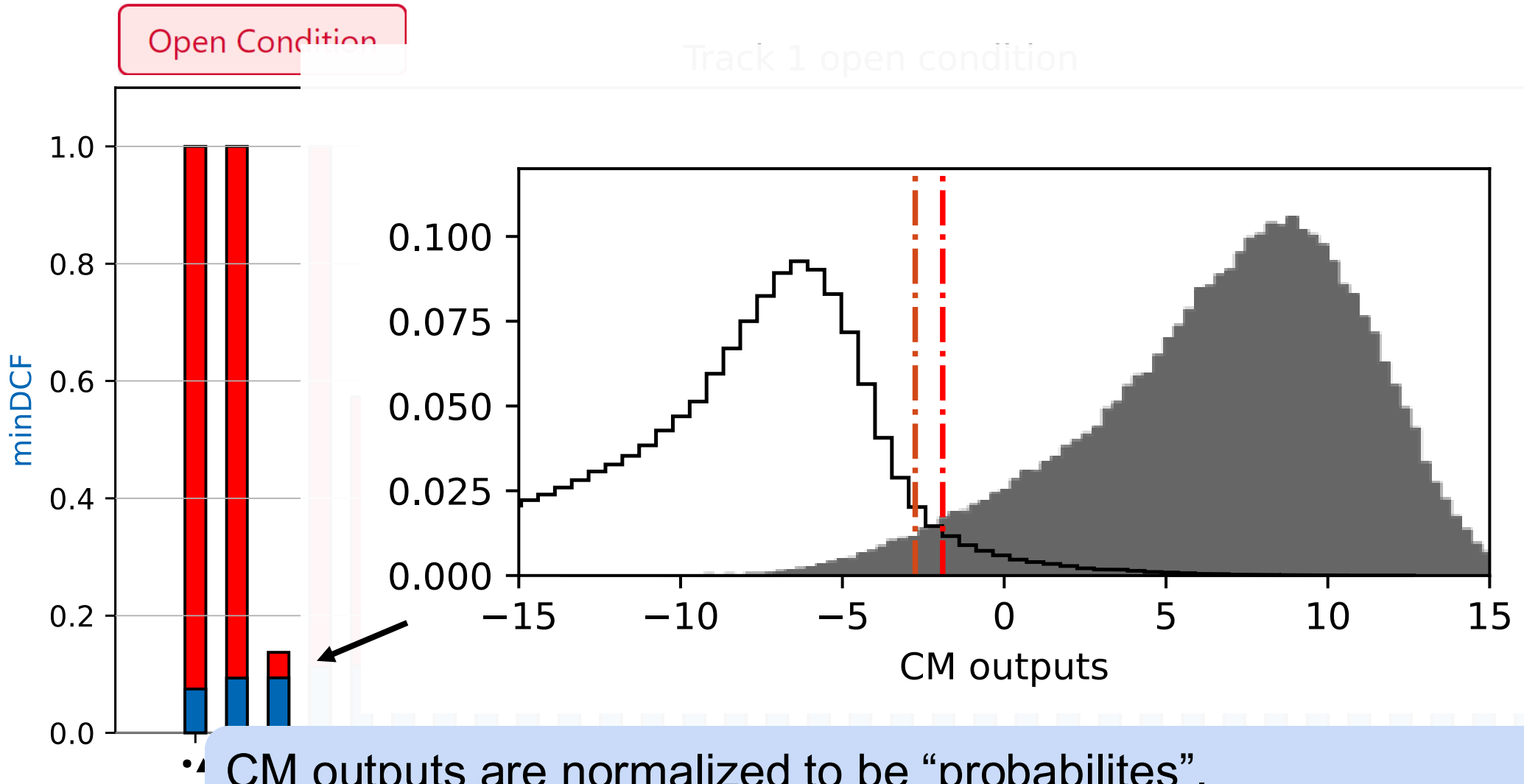


Analysis – score calibration



CM outputs are normalized to be “probabilites”.
They cannot be interpreted as LLRs for Bayesian decision.

Analysis – score calibration

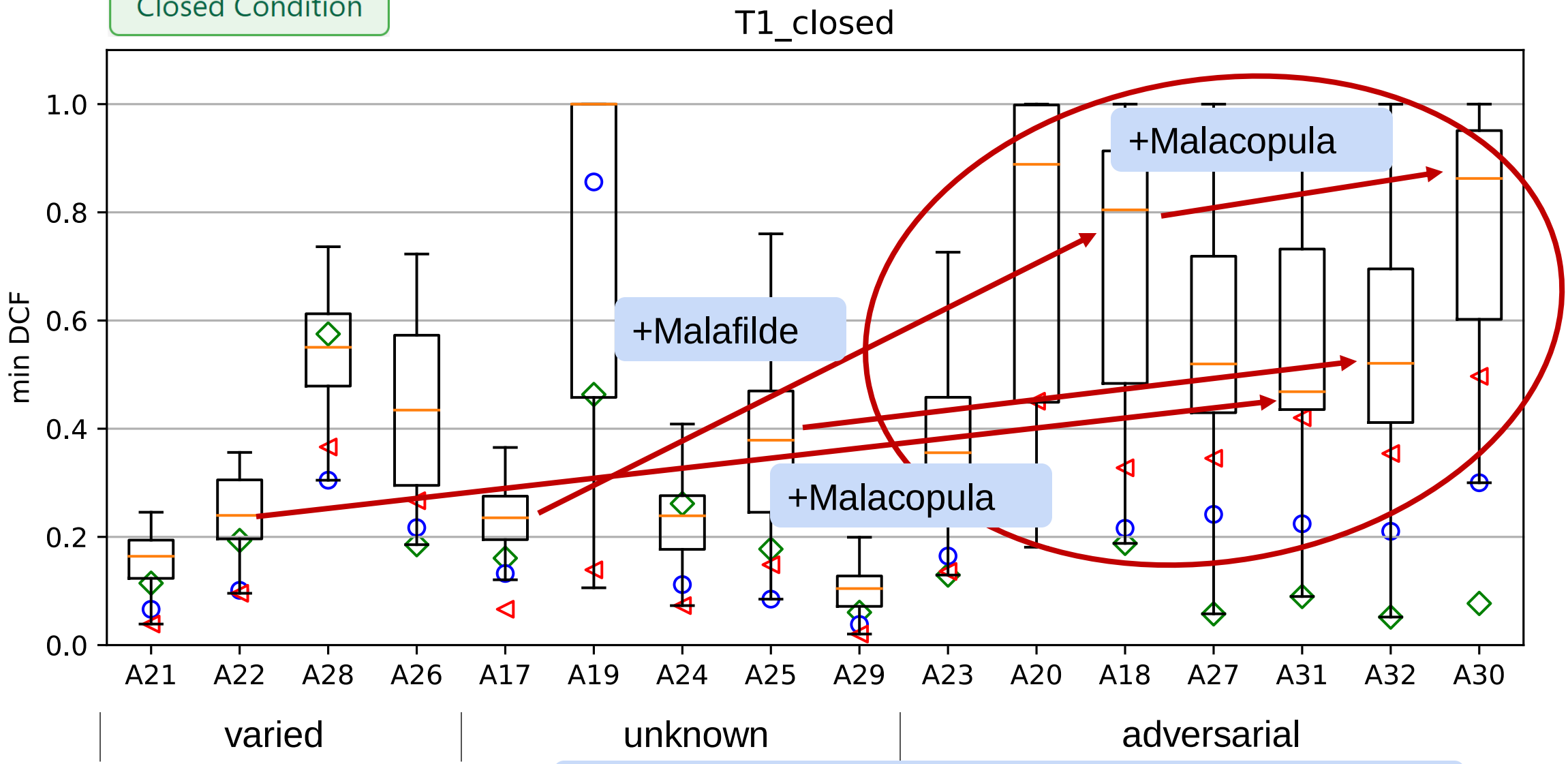


CM outputs are normalized to be “probabilites”.
They cannot be interpreted as LLRs for Bayesian decision.

Analysis & additional results

Analysis – attack

Closed Condition



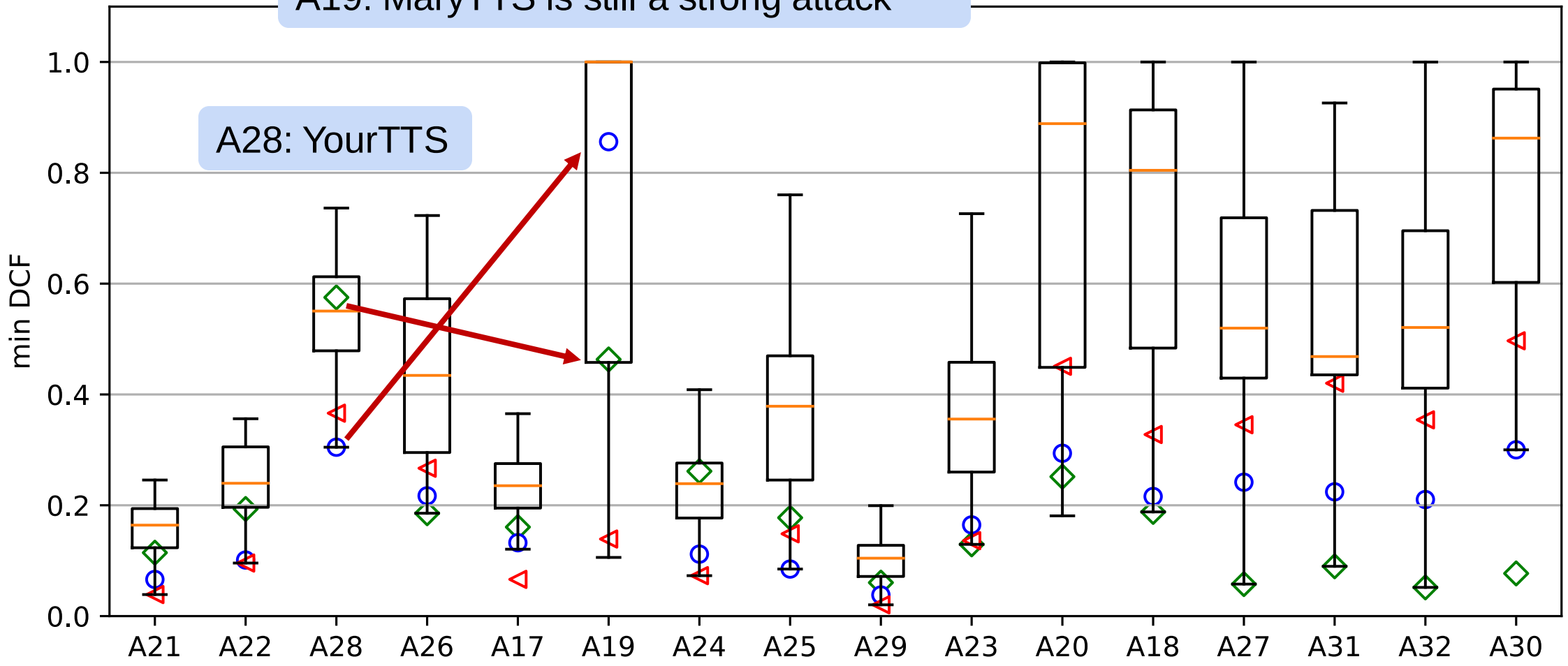
Adversarial attacks are more challenging to detect

Analysis – attack

Closed Condition

A19: MaryTTS is still a strong attack

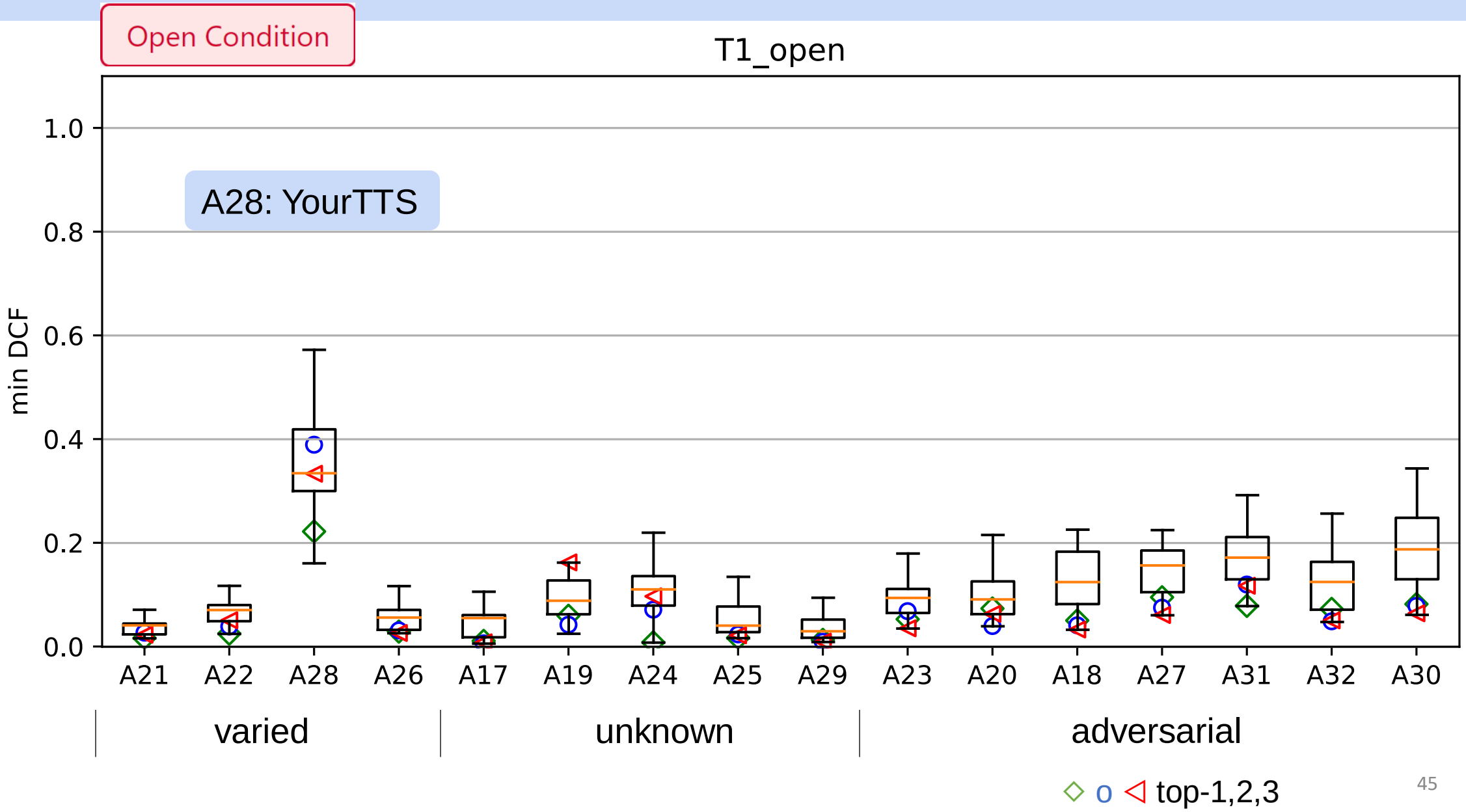
A28: YourTTS



CM overfits to different types of attacks?

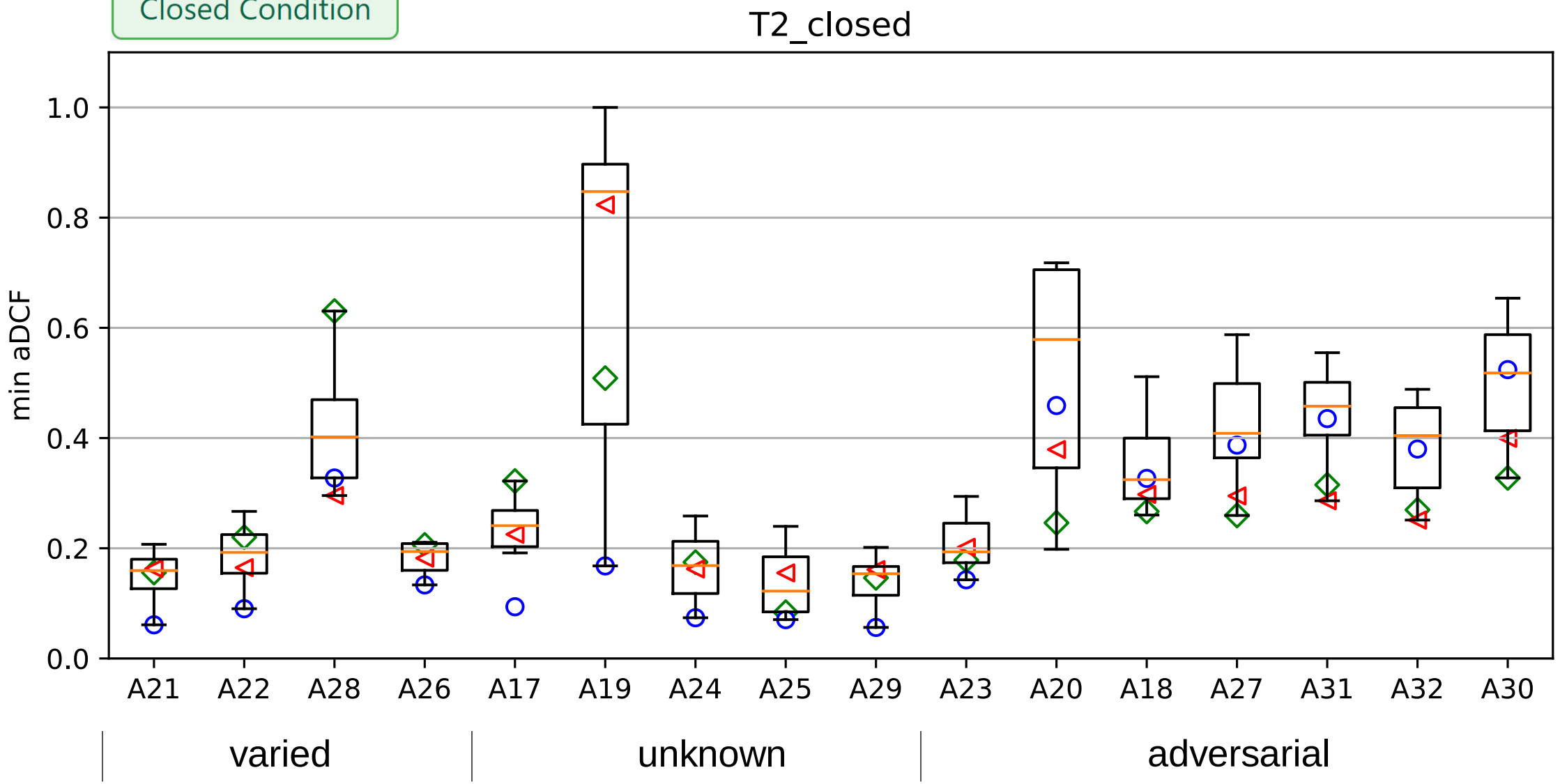
◇ o ◁ top-1,2,3

Analysis – attack



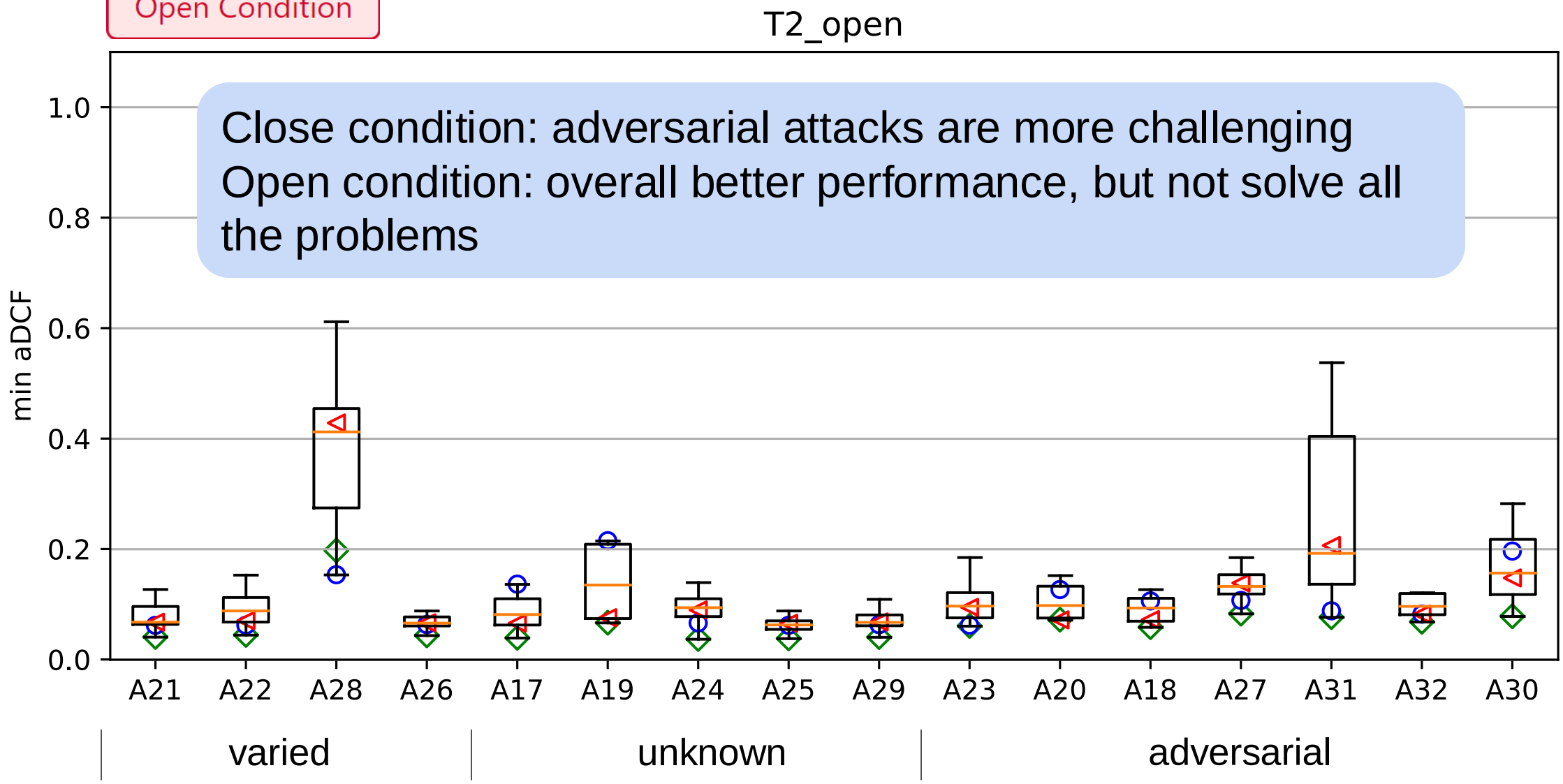
Analysis – attack

Closed Condition



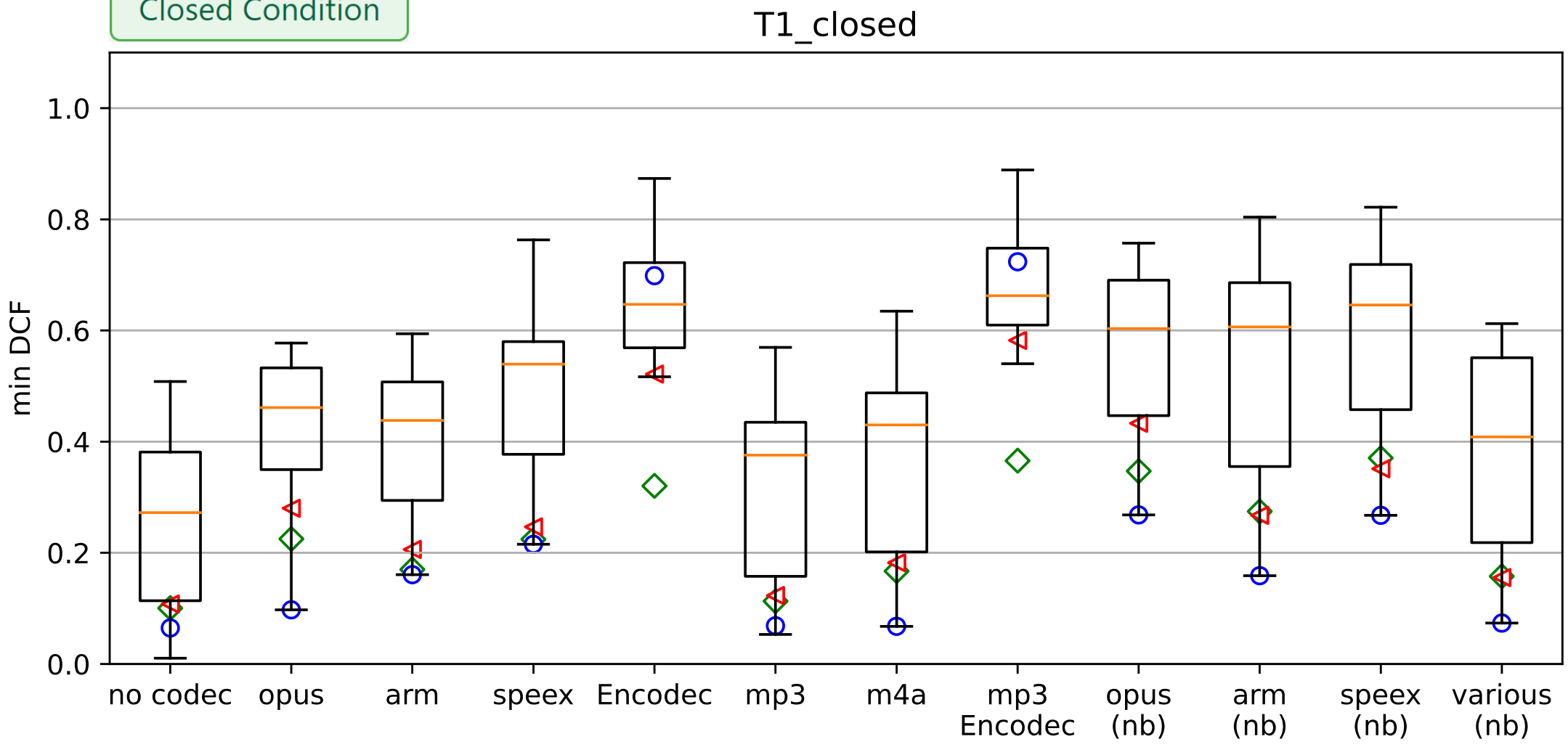
Analysis – attack

Open Condition



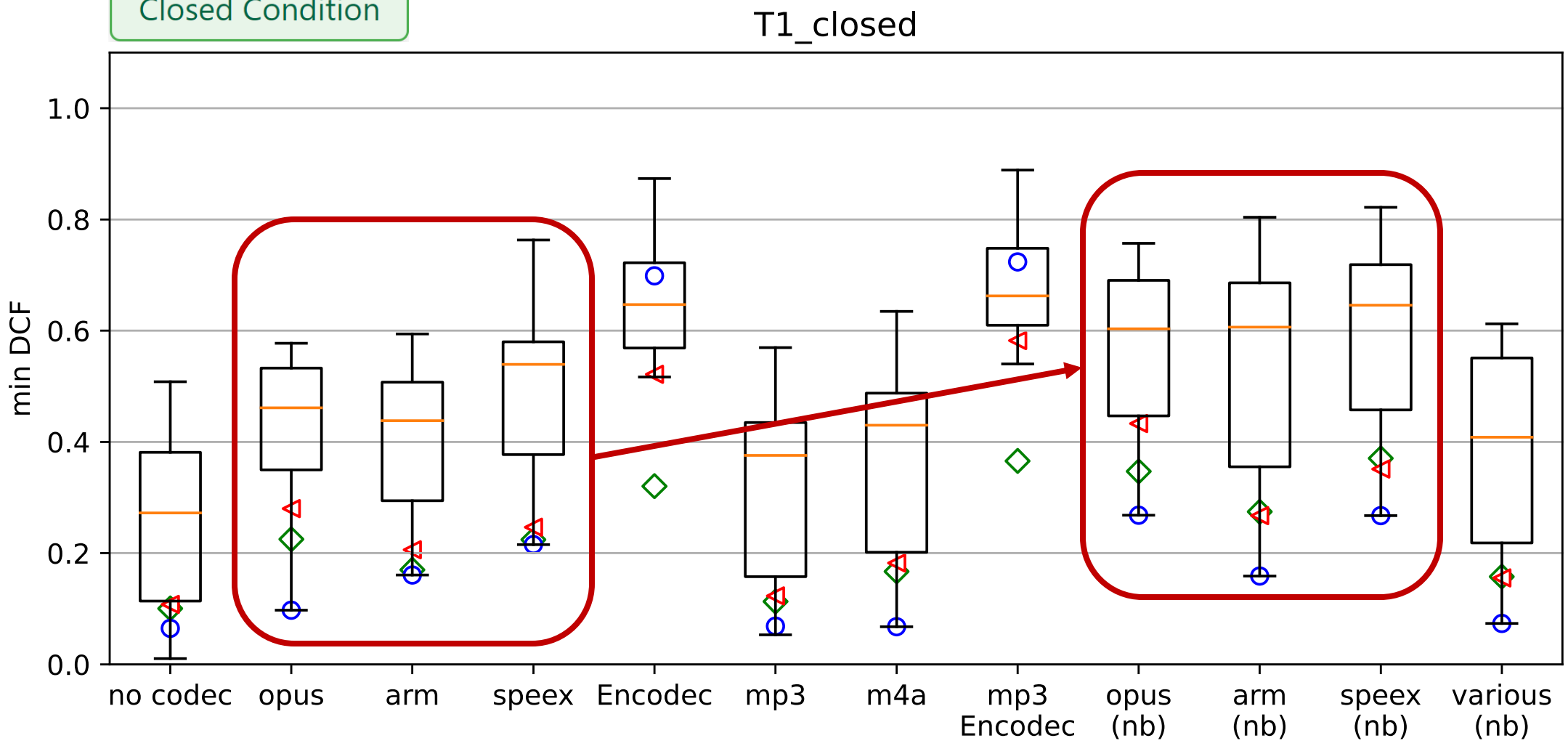
Analysis – codec

Closed Condition



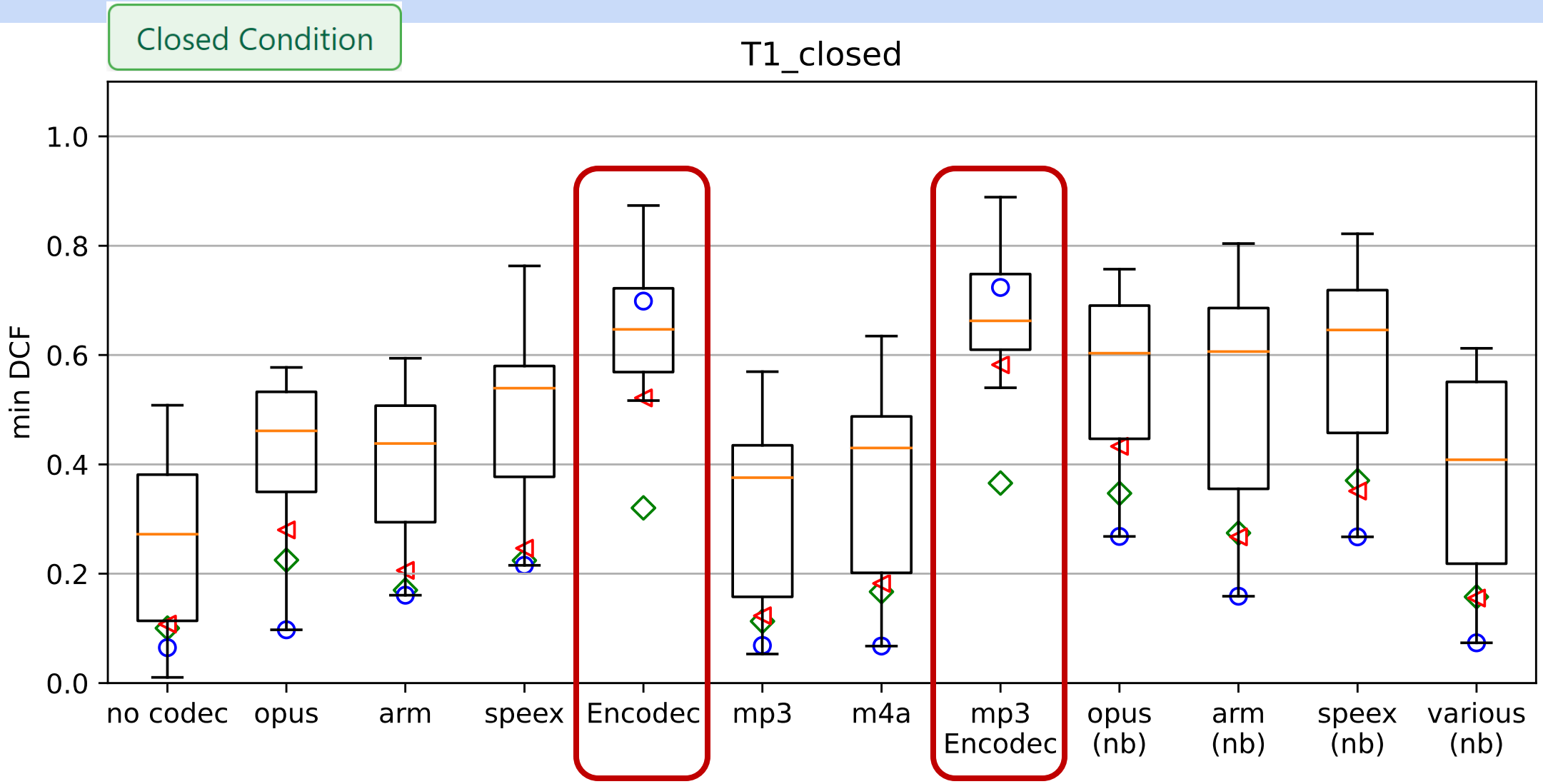
Analysis – codec

Closed Condition



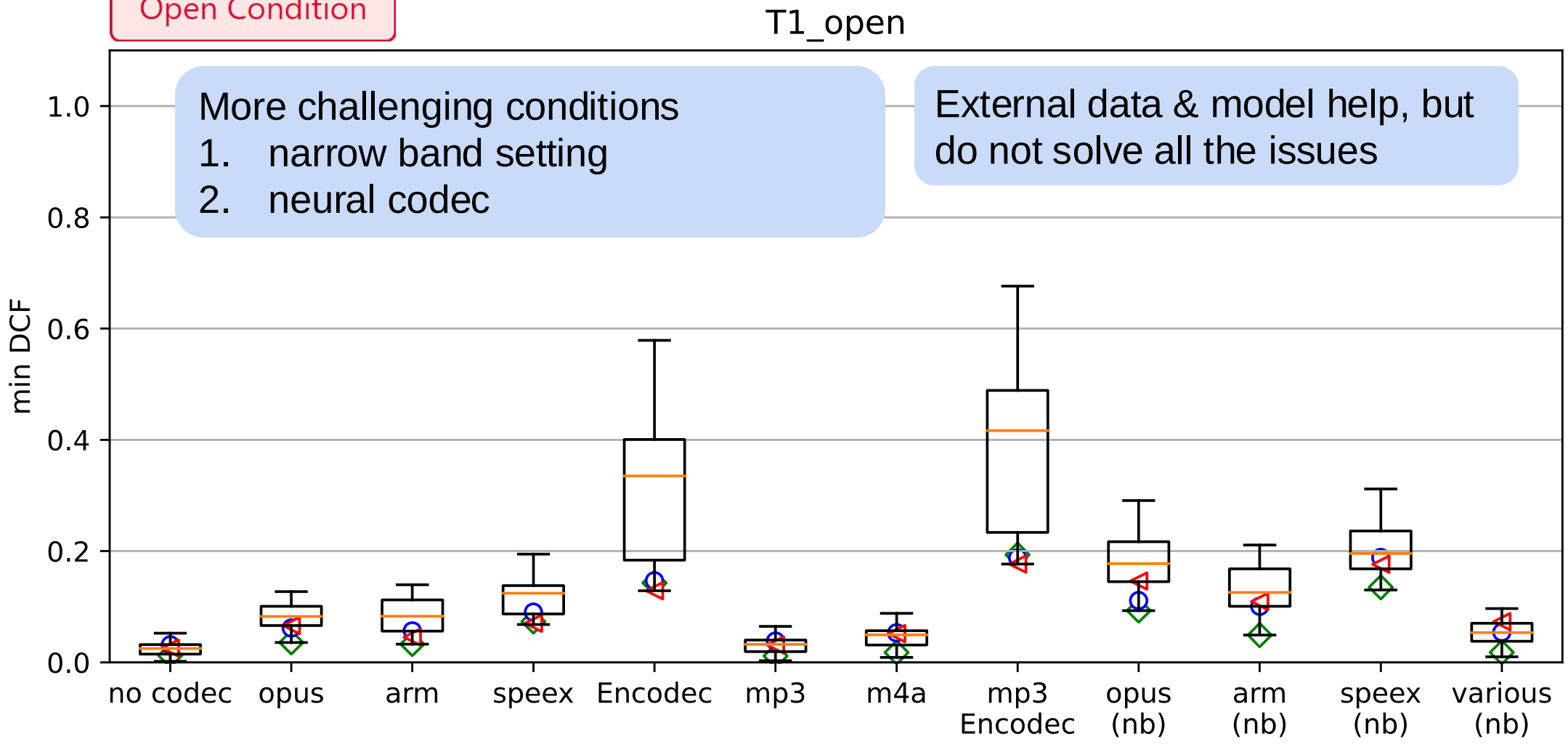
The same set of utterances across different conditions

Analysis – codec



Analysis – codec

Open Condition

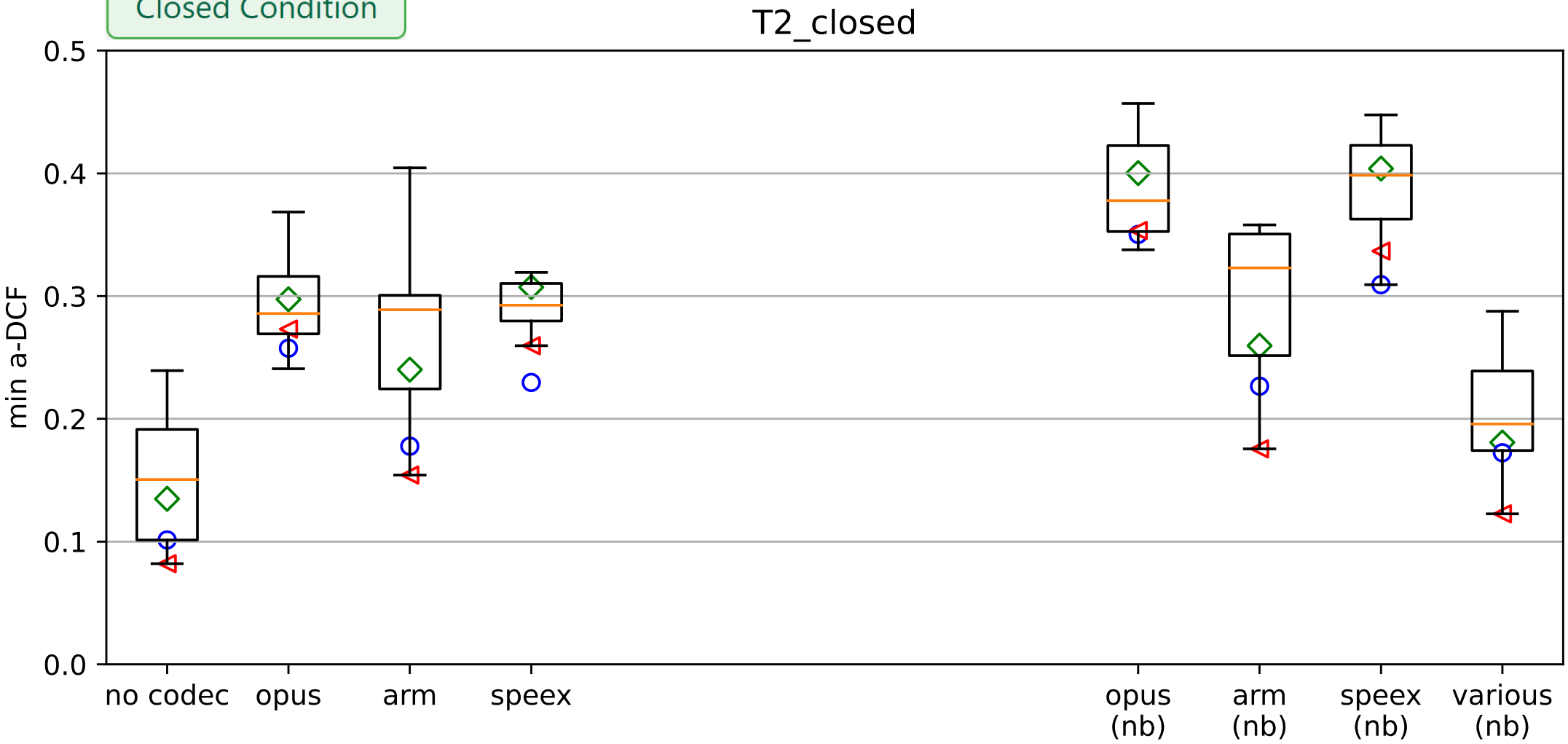


More challenging conditions
1. narrow band setting
2. neural codec

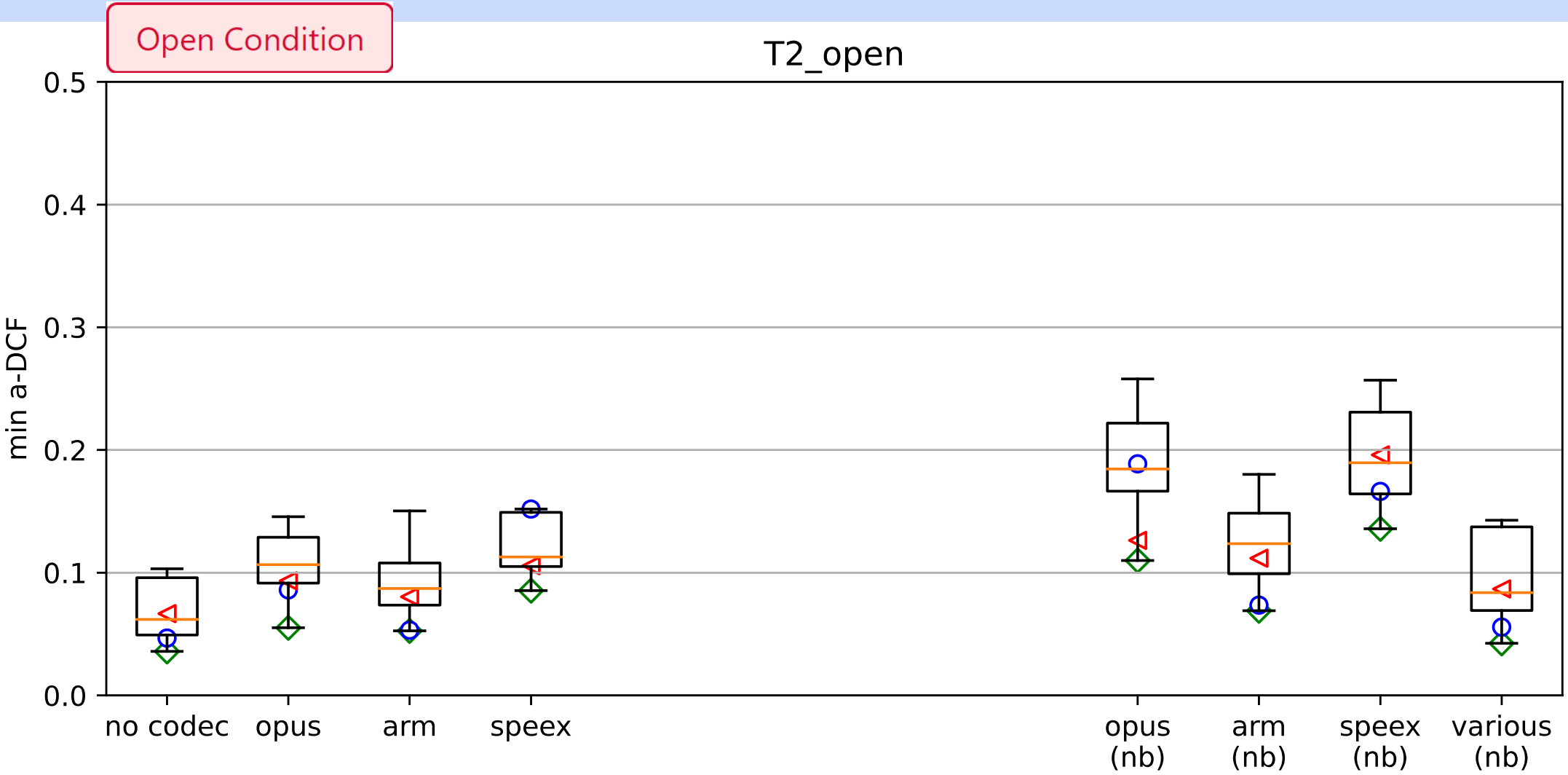
External data & model help, but do not solve all the issues

Analysis – codec

Closed Condition

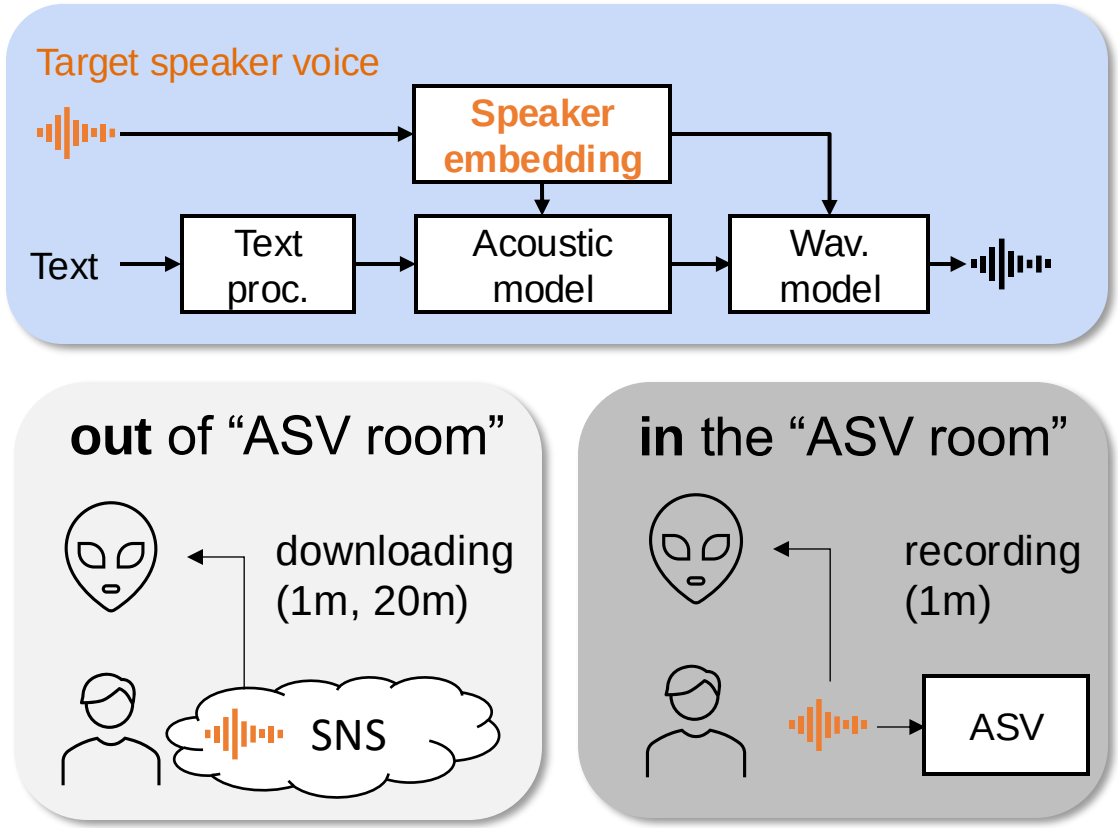


Analysis – codec

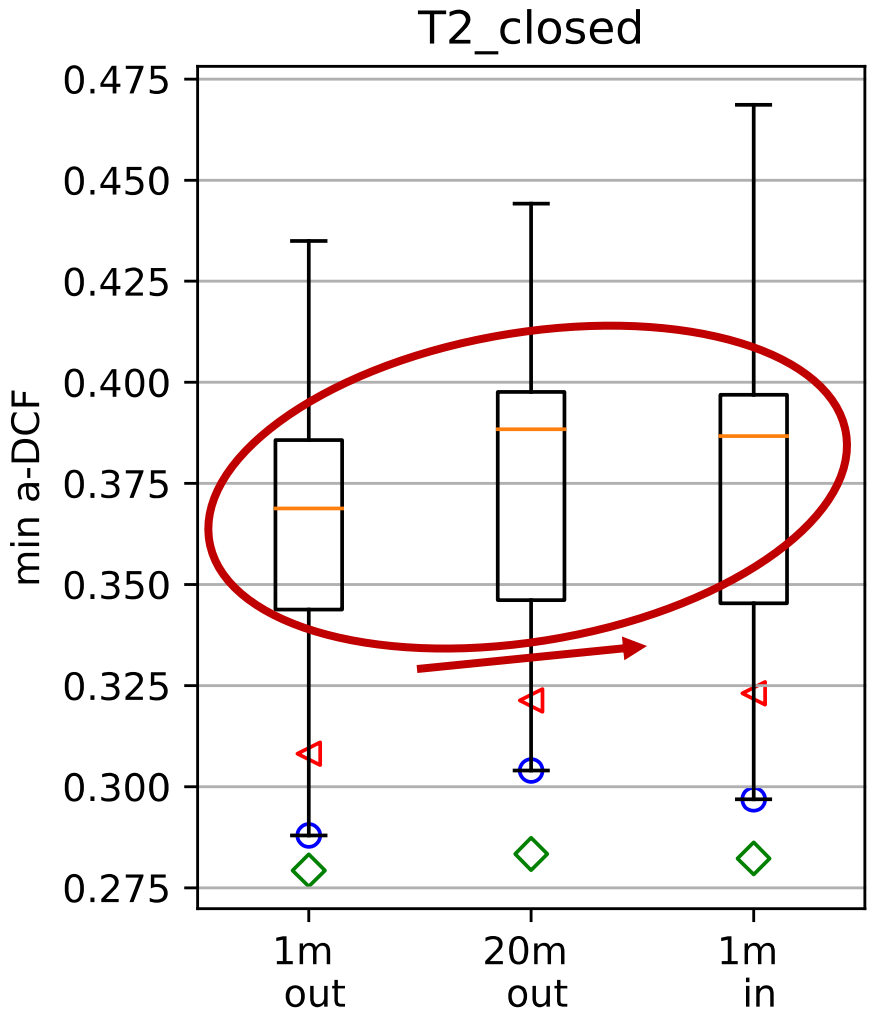


Analysis – attacker source

- Source and amount of target speaker voice

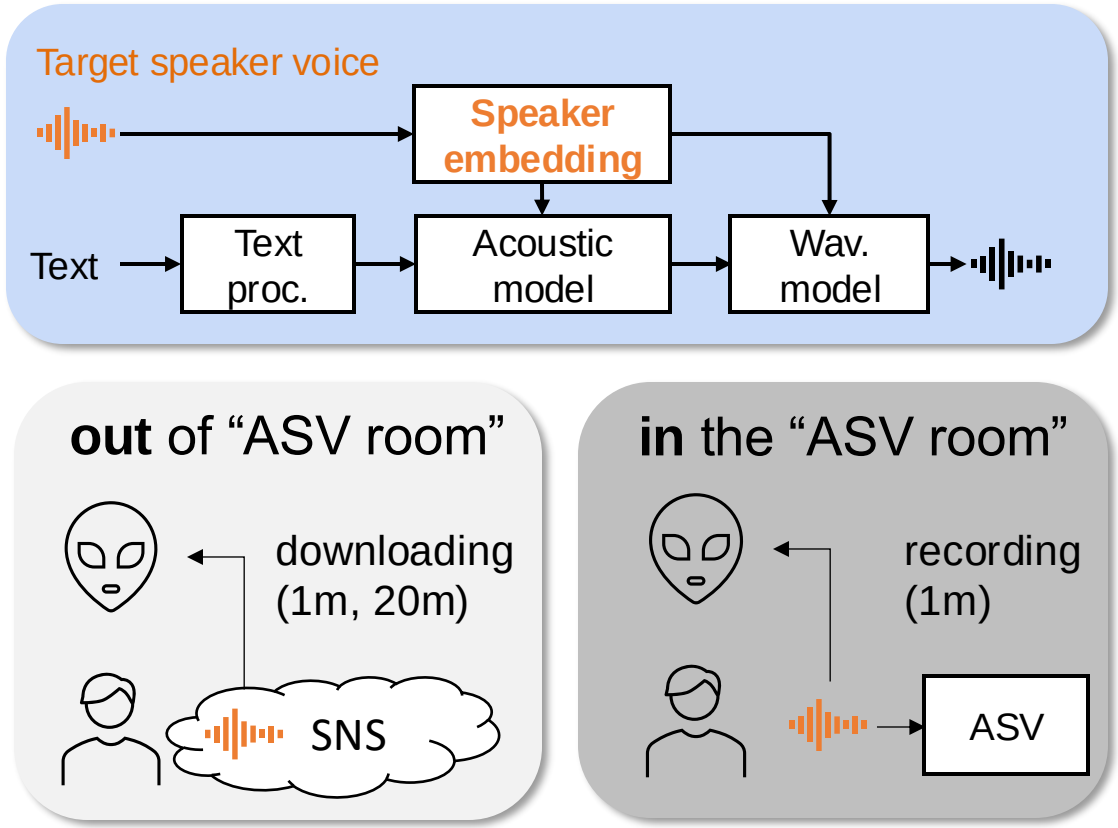


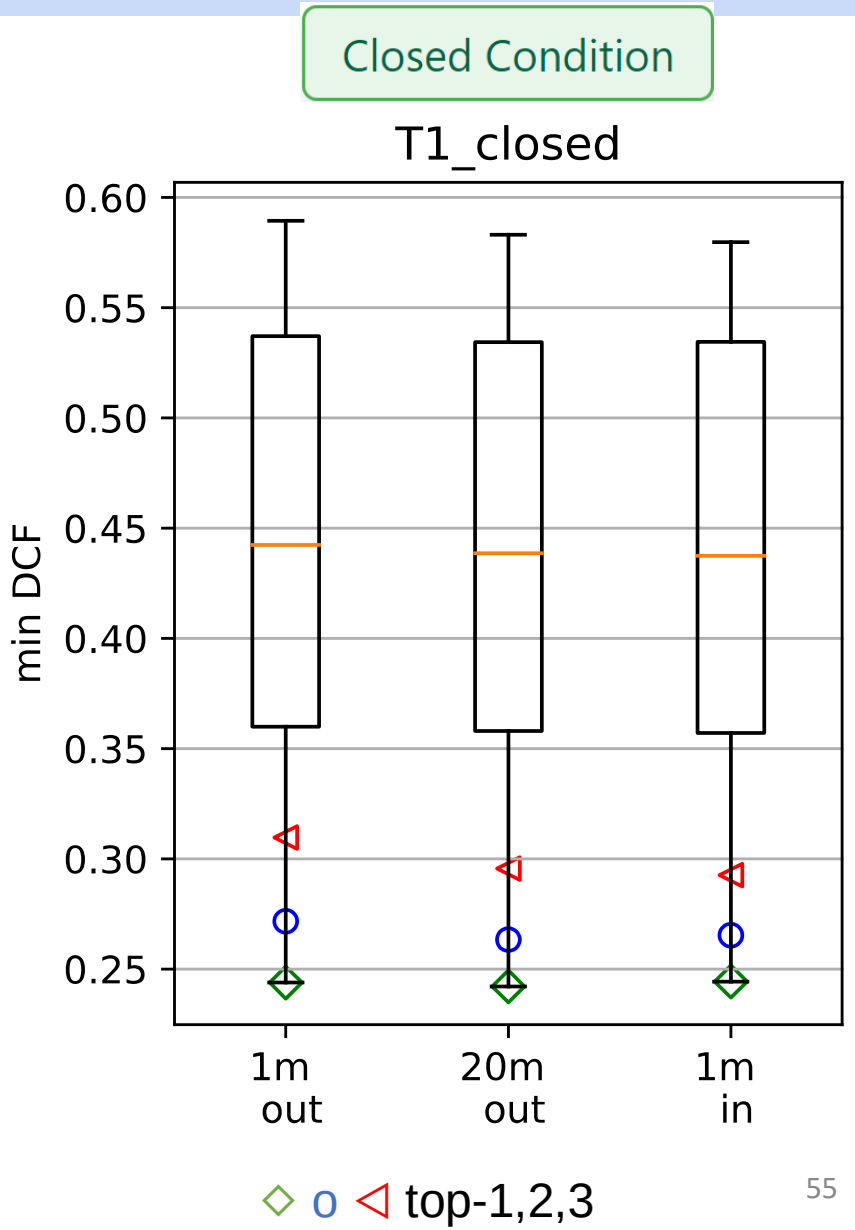
Closed Condition



Analysis – attacker source

- Source and amount of target speaker voice





Summary & dicussion

Summary

- challenges
  - non-studio quality data
  - adversarial attacks compromising ASV and CM
  - codecs, especially neural codecs
  - gap between progress and evaluation sets

- despite increased difficulty, substantial improvements
  - over baselines
  - in open conditions

- lack of score calibration in many submissions

Discussion

1. What do you think of the increased challenge/data complexity?
    1. Do you prefer to see again compressed/noisy data in future challenges – is this relevant to your research/development?
    2. How did you like the two tracks (CM and the new SASV)?
    3. Do you prefer to see neural audio codecs again – and is this "bonafide" or "spoof" anyway?
    4. Should we include more languages? Which ones (and why)?
    5. How about adversarial attacks?
2. Do you like surprises (unseen attacks, codecs etc) in eval set? Will this help us towards generalization to the unknown?
3. How did you like the inclusion of calibration-related metrics? What kind of data or tasks you'd like to see in future?
4. How well do the findings from ASVspoof challenges translate to industry practices? Are we missing anything from real-world applications?
5. Do we have life beyond SSLs and data augmentation?
6. Any fresh ideas on data collection (updating spoofing attacks and beyond)?
7. Outside of the challenge, what do think about the evolving speech generation technology? Will spoofing artifacts exist in the future as well?

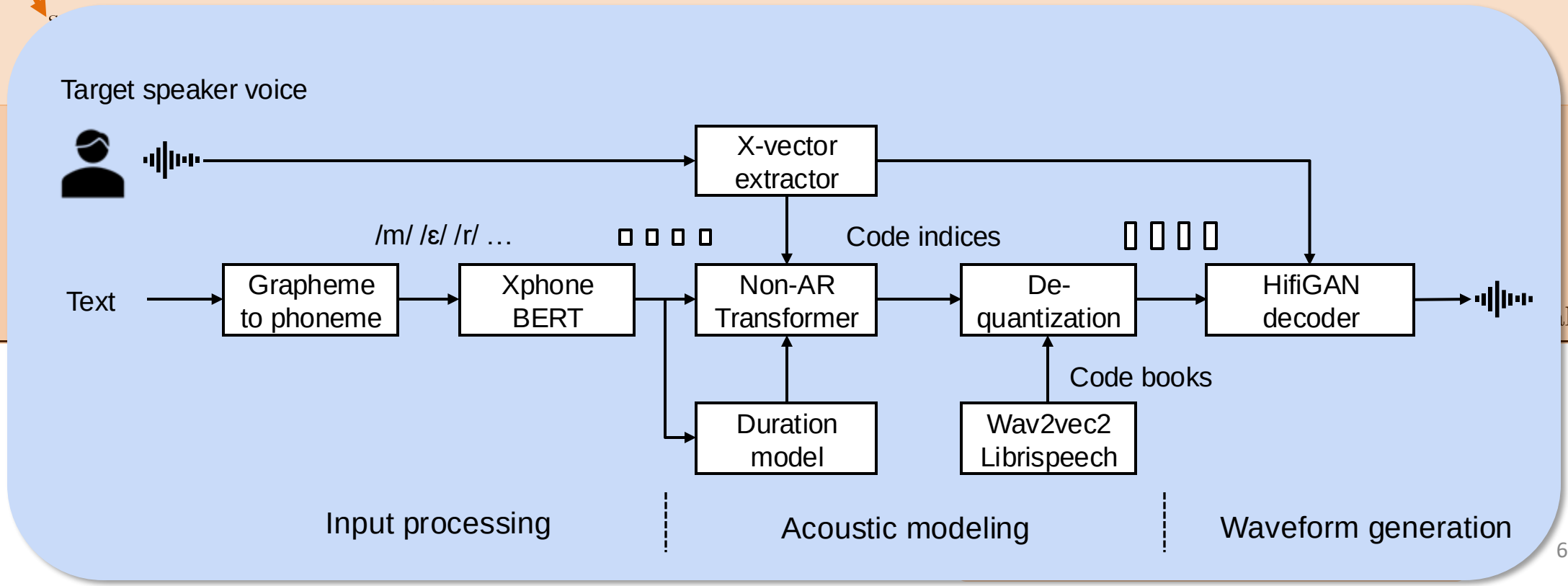
Acknowledgement

- We wish to thank:
  - **Data contributors:** Cheng Gong, Tianjin University; Chengzhe Sun, Shuwei Hou, Siwei Lyu, University at Buffalo, State University of New York; Florian Lux, University of Stuttgart; Ge Zhu, Neil Zhang, Yongyi Zang, University of Rochester; Guo Hanjie and Liping Chen, University of Science and Technology of China; Hengcheng Kuo and Hung-yi Lee, National Taiwan University; Myeonghun Jeong, Seoul National University; Nicolas Muller, Fraunhofer AISEC; Sebastien Le Maguer, University of Helsinki; Soumi Maiti, Carnegie Mellon University; Yihan Wu, Renmin University of China; Yu Tsao, Academia Sinica; Vishwanath Pratap Singh, University of Eastern Finland; Wangyou Zhang, Shanghai Jiaotong University.
  - Challenge participants/authors
  - Reviewers
  - **A⋆ STAR** (Singapore) for sponsoring CodaLab platform

Appendix

Dataset: spoofed data (eval. set)

Attack	Input	Input processor	Acoustic model	Speaker embedding	Acoustic feature	Waveform model	Post-processing
A21	text	NLP	FS-based	GST	log-spec	BigVGAN	-
A22	text	NLP	FS+ProsodyTransfer	GST	log-spec	BigVGAN	-
A28	text	DNN-encoder	YourTTS(pre.)	H/ASP	latent	HifiGAN	-
A26	speech	ASR	DNN+F0 est.	CAM++	Mel-spec	HifiGAN	original genuine noise
A17	text	BERT-based	Transformer-based	x-vector	DNN-latent	HifiGAN	-
A19	text	NLP	MaryTTS	-	-	unit-selec	-
A24							
A25							
A29							
A23							
A20							
A18							
A27							
A31							
A32							
A30							lacopula

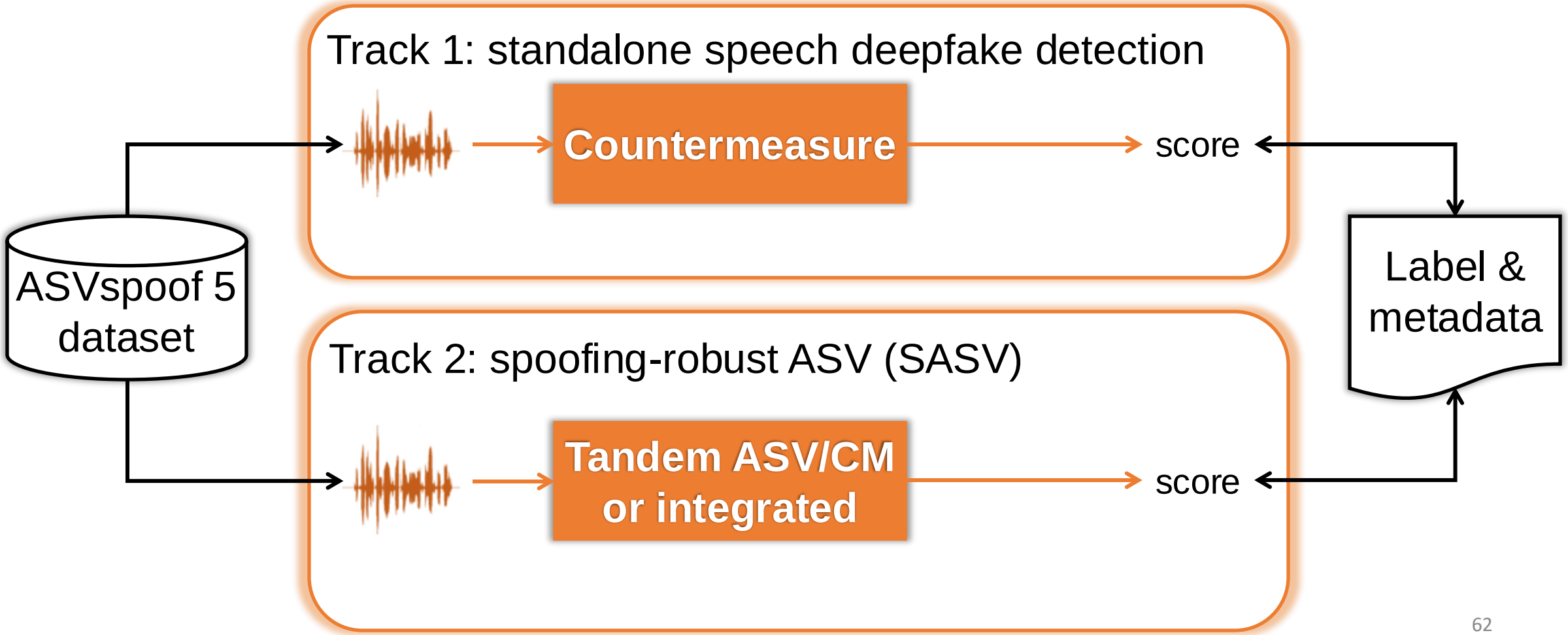


ASVspoof 5

Organisers: dataset creation

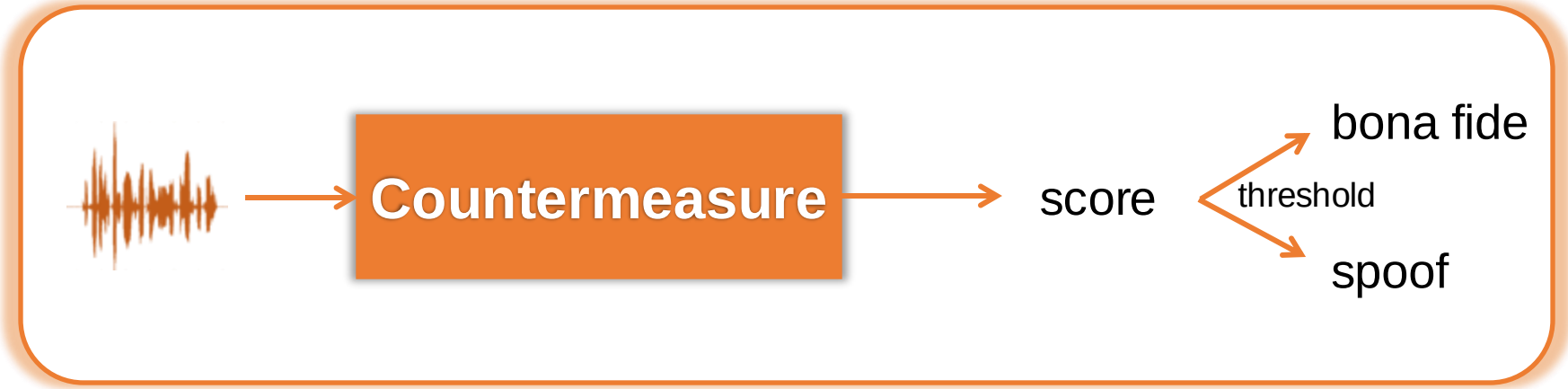
Participants: system building & scoring

Organisers: evaluation



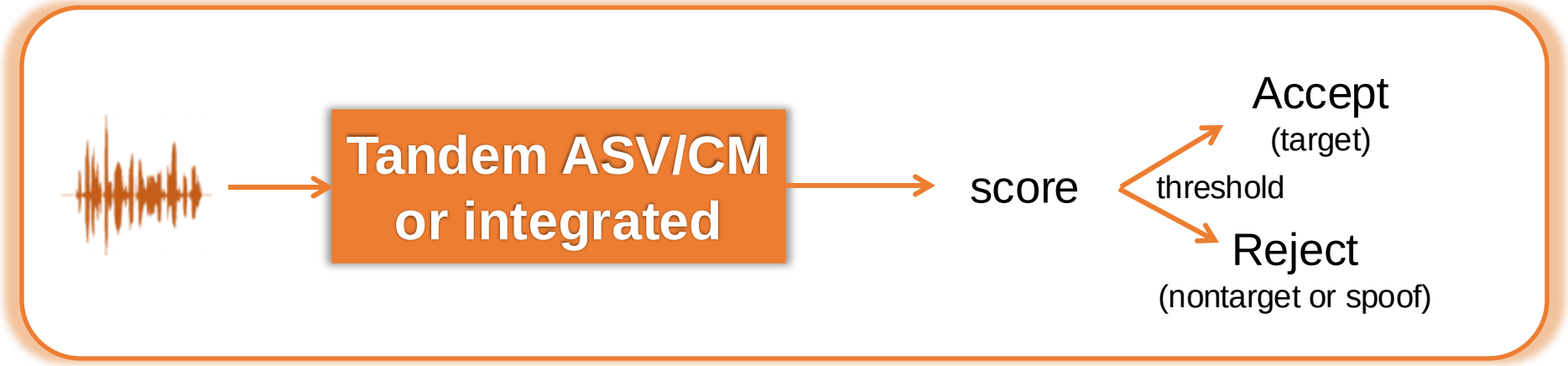
Track 1: speech deepfake detection

- Binary classification on a single audio file: bona fide or spoof
  - Inherits the DF track of previous ASVspoof challenges
  - An attacker has access to the voice data of a targeted victim
  - Conventional/neural codecs can be used



Track 2: SASV

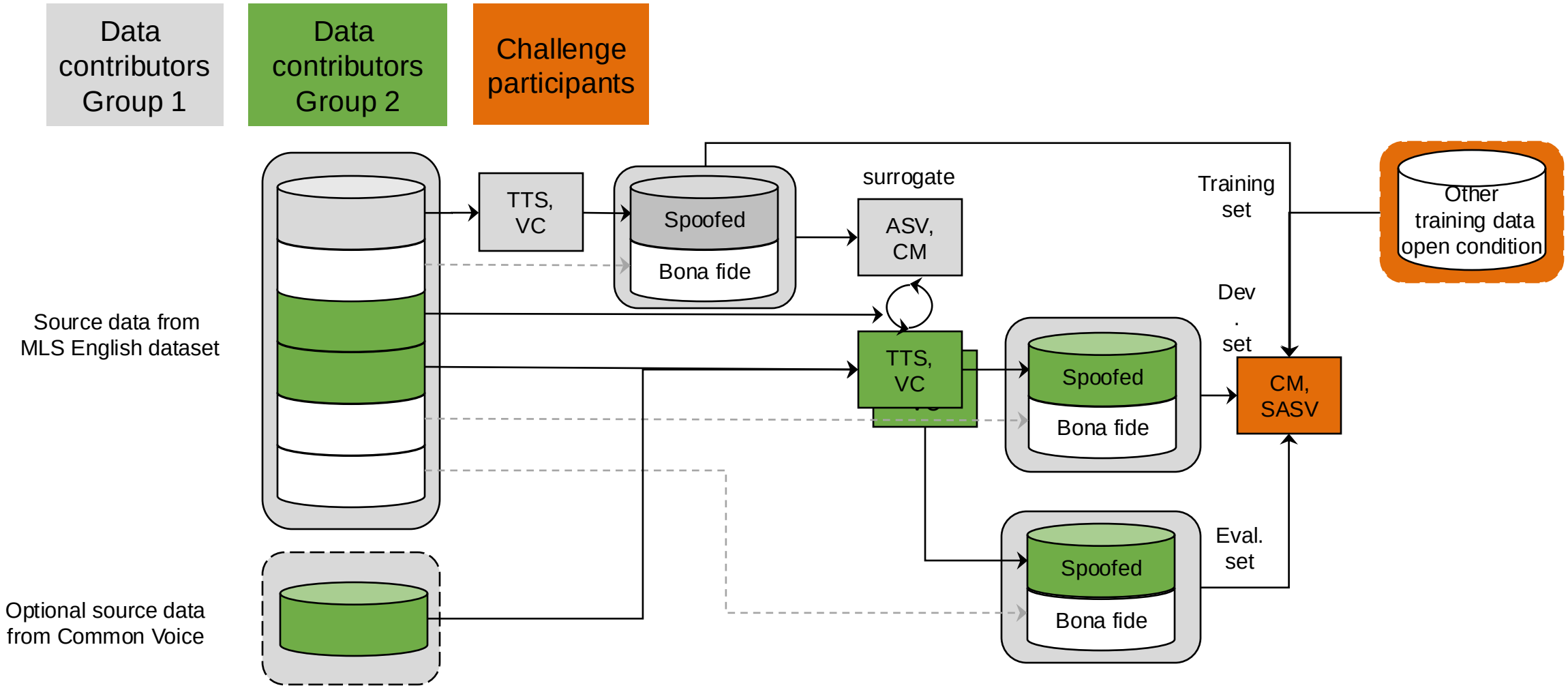
- Binary classification on a pair of audio files:
  - Inherits the LA track of previous ASVspoof and SASV2022 challenges
  - Three trial types (target / non-target / spoof), two decisions (accept / reject)
  - Systems can be a tandem fusion of ASV and CM or single models



Two conditions: open / closed

- For both tracks, there are two conditions:
  - Closed:
    - Restricted data protocols
    - Strictly train with only ASVspoof 5 training partition
    - Exception: VoxCeleb2 for speaker embedding training
  - Open:
    - External data and pre-trained models are allowed
    - Exception: overlapping data with the ASVspoof 5 evaluation partition is prohibited
    - SSL models trained on overlapping data are also not permitted

ASVspoof 5 flow chart



ASVspoof 5 dataset: spoofed data (eval. set)

Attack	Input	Input processor	Acoustic model	Speaker embedding	Acoustic feature	Waveform model	Post-processing
A21	text	NLP	FS-based	GST	log-spec	BigVGAN	-
A22	text	NLP	FS+ProsodyTransfer	GST	log-spec	BigVGAN	-
A28	text	DNN-encoder	YourTTS(pre.)	H/ASP	latent	HifiGAN	-
A26	speech	ASR	DNN+F0 est.	CAM++	Mel-spec	HifiGAN	original genuine noise
A17	text	BERT-based	Transformer-based	x-vector	DNN-latent	HifiGAN	-
A19	text	NLP	MaryTTS	-	-	unit-selec	-
A24	speech	PPG	DNN	x-vector	LSP	HifiGAN	-
A25	speech	DNN-encoder	DiffVC	latent	Mel-spec	HifiGAN	-
A29	text	DNN-encoder	XTTS(pre.)	ECAPA2	latent	HifiGAN	-
A23	A09	-	-	-	-	-	Malafide
A20	A12	-	-	-	-	-	Malafide
A18	A17	-	-	-	-	-	Malafide
A27	A26	-	-	-	-	-	Malacopula
A31	A22	-	-	-	-	-	Malacopula
A32	A25	-	-	-	-	-	Malacopula
A30	A18	-	-	-	-	-	Malafide+Malacopula

A21	A22	A28	A26	A17
A19	A24	A25	A29	A23
A20	A18	A27	A31	A32
		A30		

Condition C11

- Telephone simulation by a swept sine approach. The swept sine signal is transmitted though a call to a call center.
- Captured methods:
  - Microsoft Teams call. Audio digitally injected using a virtual audio cable driver.
  - Poco F4. Audio digitally injected via Bluetooth.
  - Redmi Note 8 Pro. Audio digitally injected via Bluetooth.
  - Redmi Note 8 Pro. Audio injected via cable to input jack.
  - Samsung Galaxy A12. Audio digitally injected via Bluetooth.
  - Samsung Galaxy A12. Audio injected via cable to input jack.
  - Samsung Galaxy S23 Ultra. Audio digitally injected via Bluetooth.

Codalab

- Codalab platform
- Progress period (06/12 – 07/21)
  - ~1 month
  - **subset of evaluation data**
  - 4 submissions per day
- Evaluation period (07/21 – 07/24)
  - 3 days
  - one submission only

Progress data

Attack	Input	Input processor	Acoustic model	Speak
A21	text	NLP	FS-based	GST
A22	text	NLP	FS+ProsodyTransfer	GST
A28	text	DNN-encoder	YourTTS(pre.)	H/AS
A26	speech	ASR	DNN+F0 est.	CAM
A17	text	BERT-based	Transformer-based	x-vec
A19	text	NLP	MaryTTS	-
A24	speech	PPG	DNN	x-vec
A25	speech	DNN-encoder	DiffVC	laten
A29	text	DNN-encoder	XTTS(pre.)	ECAI
A23	A09	-	-	-
A20	A12	-	-	-
A18	A17	-	-	-
A27	A26	-	-	-
A31	A22	-	-	-
A32	A25	-	-	-
A30	A18	-	-	-

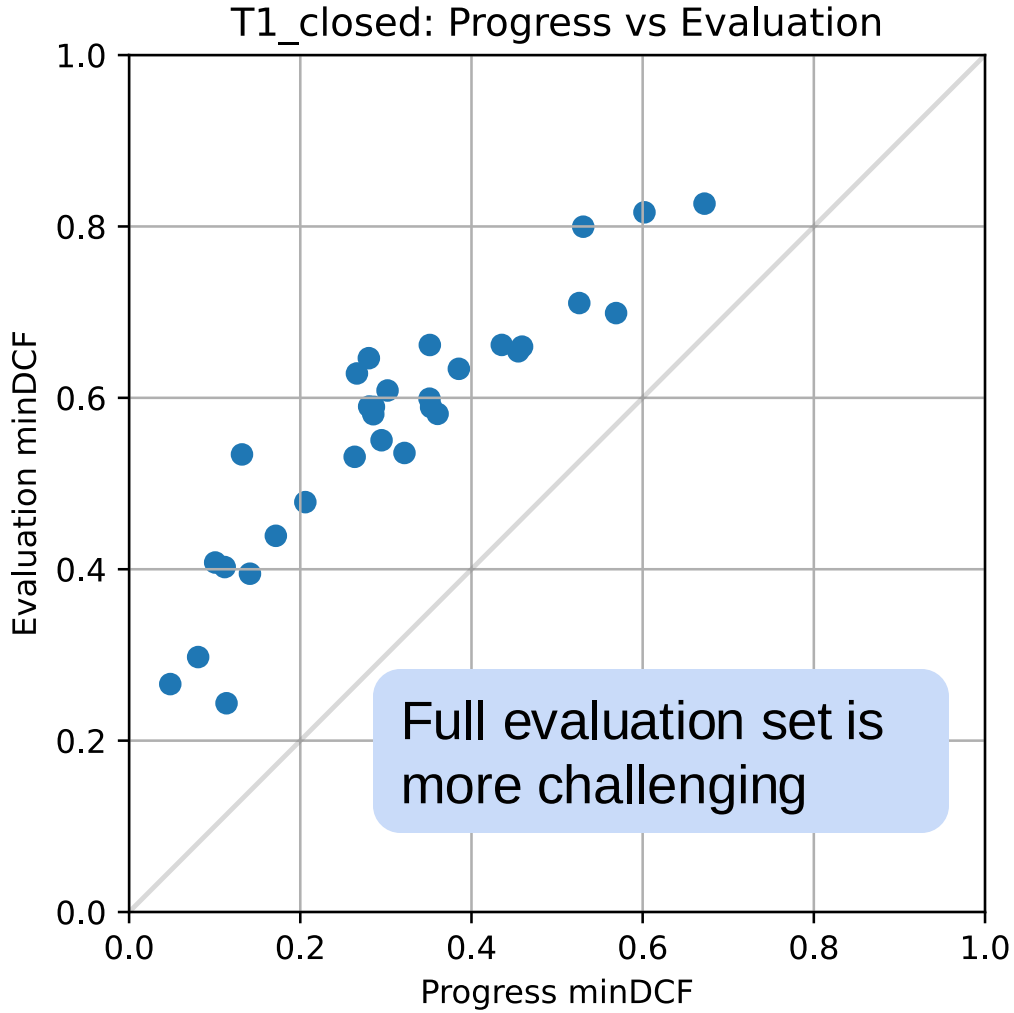
Codecs: no codec, opus, arm, opus(nb), arm (nb)

Progress of developing

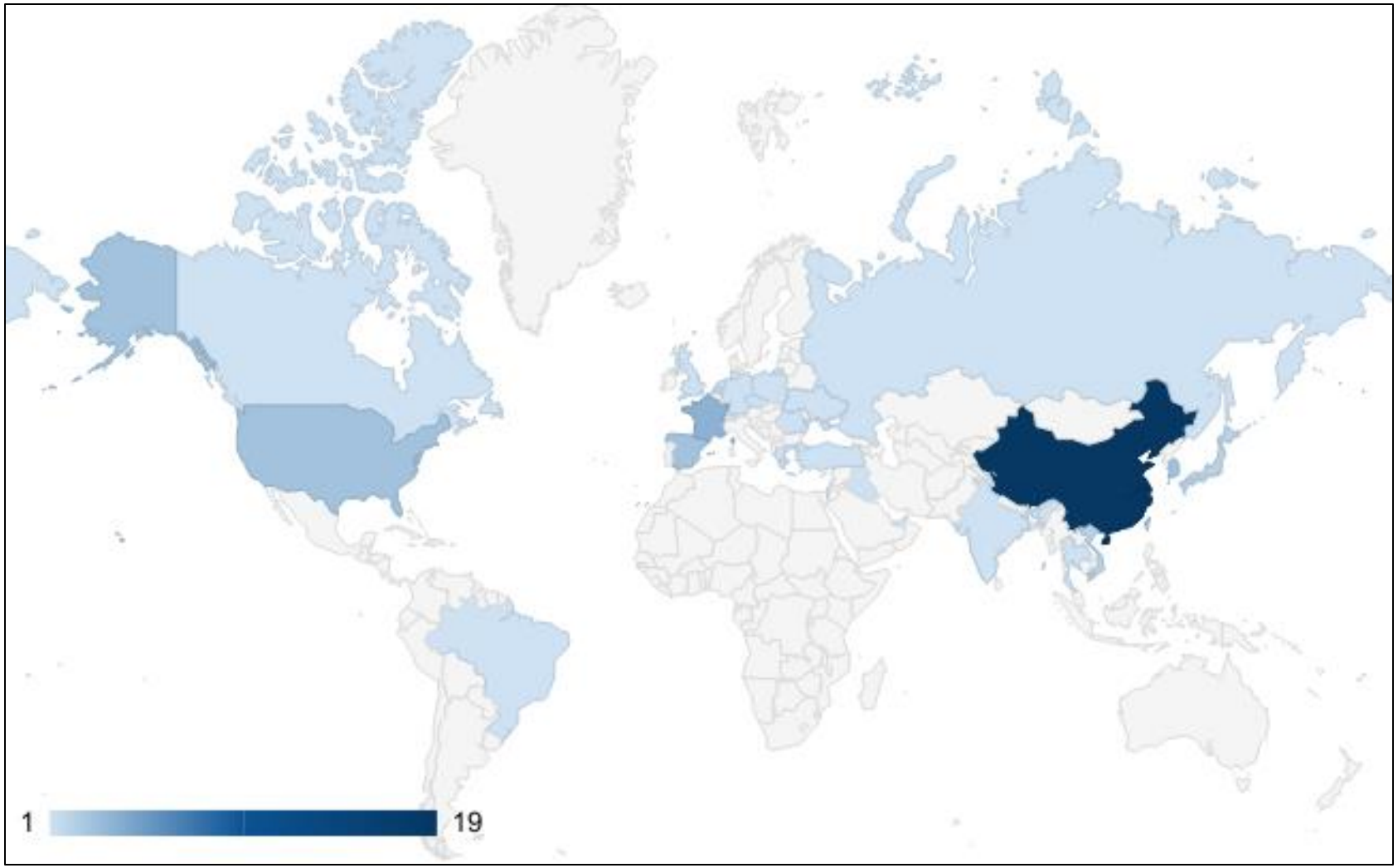
- Codalab

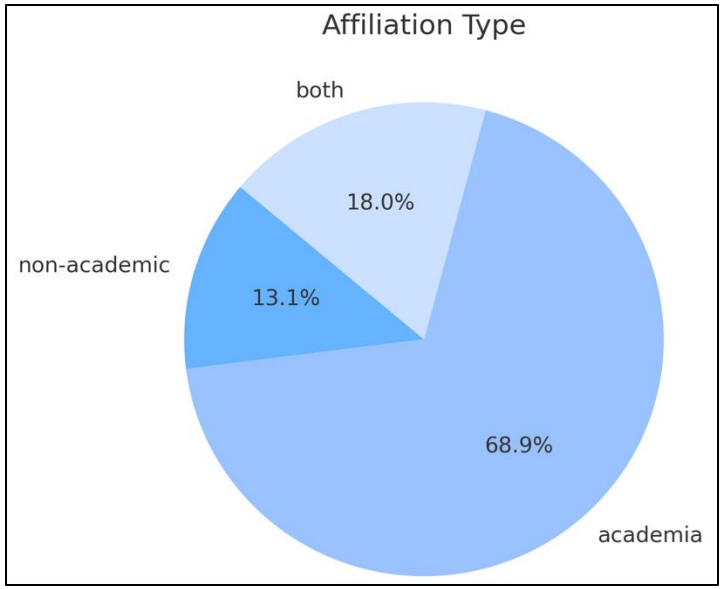
- Progress period
  - ~1 month
  - subset of evaluation data
  - 4 submissions per day

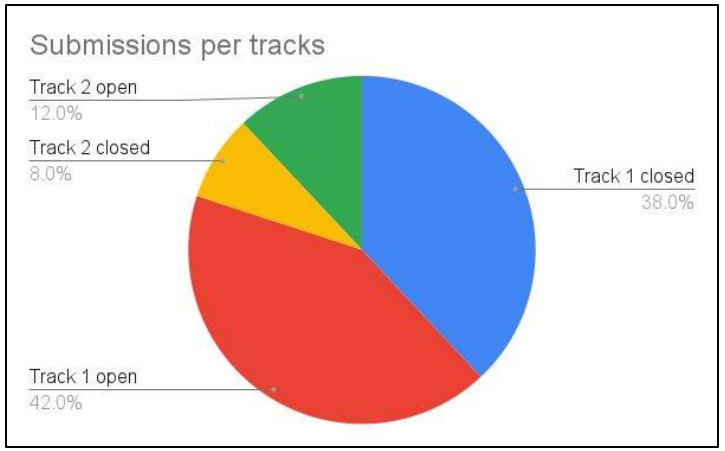
- Evaluation period
  - 3 days
  - one submission only



Participants submitted in eval. phase

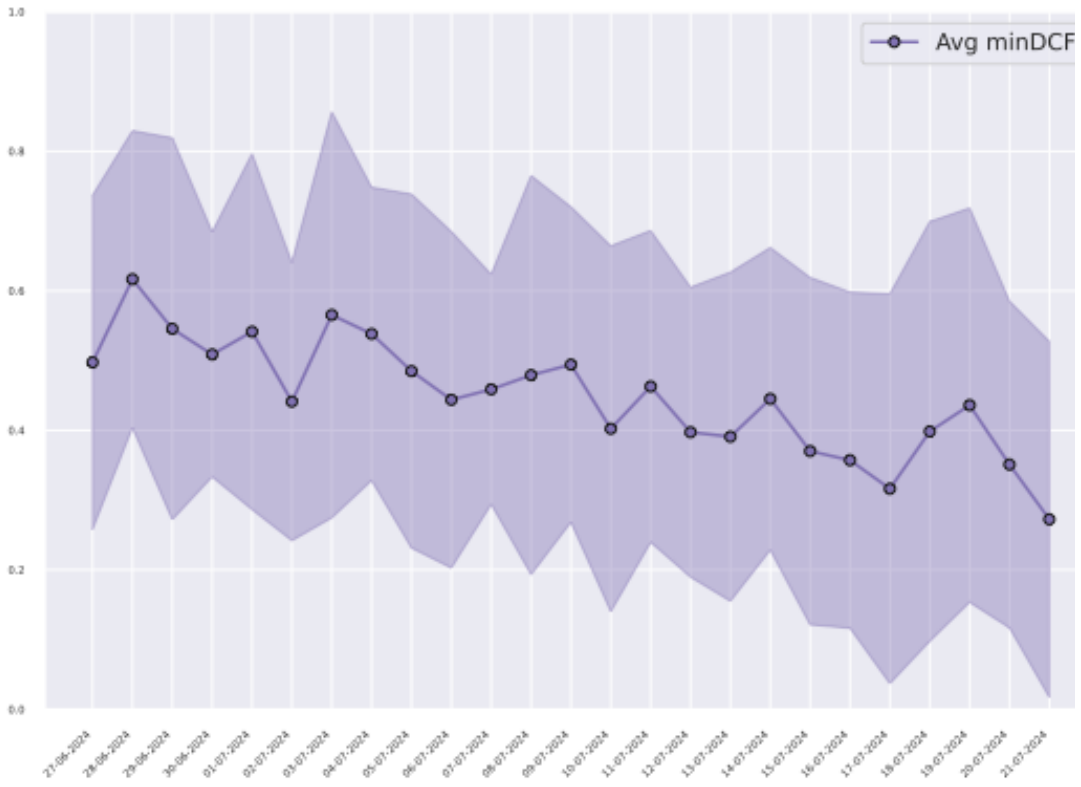




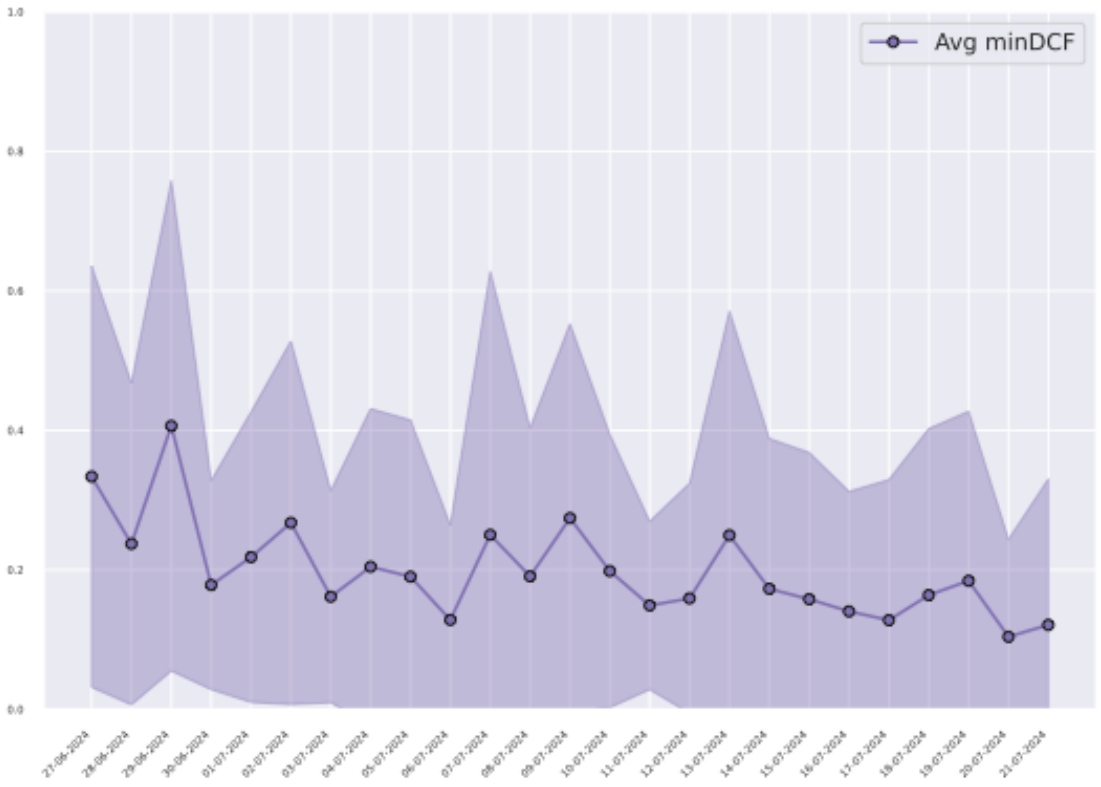


Codalab

Track 1 closed

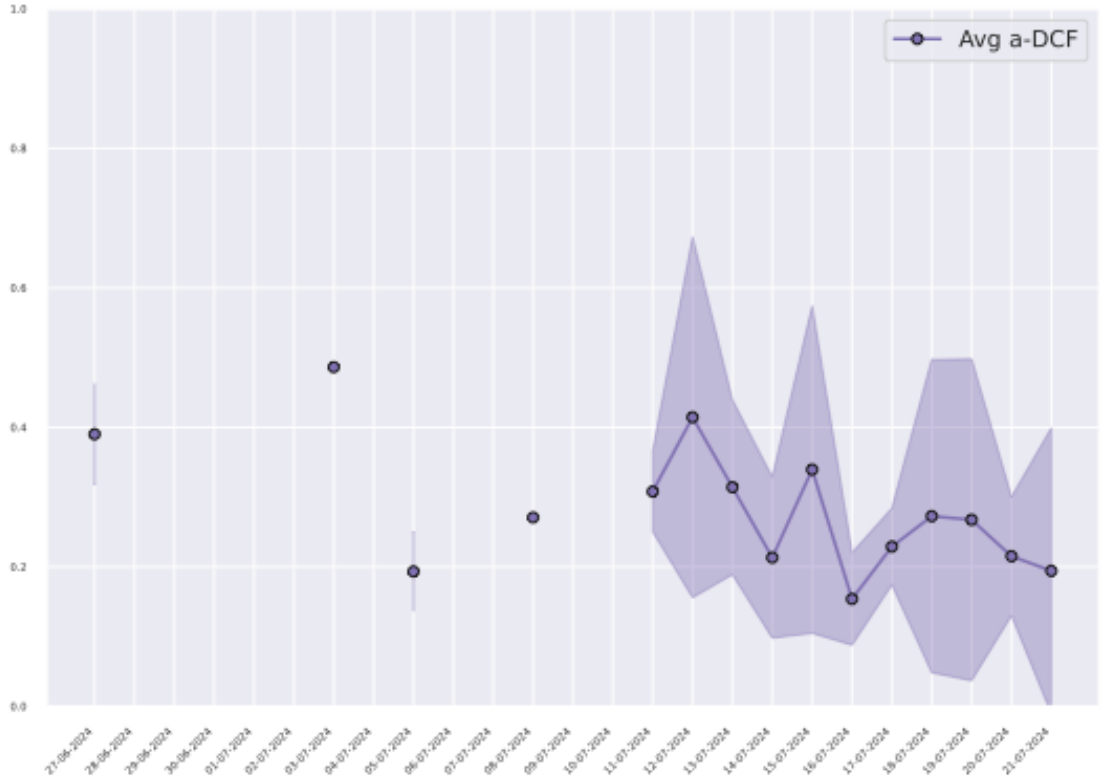


Track 1 open

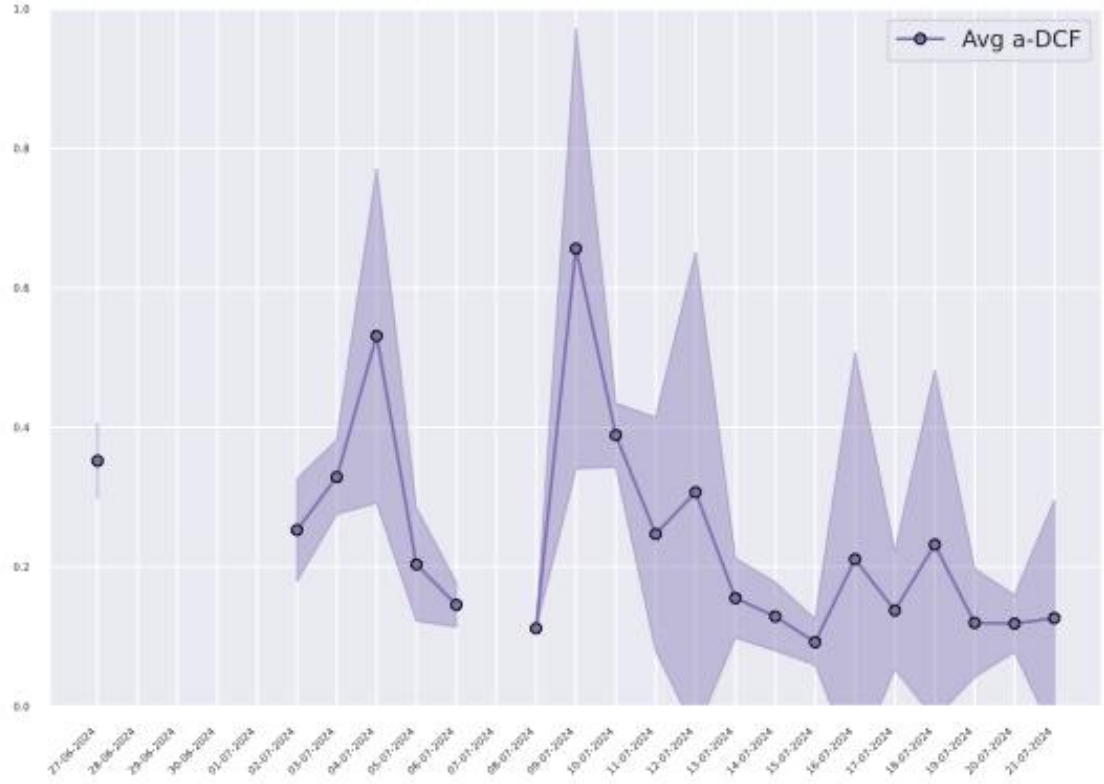


Codalab

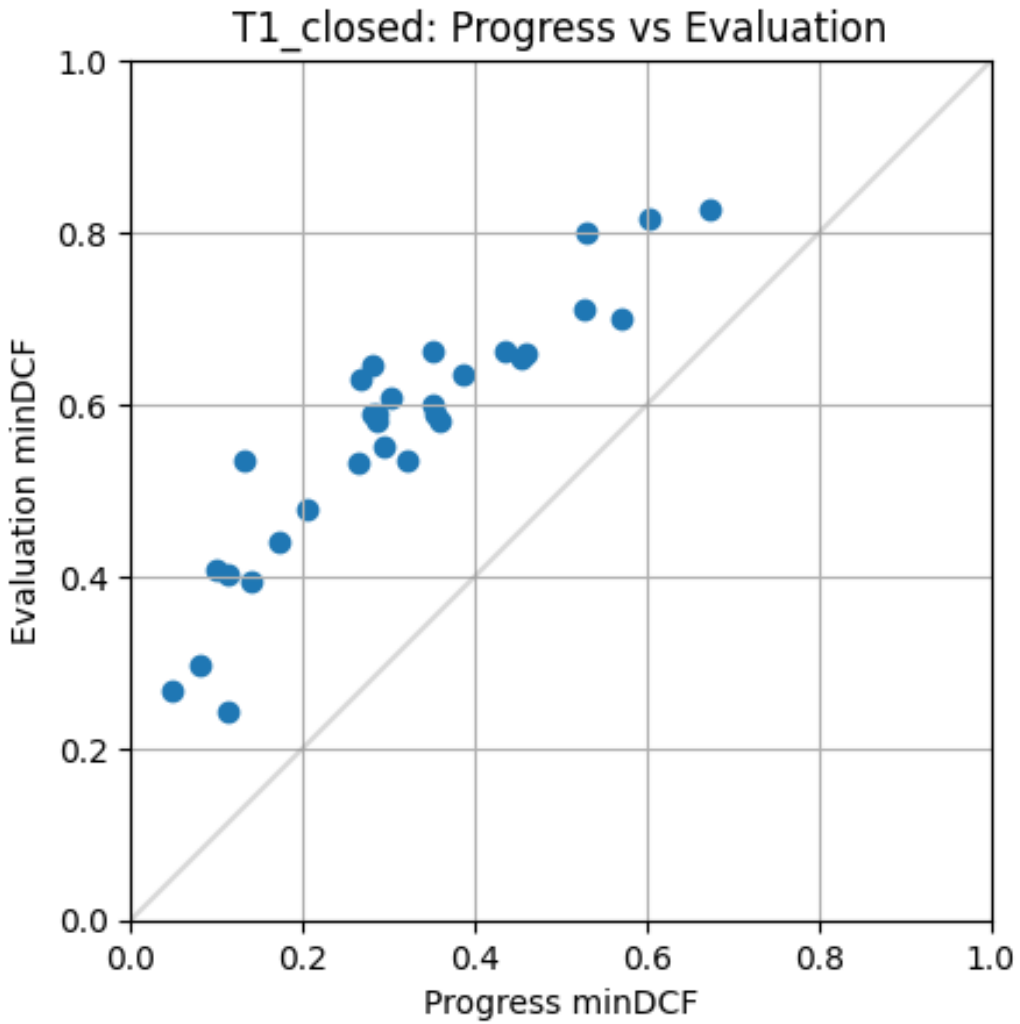
Track 1 closed

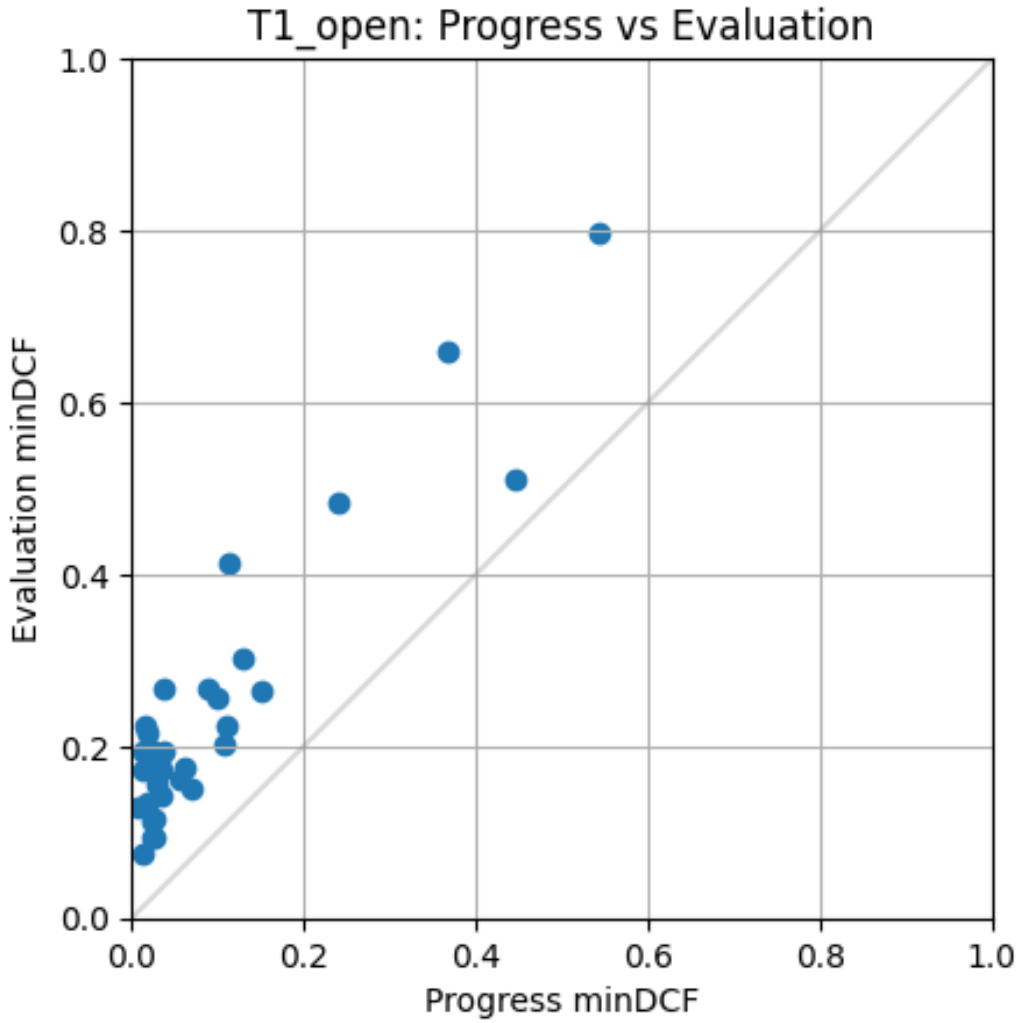


Track 1 open

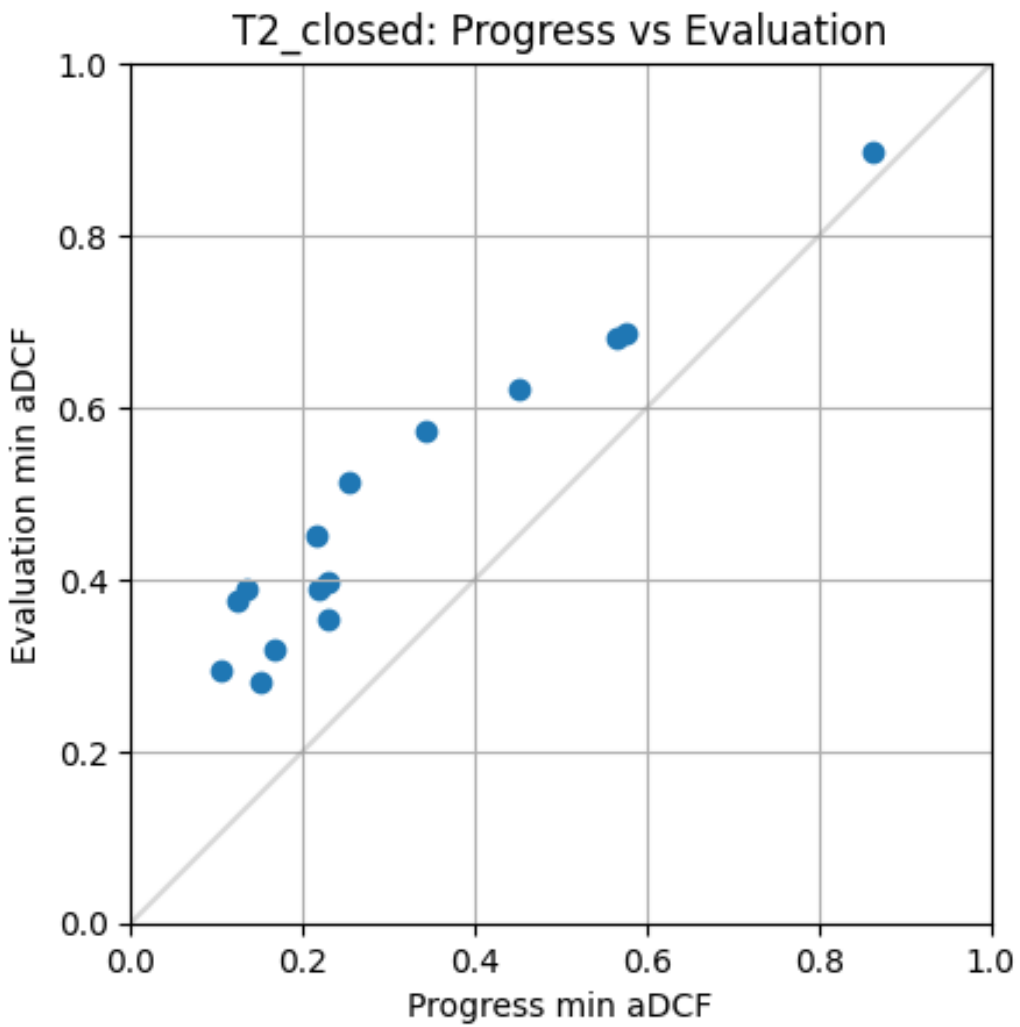


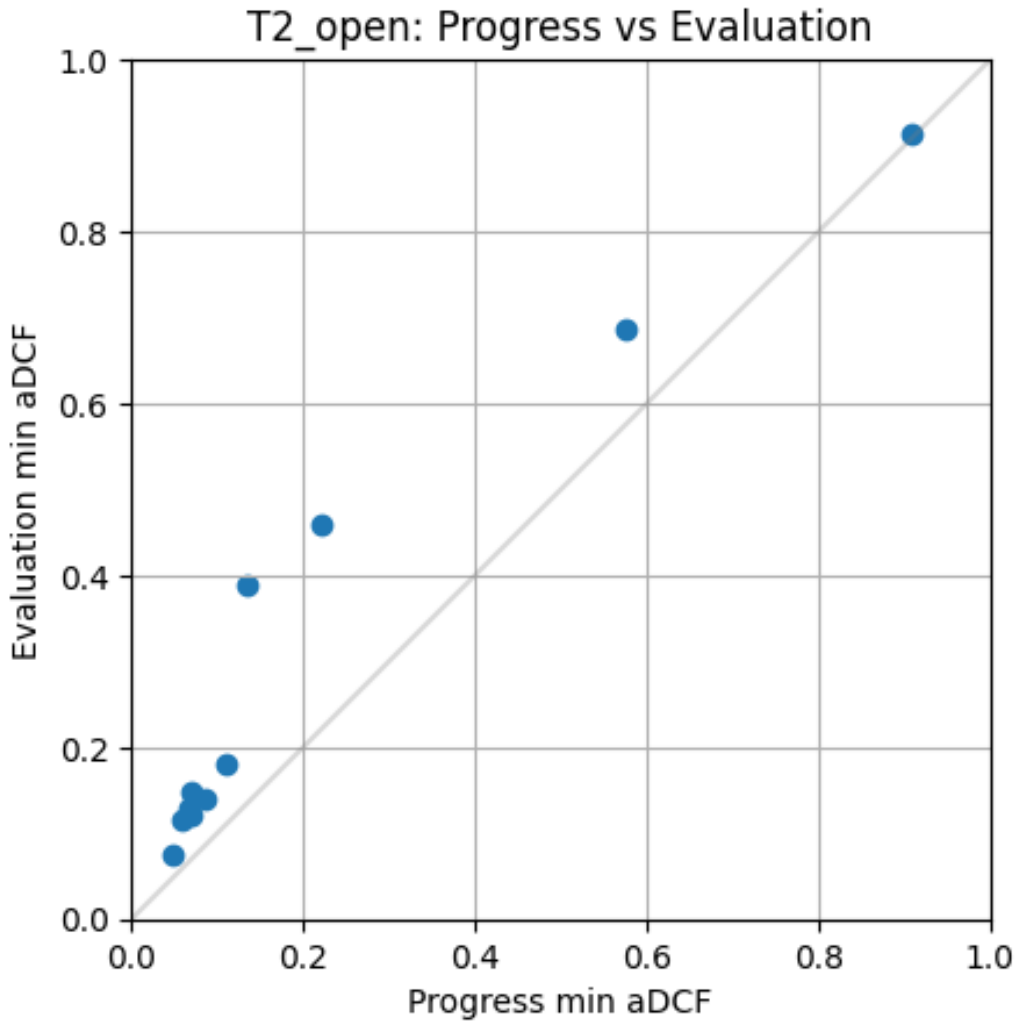
Track 1, Progress vs Evaluation minDCF





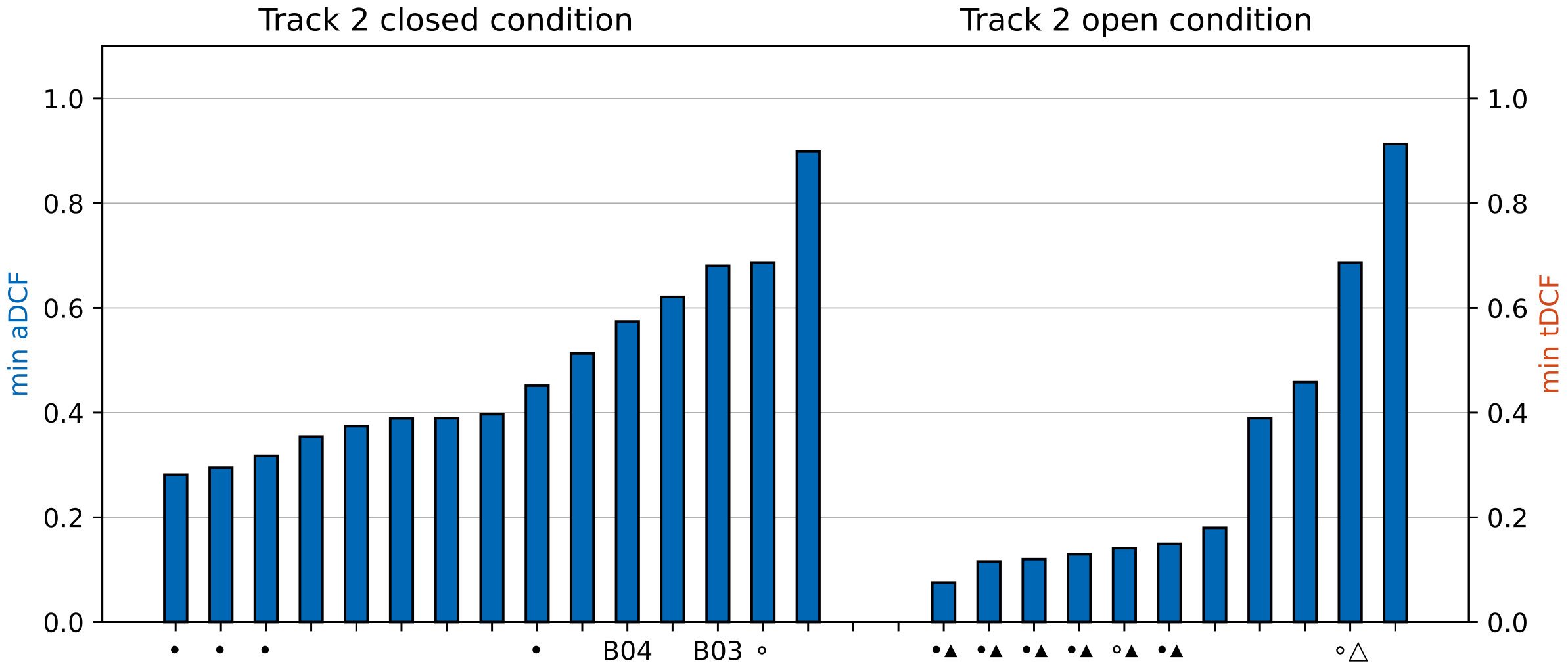
Track 2, Progress vs Evaluation minDCF



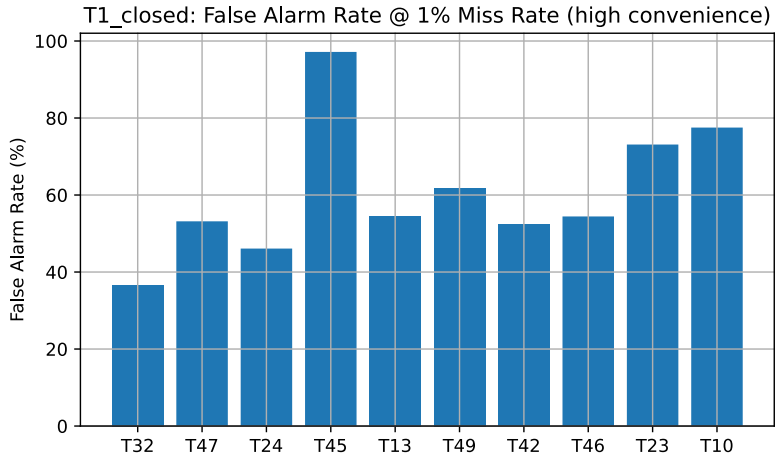
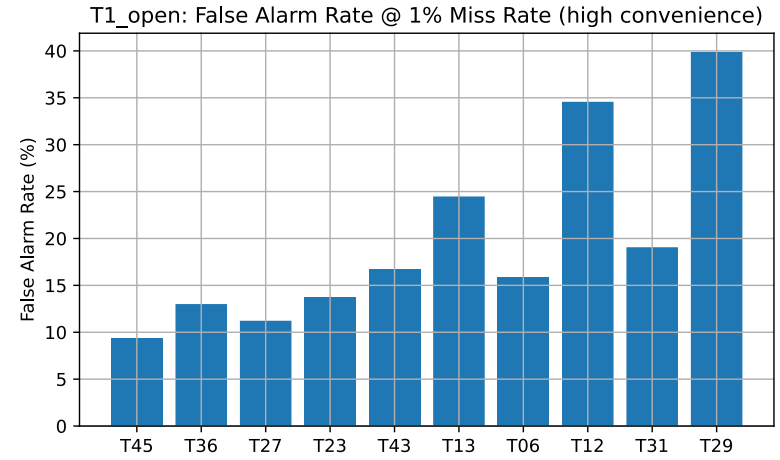
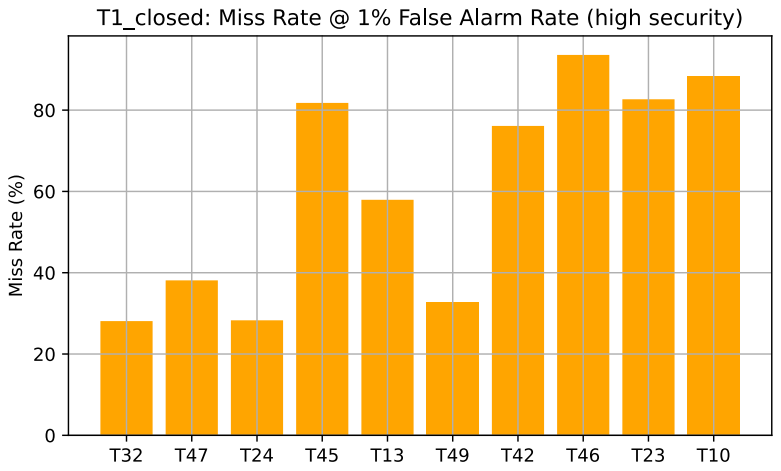
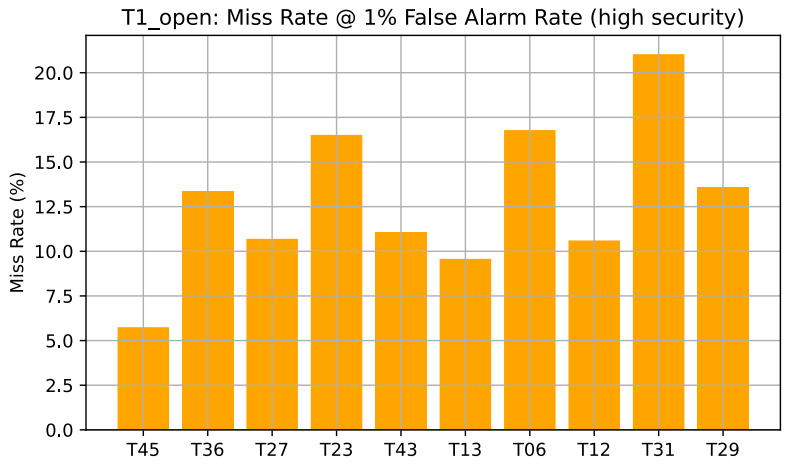


Appendix - results

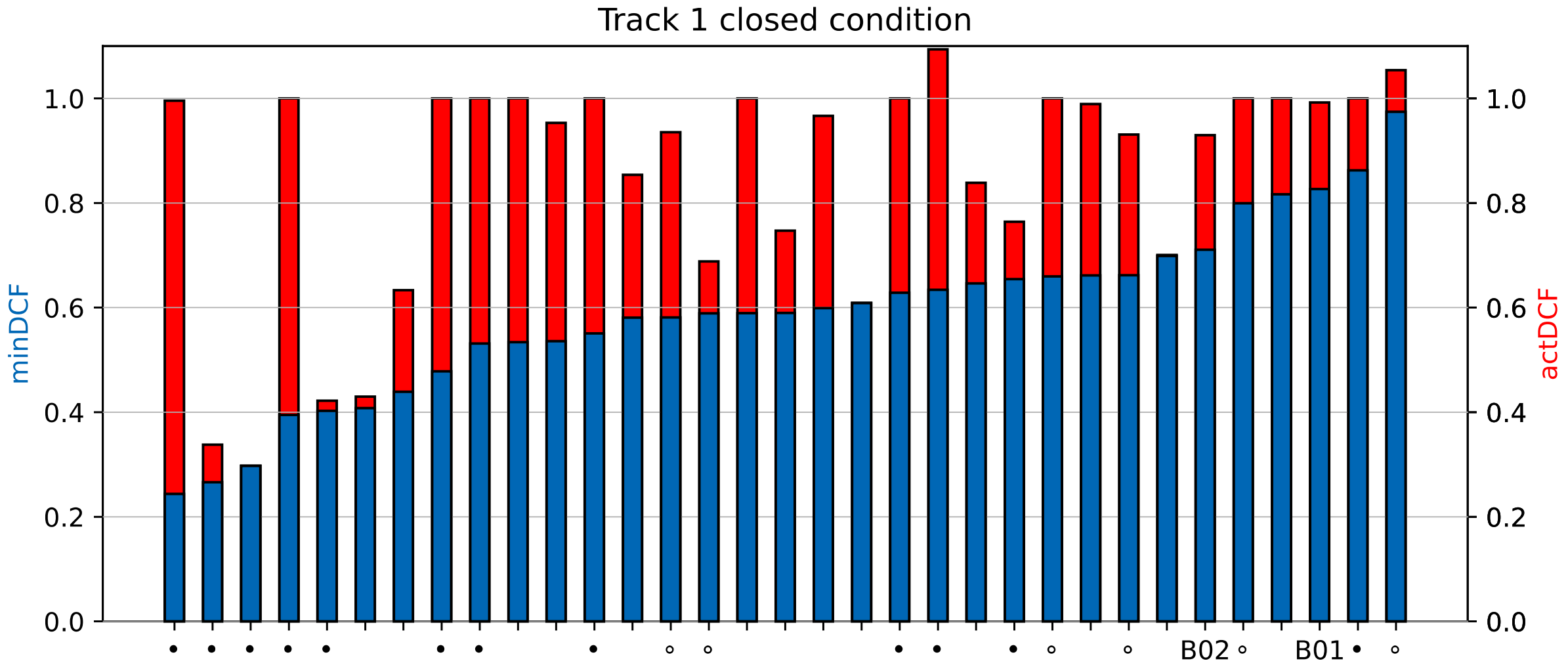
Track 2 - overall results



Track 1 additional operation points

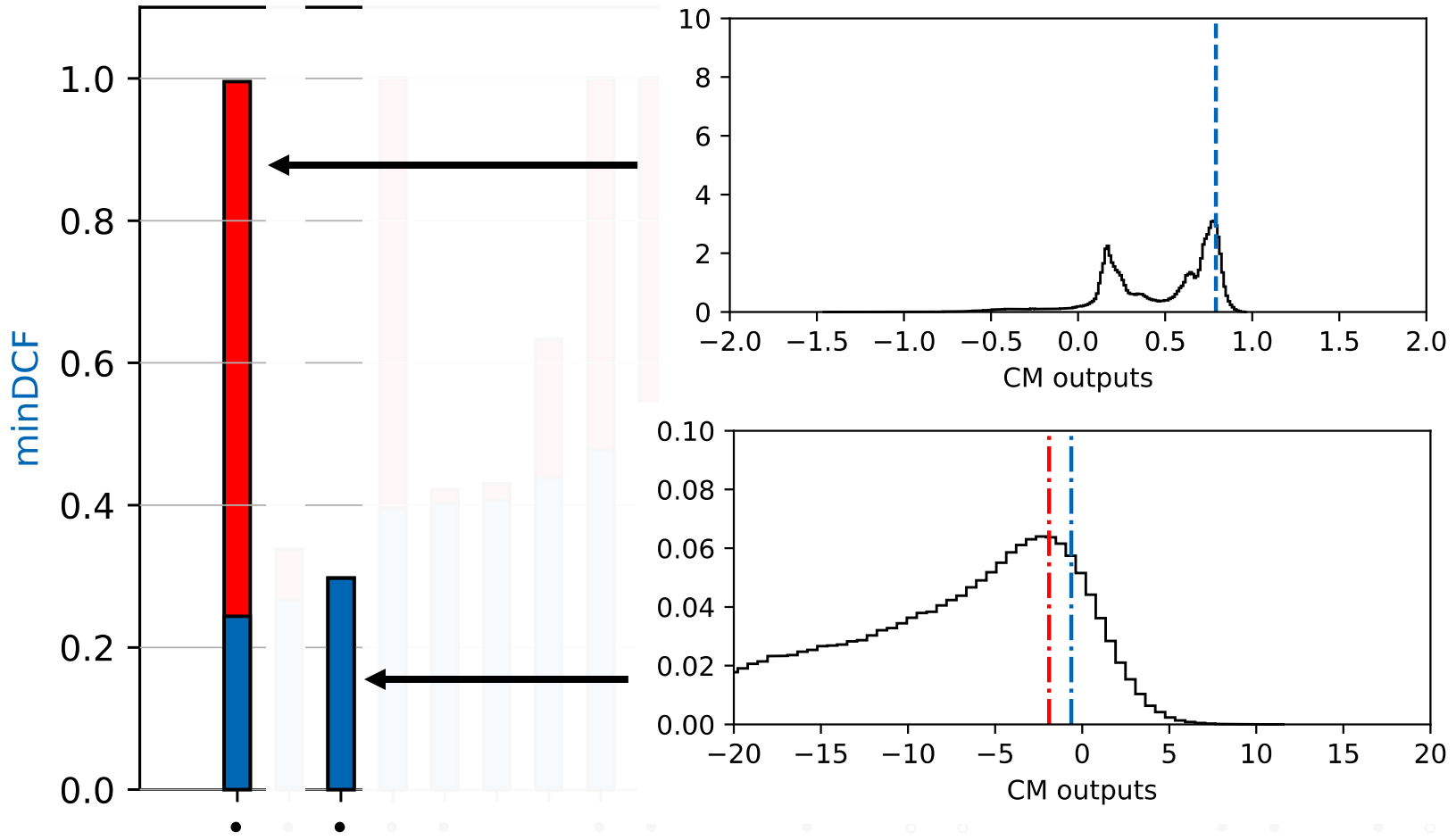
	Track 1 Closed	Track 1 Open
False Alarm Rate @ 1% Miss Rate High user convenience		
Miss Rate @ 1% False Alarm Rate High security		

Calibration



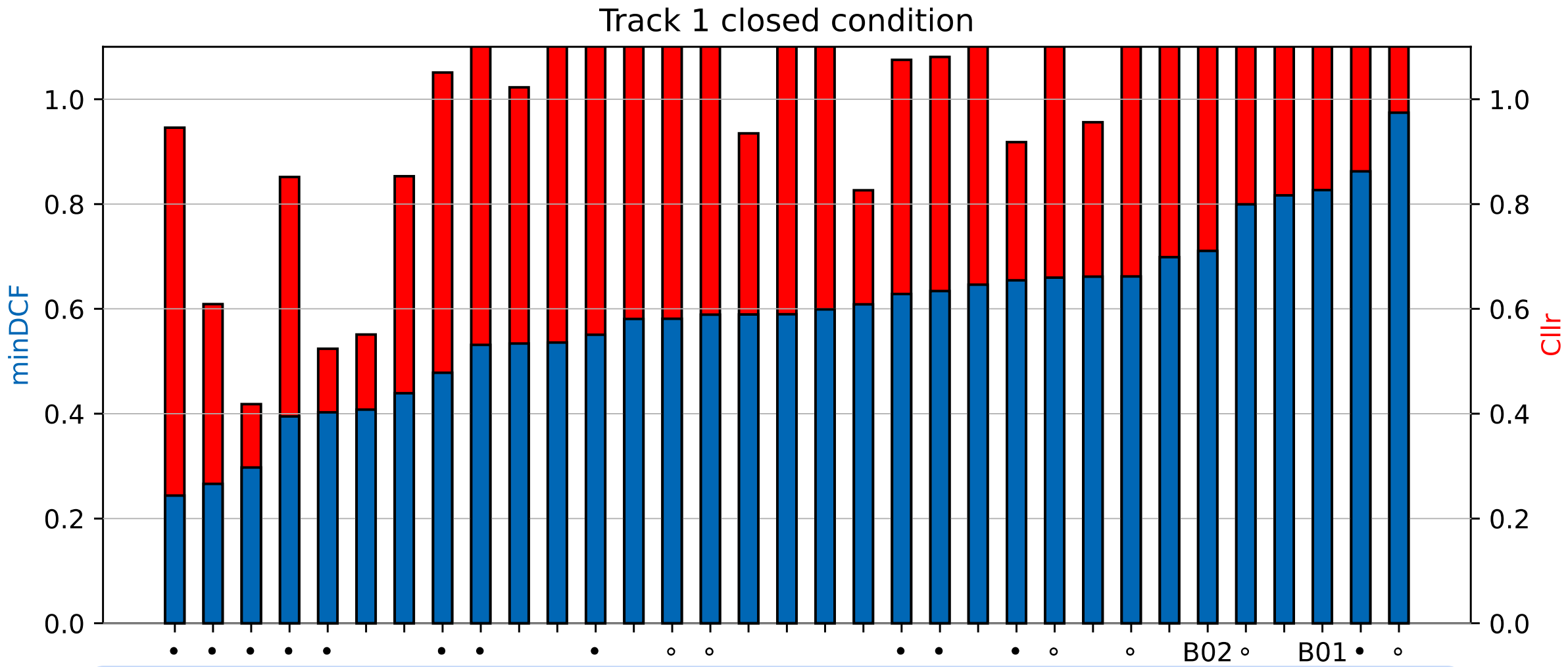
A few systems did quite well on calibration

Calibration



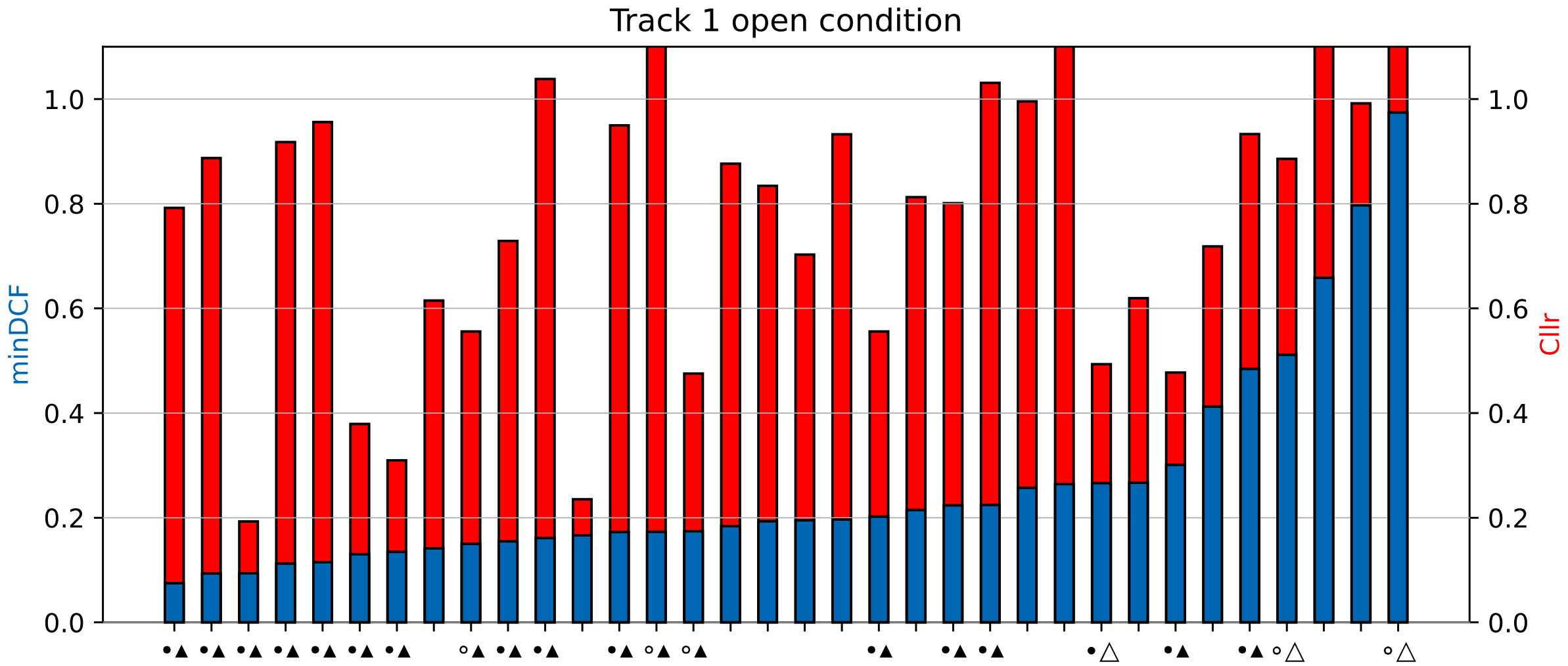
A few systems did quite well on calibration

Calibration



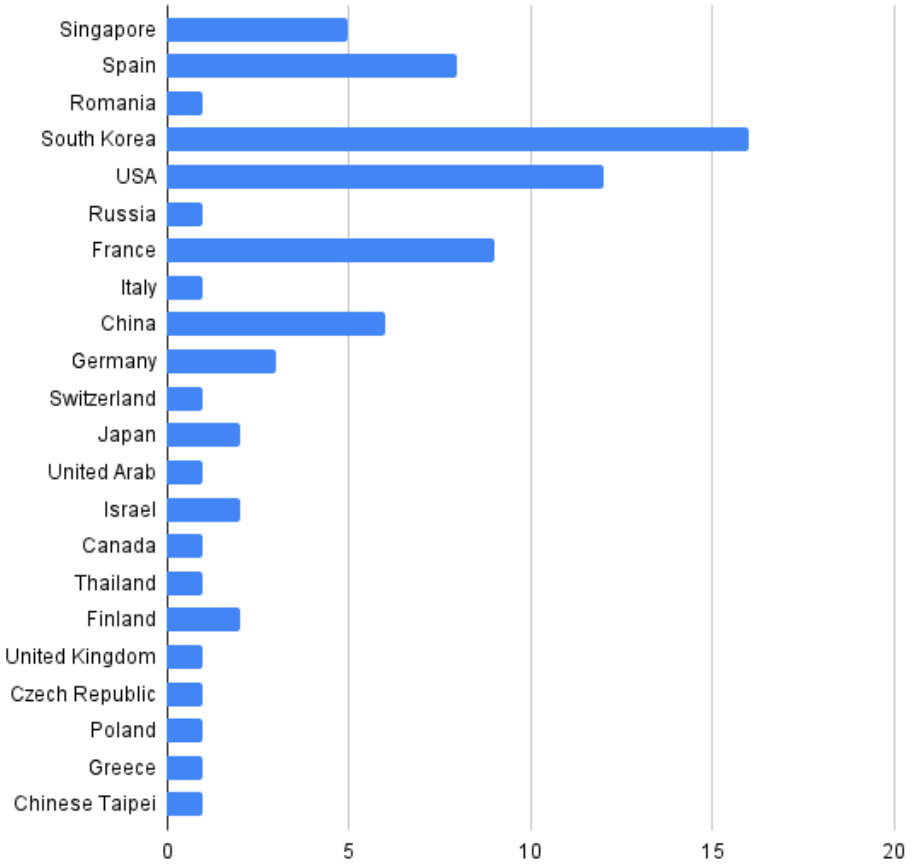
Systems with Cllr > 1 are poorly calibrated, and decisions are better made by omitting these systems ... Cllr = 1 are as good as a coin toss (on average) *(Nautsch 2019)*

Calibration

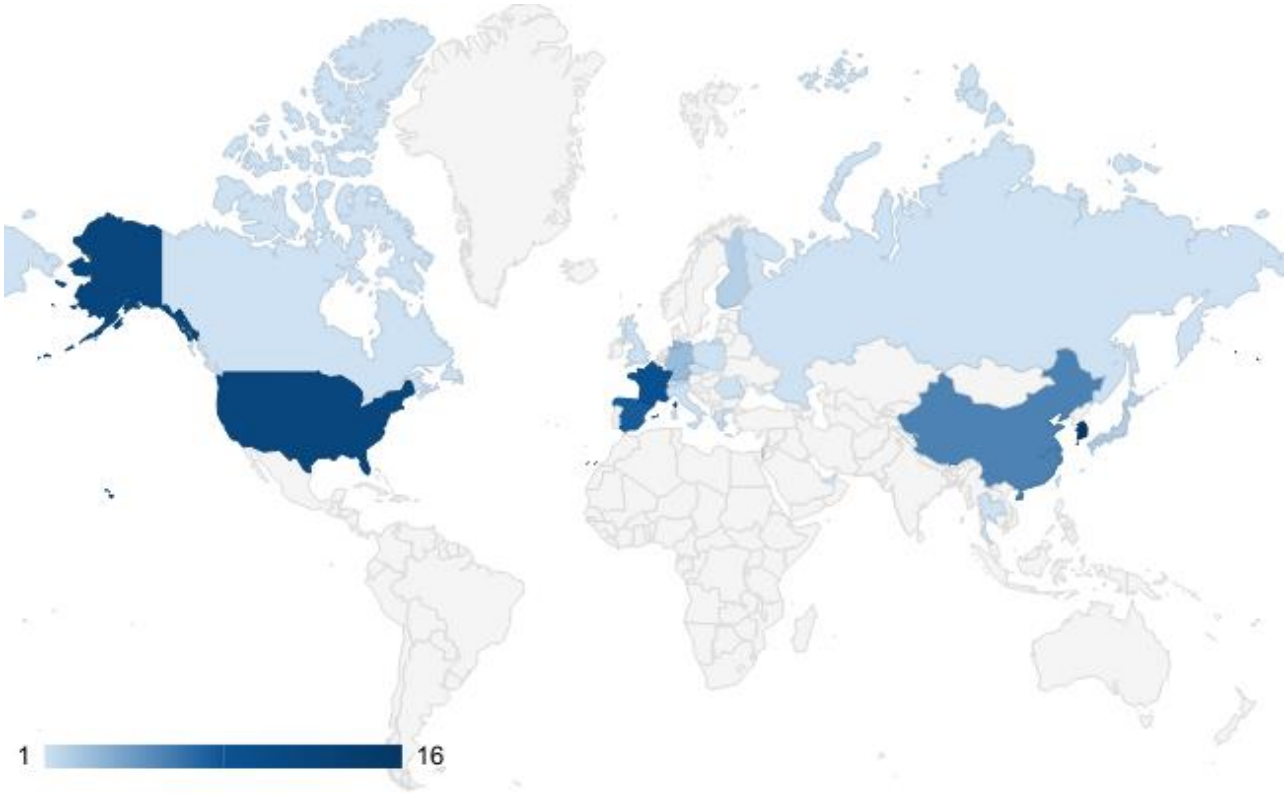


Appendix - misc

ASVspoof 5 workshop participants



Registrations by country



Acknowledgement

- We wish to thank:
  - **Phase-1 data contributors:** Cheng Gong, Tianjin University; Chengzhe Sun, Shuwei Hou, Siwei Lyu, University at Buffalo, State University of New York; Florian Lux, University of Stuttgart; Ge Zhu, Neil Zhang, Yongyi Zang, University of Rochester; Guo Hanjie and Liping Chen, University of Science and Technology of China; Hengcheng Kuo and Hung-yi Lee, National Taiwan University; Myeonghun Jeong, Seoul National University; Nicolas Muller, Fraunhofer AISEC; Sebastien Le Maguer, University of Helsinki; Soumi Maiti, Carnegie Mellon University; Yihan Wu, Renmin University of China; Yu Tsao, Academia Sinica; Vishwanath Pratap Singh, University of Eastern Finland; Wangyou Zhang, Shanghai Jiaotong University.
  - Challenge participants/authors
  - Reviewers
  - **A⋆ STAR** (Singapore) for sponsoring CodaLab platform,
  - **Pindrop** (USA) and **KLASS Engineering** (Singapore) for sponsoring the ASVspoof 2024 Workshop