

Emotion-Aware Speaker Anonymization with Zero-Shot TTS Foundation Models for VoicePrivacy 2026

Alef Iury S. Ferreira¹, Lucas H. Ueda², Lucas R. S. Gris¹,
Frederico S. de Oliveira³, Augusto S. da Rosa⁴, Alexandre C. Ferro Filho¹, Victor Moreno²

¹Federal University of Goiás, Brazil

²Universidade Estadual de Campinas (Unicamp), Brazil

³Federal University of Mato Grosso (UFMT), Brazil

⁴São Paulo State University (Unesp), Brazil



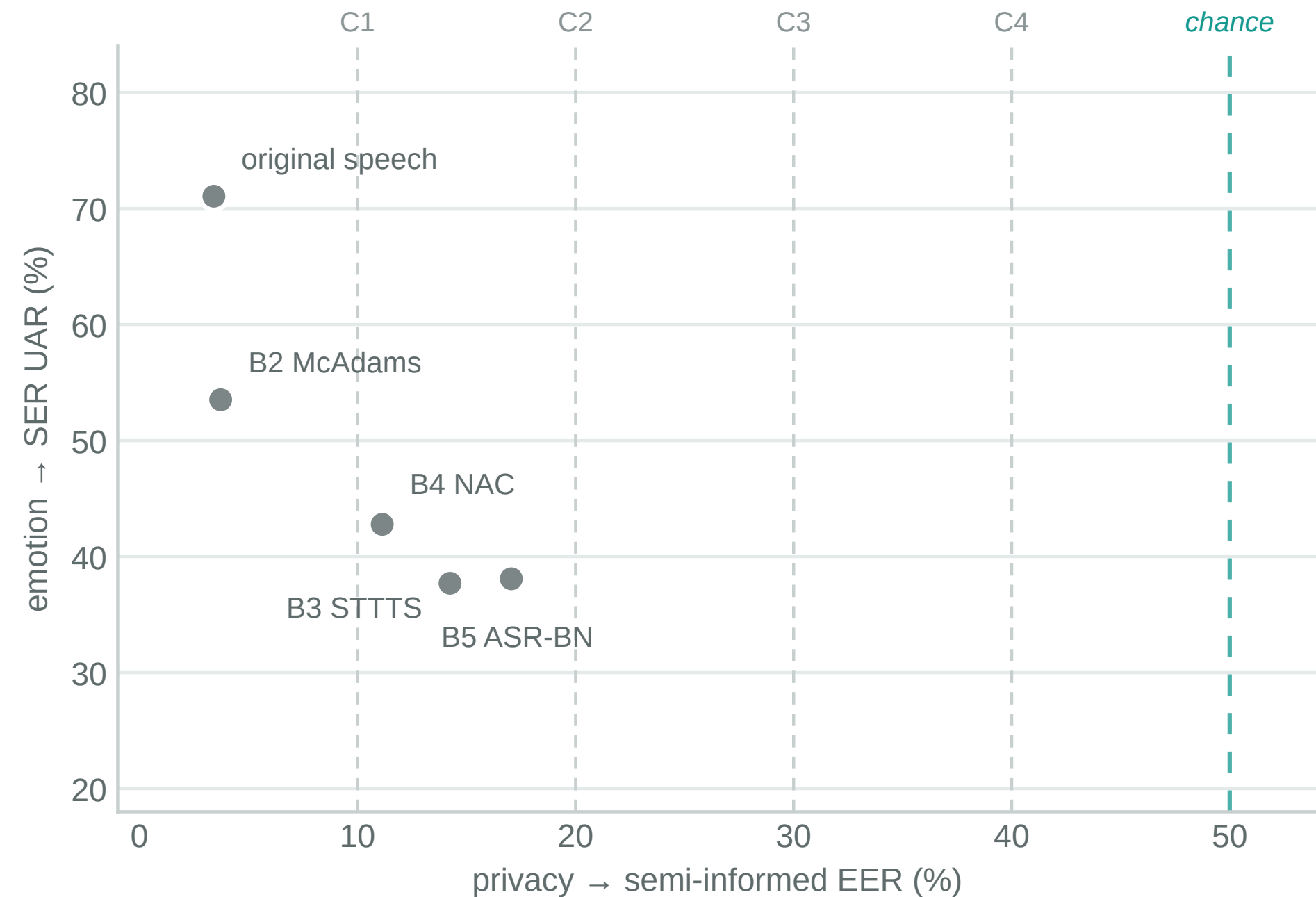
CEIA
CENTRO DE EXCELÊNCIA EM
INTELIGÊNCIA ARTIFICIAL



unesp

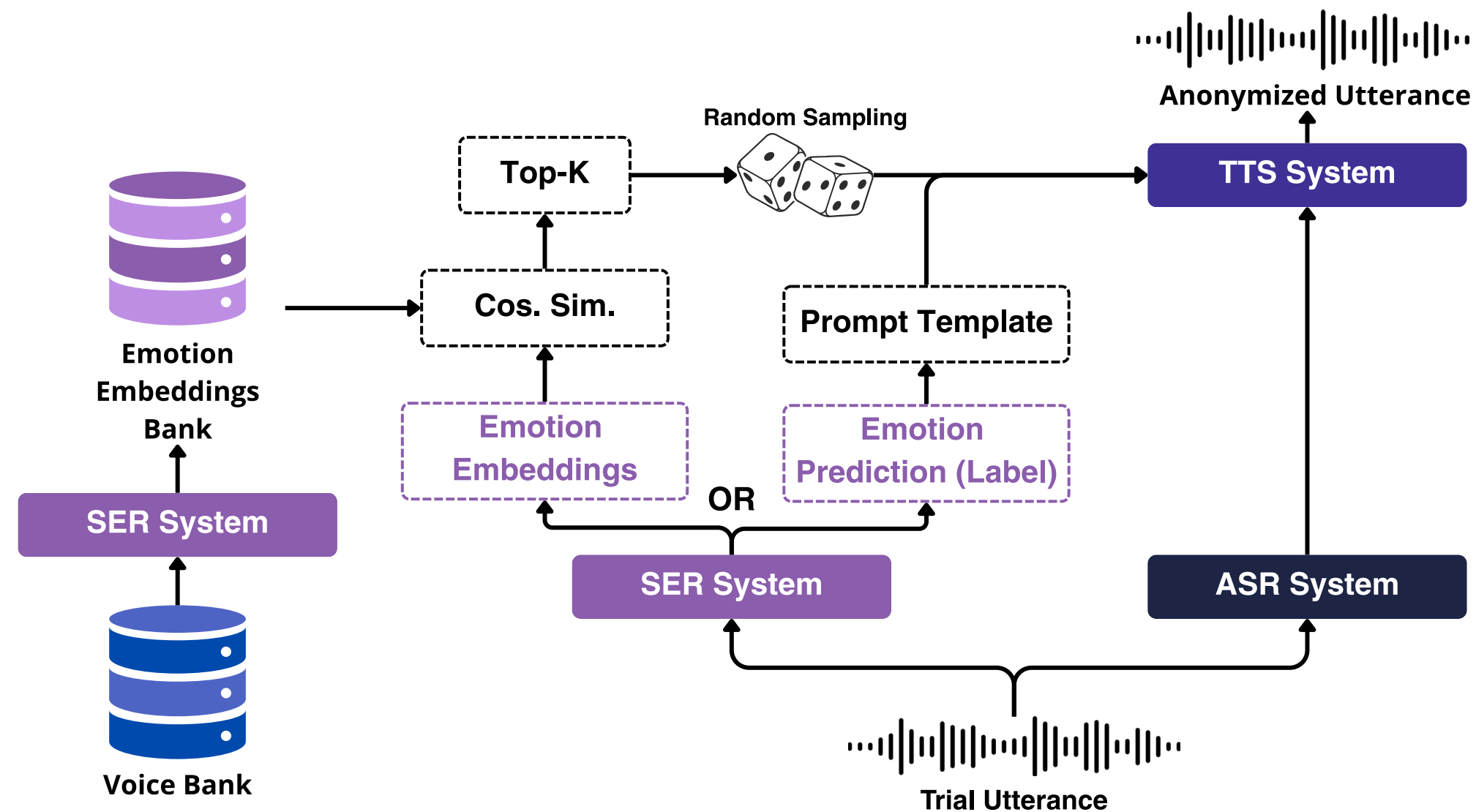


The Privacy–Emotion Trade-off



Main Goal: Move Right Without Falling Down

Training-Free ASR-TTS Pipeline



Reference-based Cloning

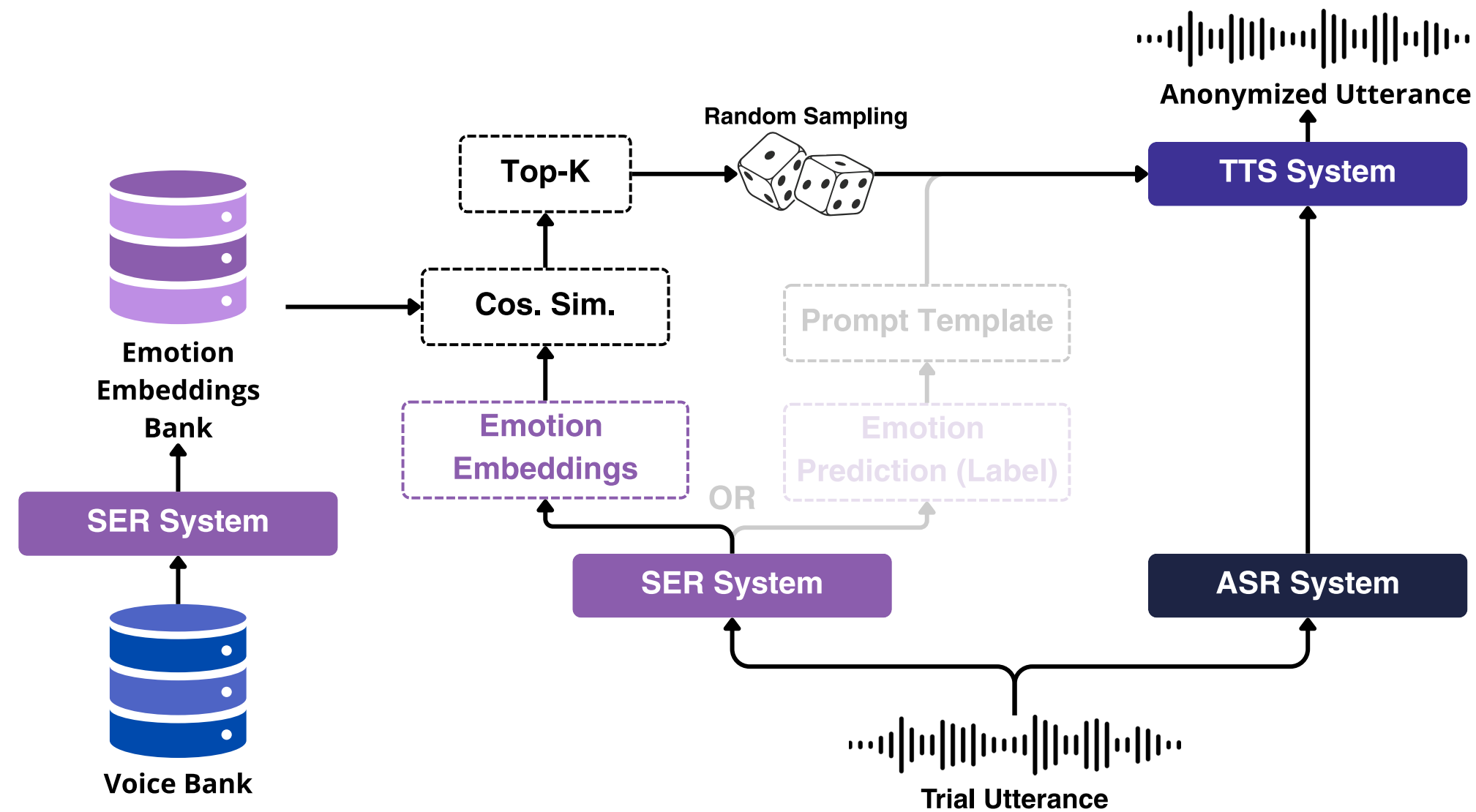
- Emotional Audio Reference

Prompt-based Voice Design

- Voice Description Only

"...in a clearly angry voice with firm emphasis."

Training-Free ASR-TTS Pipeline



Reference-based Cloning

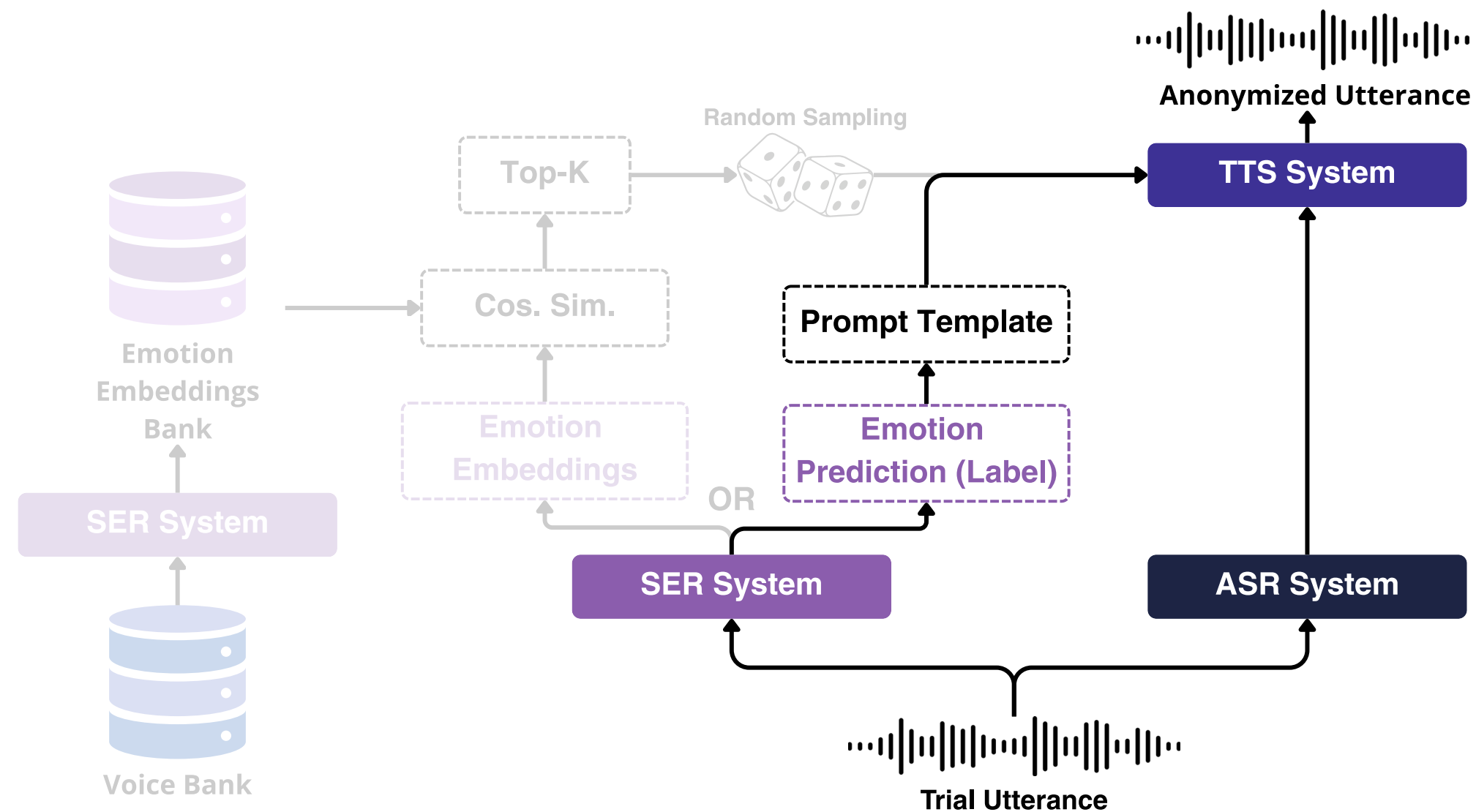
- Emotional Audio Reference

Prompt-based Voice Design

- Voice Description Only

"...in a clearly angry voice with firm emphasis."

Training-Free ASR-TTS Pipeline



Reference-based Cloning

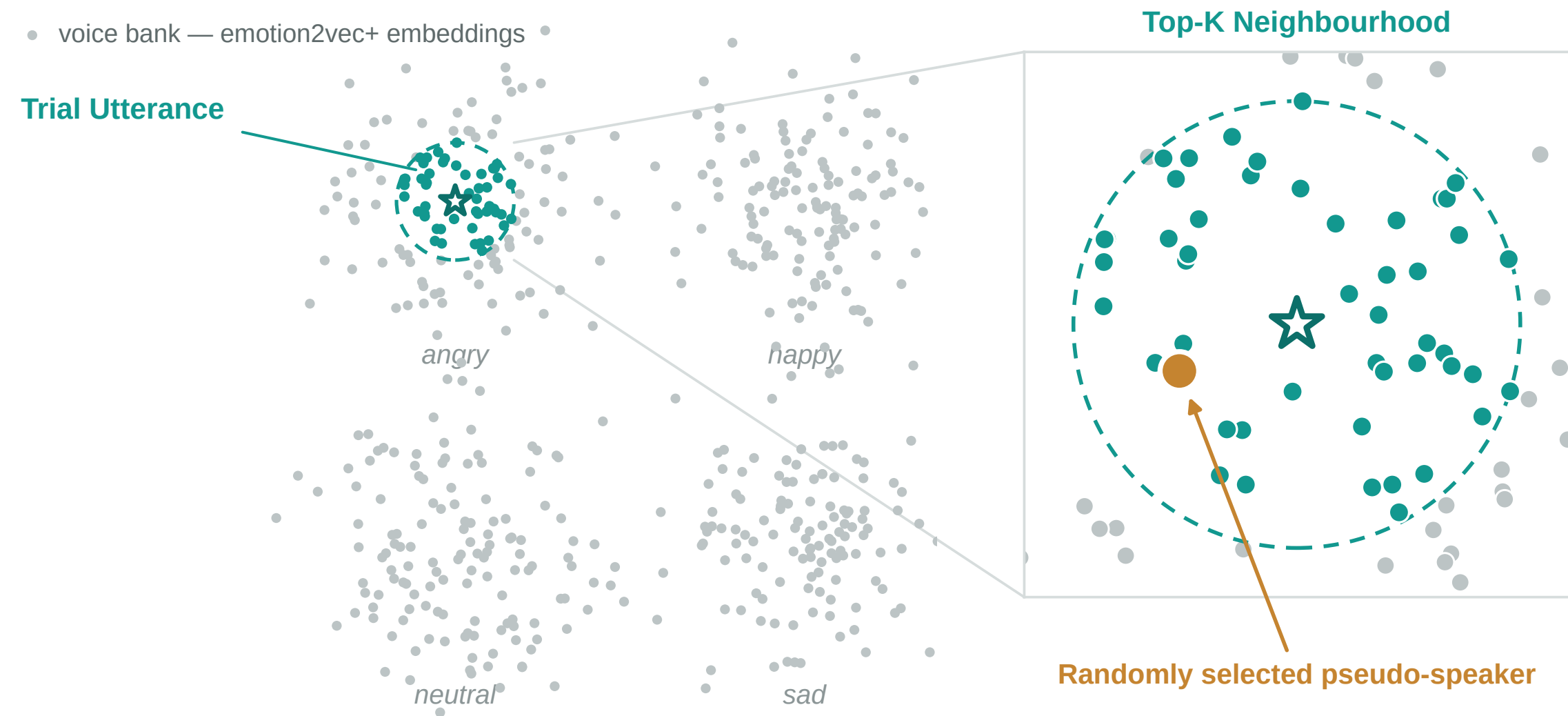
- Emotional Audio Reference

Prompt-based Voice Design

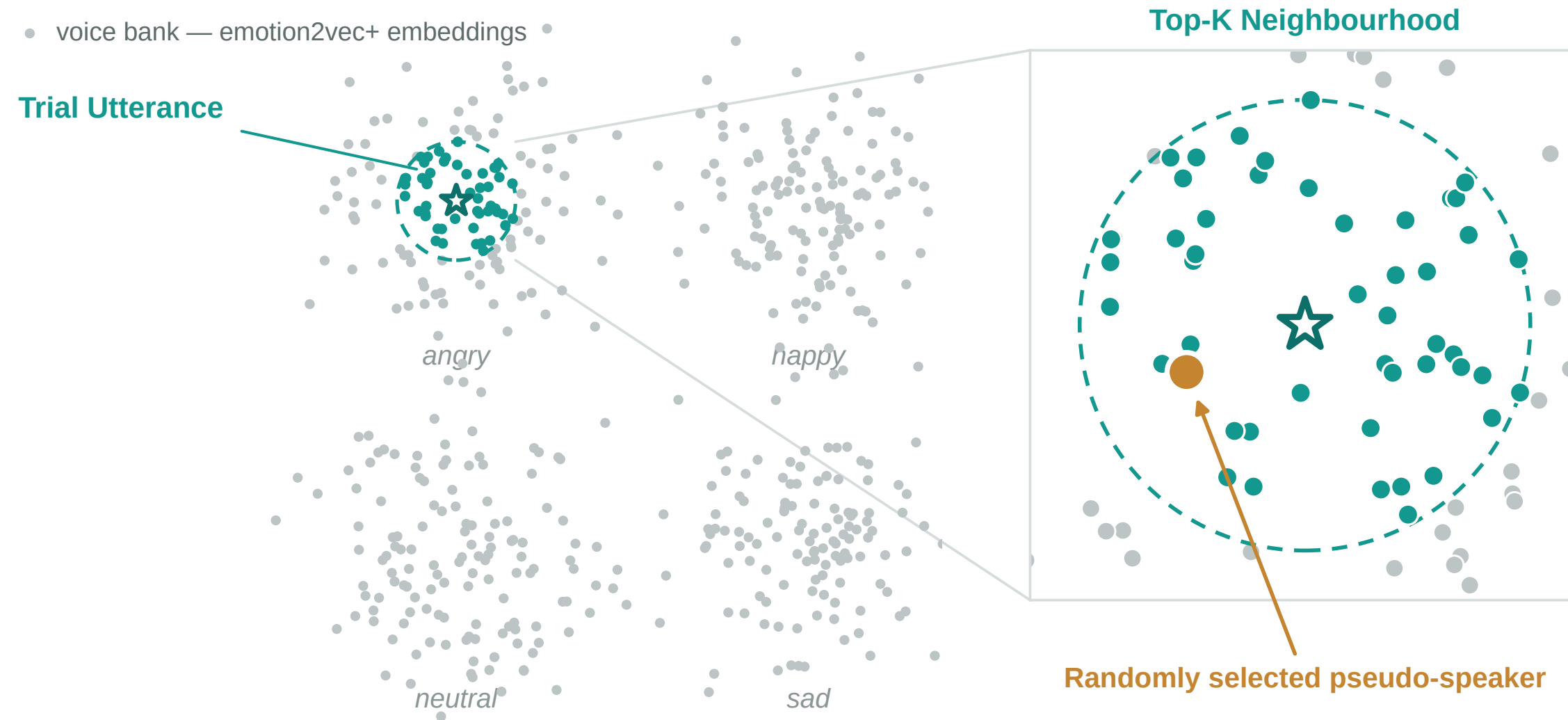
- Voice Description Only

"...in a clearly angry voice with firm emphasis."

Emotion-Aware Reference Selection



Emotion-Aware Reference Selection

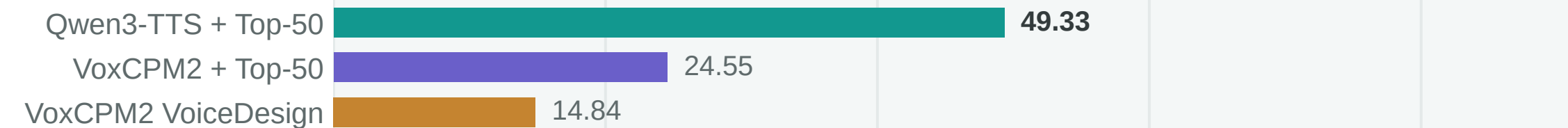


- Voice Bank

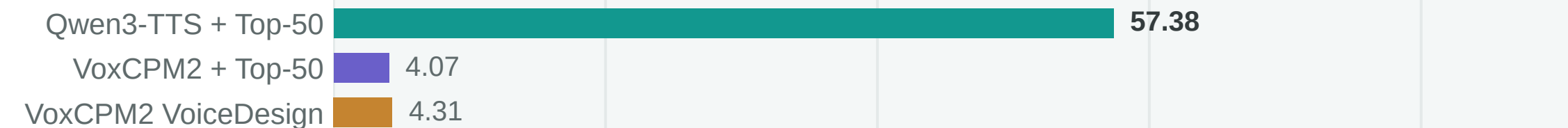
- Track 1: MSP-Podcast (Spont.) + **CAMEO** (Prep.) + **MEAD** (Prep.)
- Track 2: MSP-Podcast (EN) + **CAMEO** (EN + Other Languages)

No Single System Wins in Every Emotion

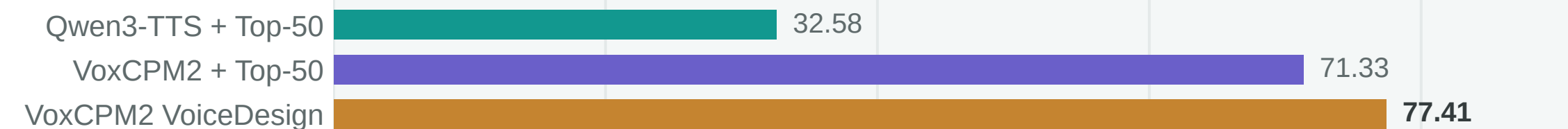
NEUTRAL



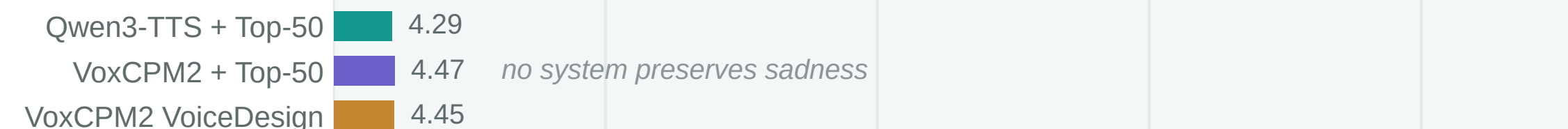
ANGRY



HAPPY



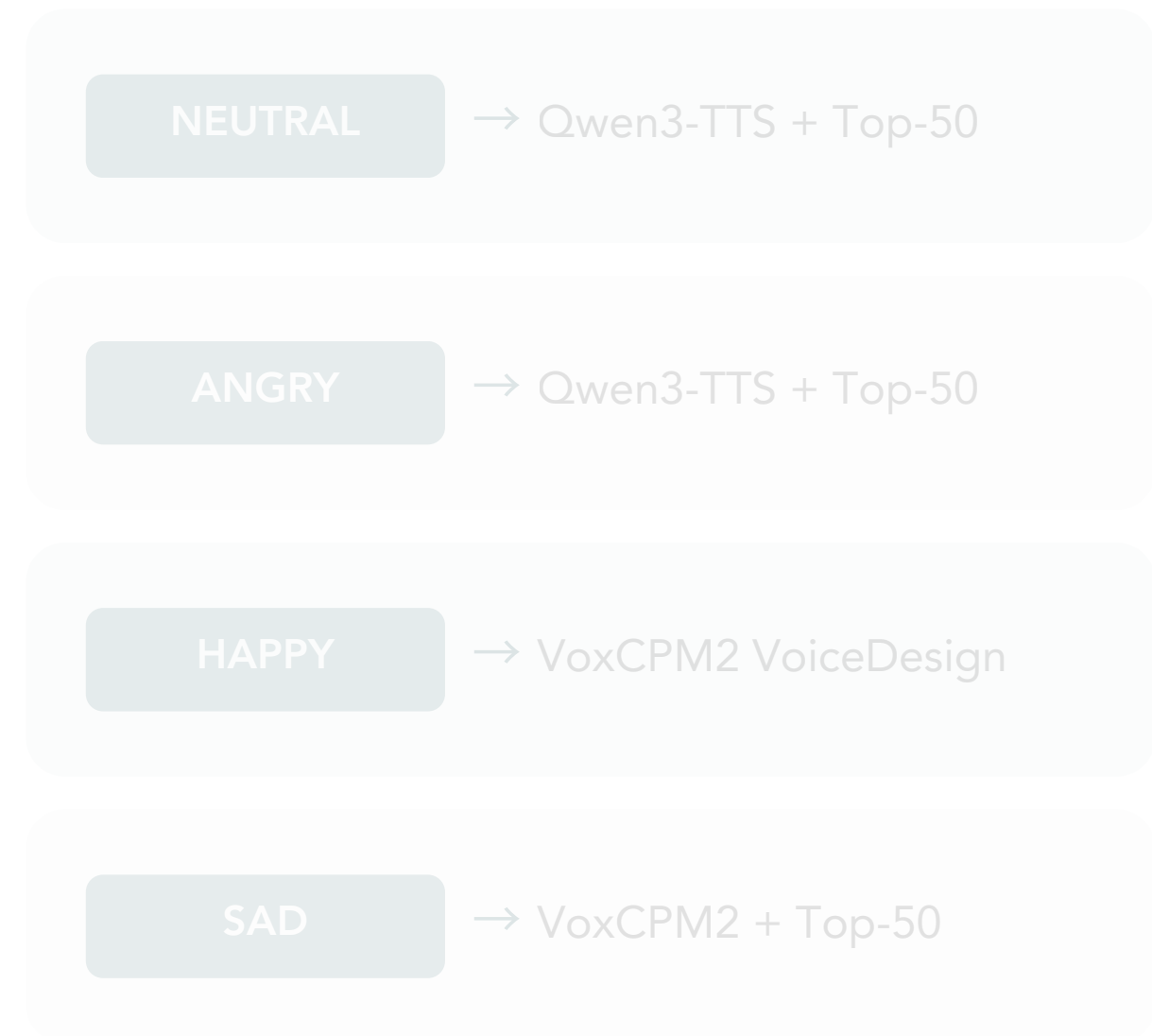
SAD



class accuracy on the development set (%)

bold value = best system for that emotion

Ensemble: Routing by Predicted Emotion



No Single System Wins in Every Emotion



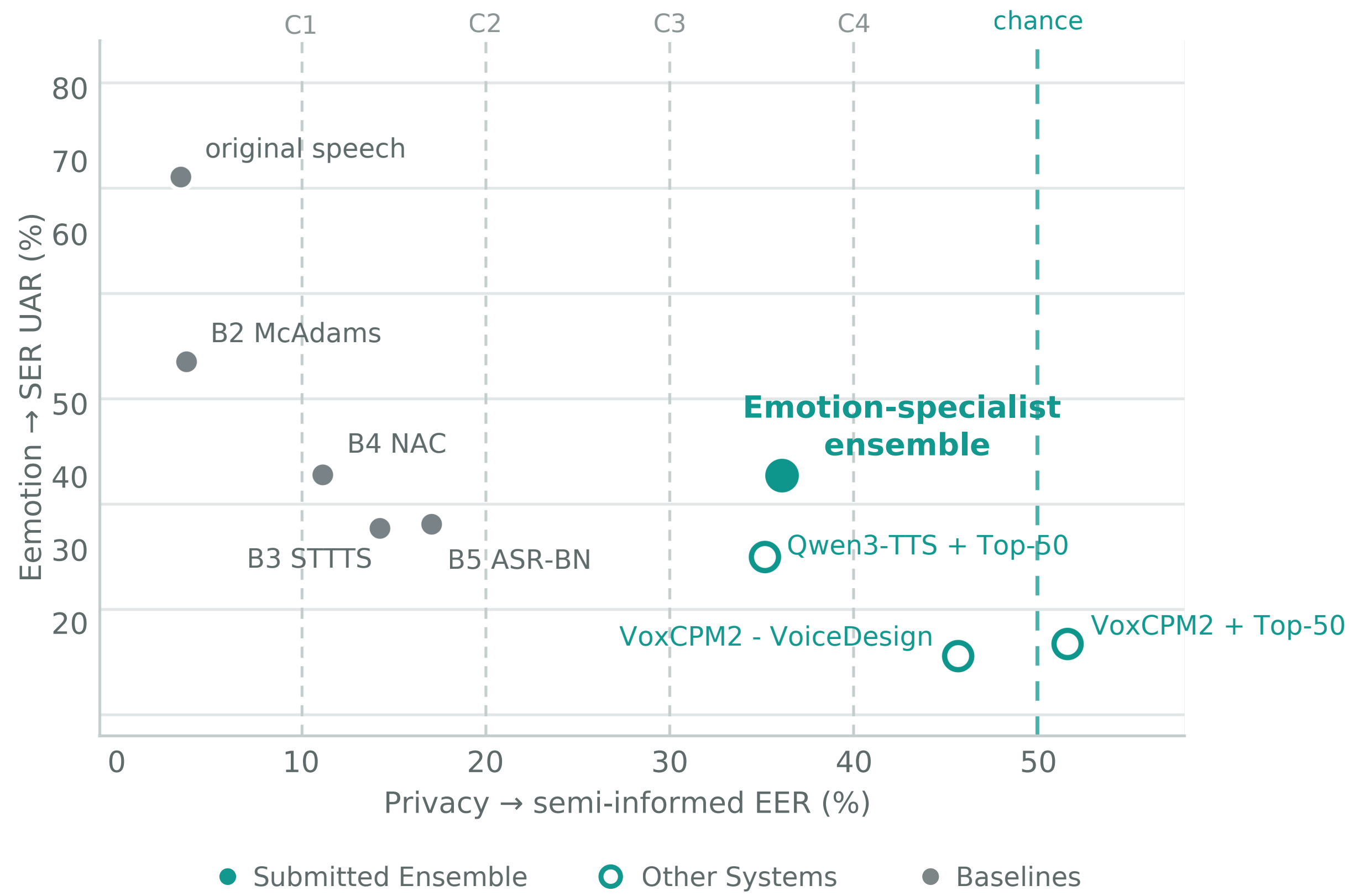
no system preserves sadness

bold value = best system for that emotion

Ensemble: Routing by Predicted Emotion



Track 1: English Results

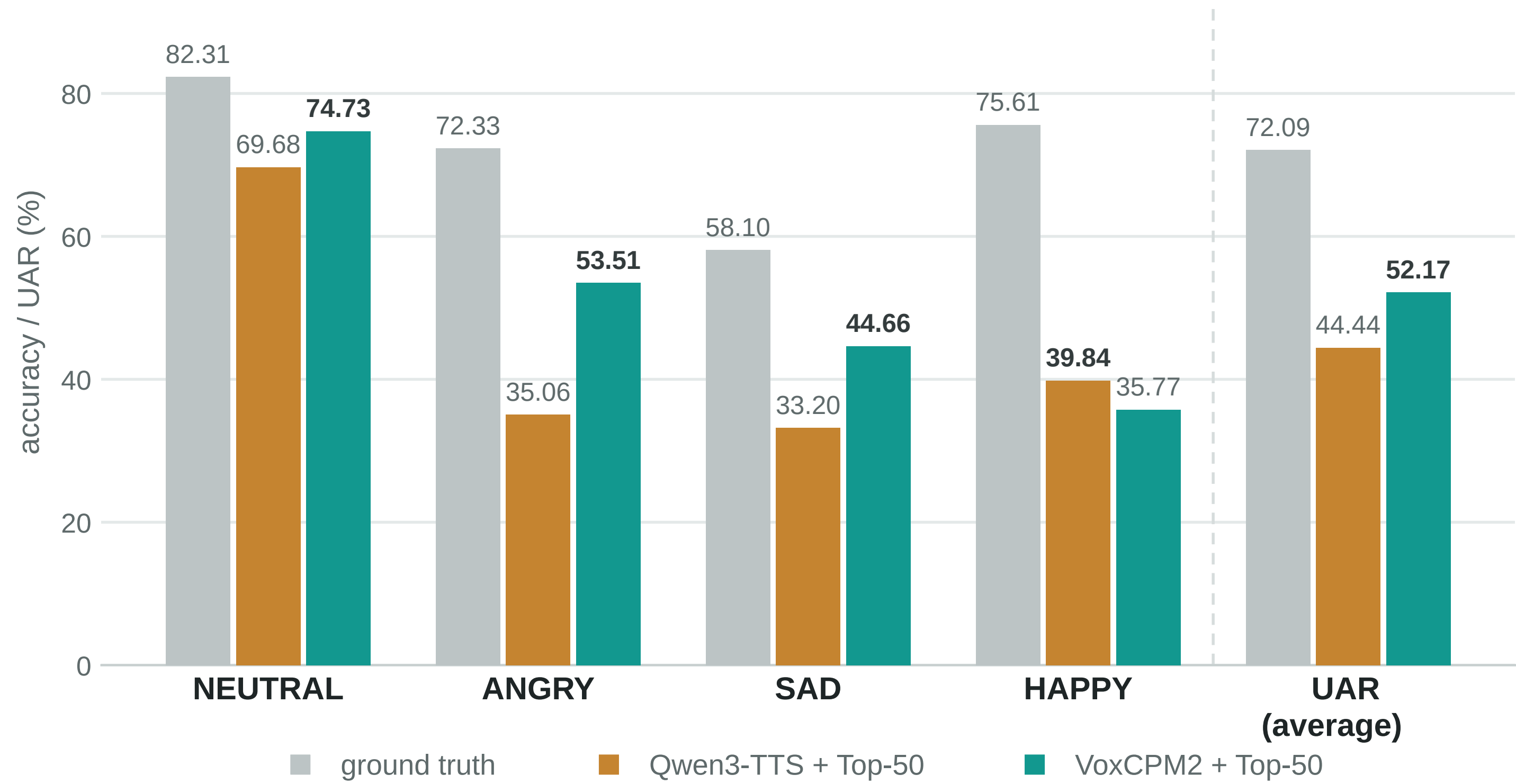


36.11%
semi-informed EER
2.1× the best baseline

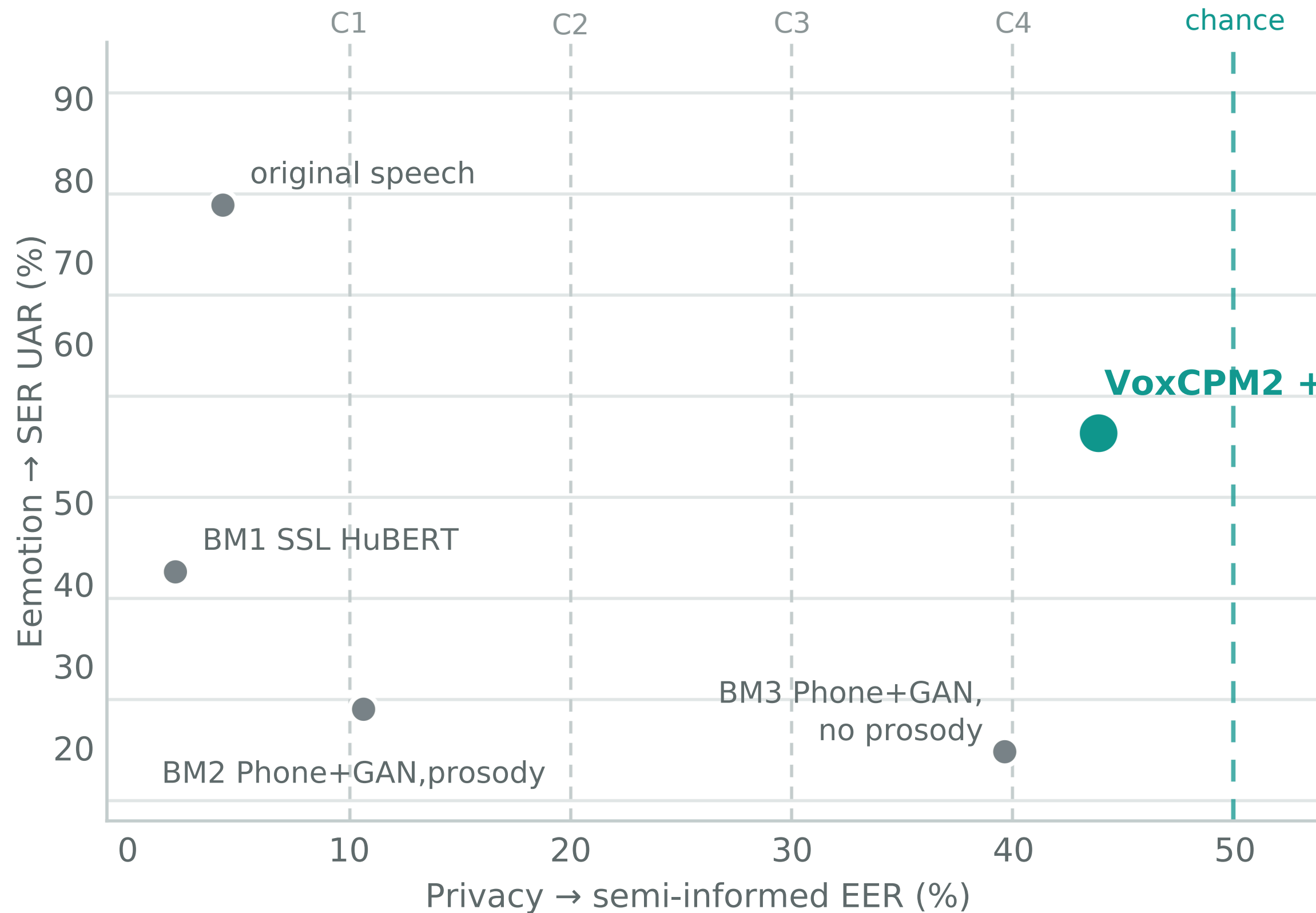
4.01%
WER

42.71%
UAR
Above every ASR-TTS baseline

Track 2: VoxCPM2 Better Across the Board



Track 2: Multilingual Results



43.90%

semi-informed EER
Highest average of the track

7.85%

WER

56.34%

SER UAR
Best of systems that reaches C4

8

Lessons Learned and Limitations

Lessons

Foundation TTS Models are Enough

Strong anonymization off the shelf

Emotion is the Bottleneck

No configuration came close to original speech

Ensemble When Systems Differ

Widens emotional coverage across backbones
Worth it when no single system dominates

Limitations

- **ASR–TTS Pipeline**

- Possible error propagation
- Exact prosody and timing are gone

- **Sad Emotion Could Not be Modeled**

- Bad synthesis representation of low-arousal speech

• • • • • • • • • •
• • • • • • • • • •
• • • • • • • • • •
• • • • • • • • • •
• • • • • • • • • •

Thank You!



CEIA
CENTRO DE EXCELÊNCIA EM
INTELIGÊNCIA ARTIFICIAL



OKCIT
CENTRO DE COMPETÊNCIA EMBRAPII
EM TECNOLOGIAS IMERSIVAS



UFG
UNIVERSIDADE
FEDERAL DE GOIÁS



unesp



UFMT

Backup Slides

Track 1 System Comparison

System	SER UAR ↑	ASV EER ↑	WER ↓	Semi EER ↑	Trade-off ↑
Qwen3-TTS + Random	31.50	<u>47.94</u>	5.23	34.60	32.98
Qwen3-TTS + Top-1	33.33	47.73	8.76	34.30	33.81
Qwen3-TTS + Top-50	<u>34.98</u>	47.34	4.24	35.18	<u>35.08</u>
VoxCPM2 + Top-50	26.72	49.55	5.23	51.64	35.22
VoxCPM2 + VoiceDesign	25.57	47.45	2.67	<u>45.69</u>	32.79
Emotion-specialist ensemble	42.71	44.53	<u>4.01</u>	36.91	39.60

Trade-off = harmonic mean of SER UAR and semi-informed EER.
The ensemble was submitted because it maximizes this balance, not raw EER.

Track 2 System Comparison

	EN	DE	ES	FR	avg
EER ↑ (ours)	44.81	42.35	43.87	44.55	43.90
EER ↑ (BM3)	20.40	46.05	45.92	46.25	39.65
WER ↓ (ours)	7.27	10.75	5.38	8.00	7.85

UAR ↑	SAD	ANG	HAP	NEU	avg
Original	73.50	83.40	68.08	90.64	78.90
Qwen3 + Top-50	46.15	39.25	36.92	74.89	49.30
VoxCPM2 + Top-50	49.15	55.85	44.62	75.75	56.34

Perspectives for the Next Challenge

Attacker Challenge

Reproducible attacks, including content-based ones

Beyond Four Emotions

Arousal / valence and expressiveness, not only discrete

Data & Languages

More emotional multilingual data