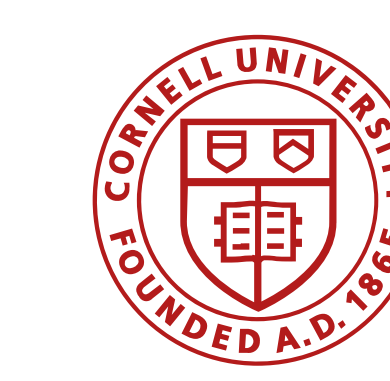


When Audio-Language Models **Fail to Leverage Multimodal Context** for Dysarthric Speech Recognition

Benchmarking, diagnosing, and repairing the clinical-context failure

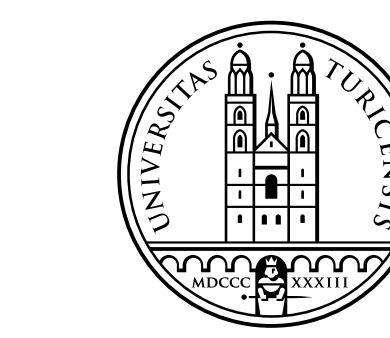
Pehuén Moure^{1*} · Niclas Pokel^{1*} · Bilal Bounajma² · Yingqiang Gao³ · Roman Böhlinger¹ · Longbiao Cheng¹ · Shih-Chii Liu¹

¹ Institute of Neuroinformatics, University of Zürich & ETH Zürich · ² Dept. of Mathematics, ETH Zürich · ³ Dept. of Computational Linguistics, University of Zürich · * equal contribution · P.M. is now at Cornell Tech



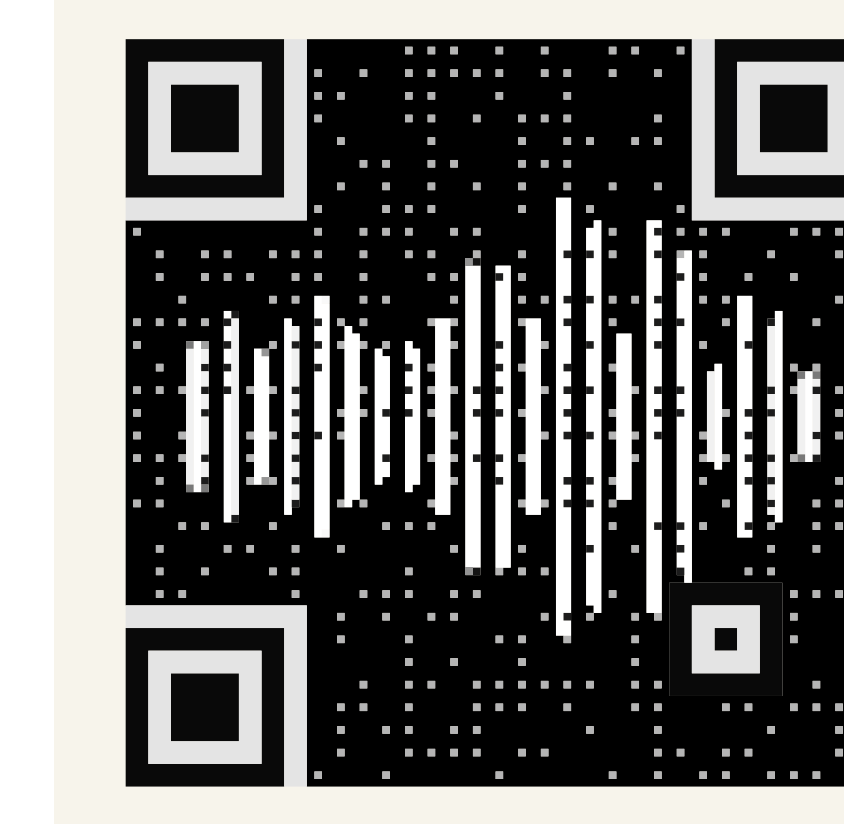
**CORNELL
TECH**

COLM



University of
Zurich ^{UZH}

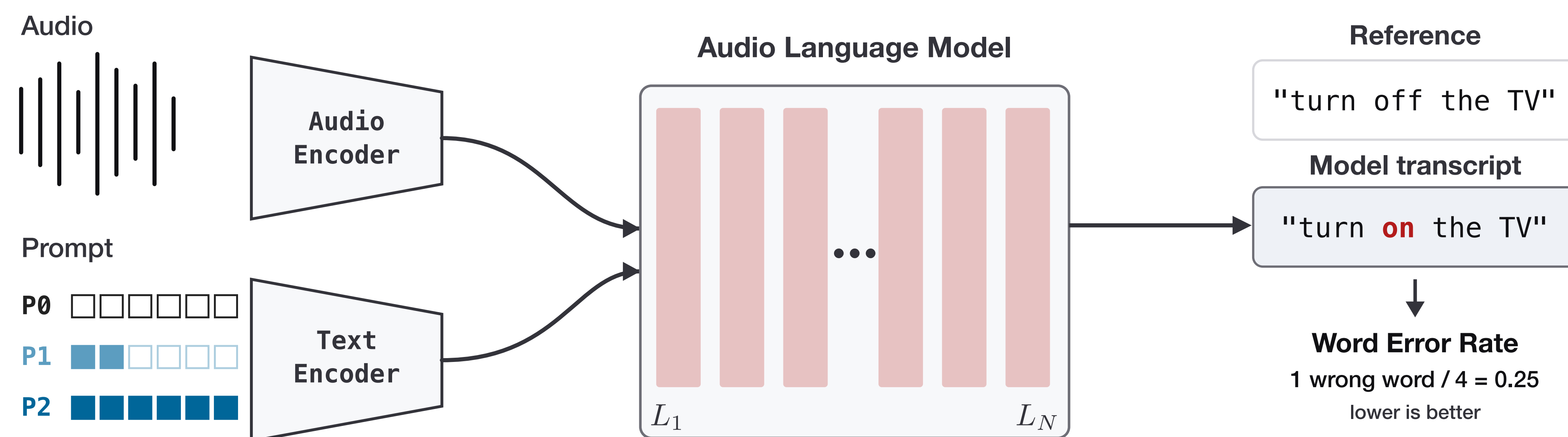
ETH zürich



Paper
arXiv:2605.02782

The Experimental Setup

Does telling a frozen audio-language model the speaker's clinical profile help it transcribe?



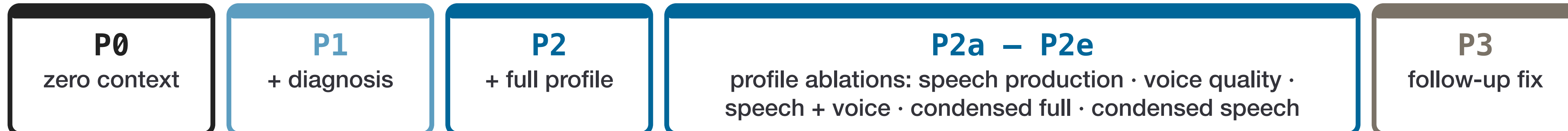
Three of the prompts, in full

P0 Audio only · ~11 tokens
Transcribe the audio in English.

P1 + Diagnosis · ~70 tokens
The speaker has Cerebral Palsy, which often causes imprecise consonants, distorted vowels, and irregular rhythm.

P2 + Clinical profile · ~160 tokens
[...P1...] Clinical speech profile (1–7): imprecise consonants 4/7 · monopitch 4/7 · nasalness 5/7.

The full sweep: nine prompt conditions (the x-axis of every sweep plot)

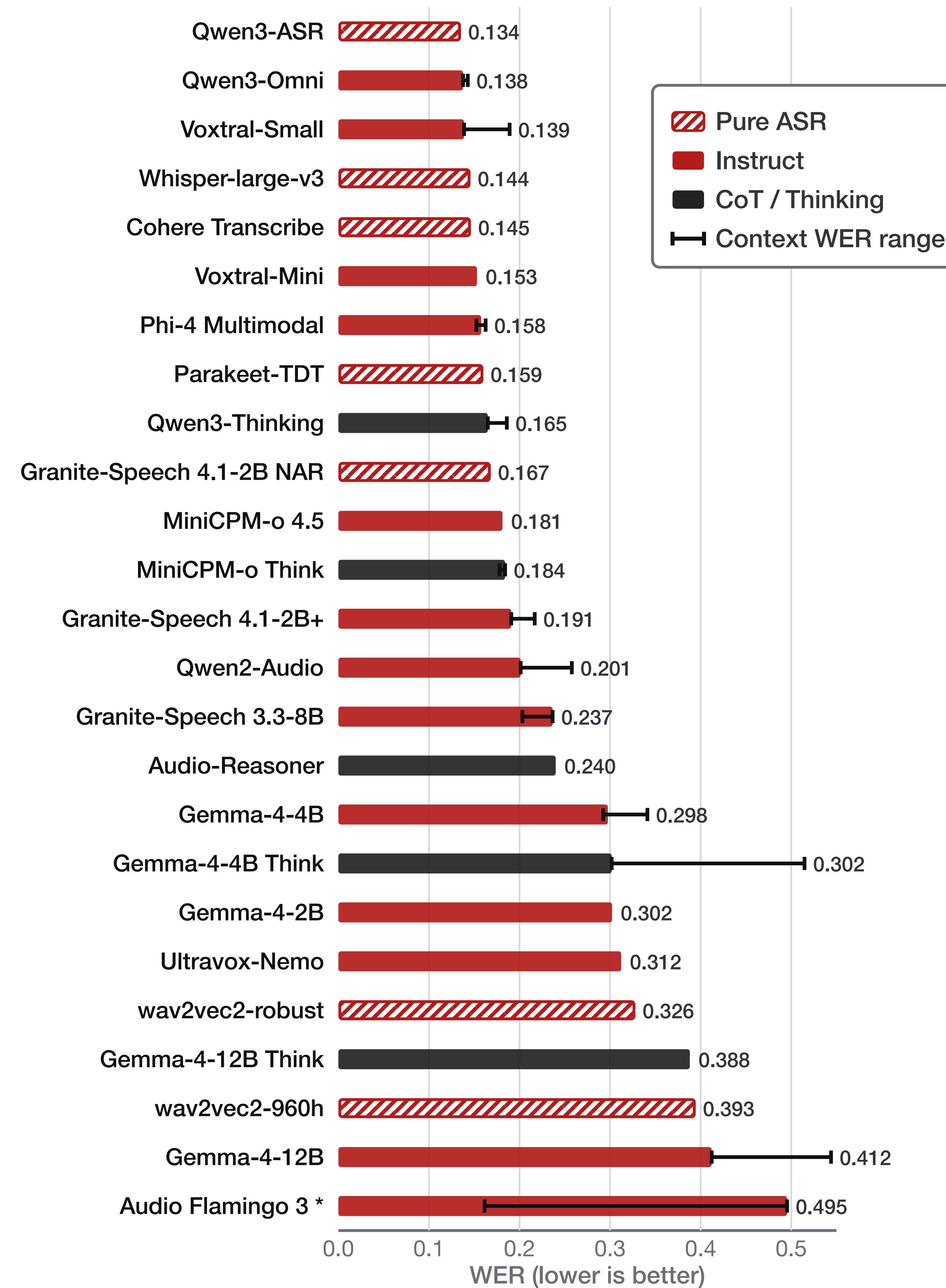


The data: 11,218 clinician-rated SAP utterances from 437 speakers

Parkinson's disease	67%	7,504 utt. · 249 speakers
Cerebral palsy	16%	1,762 · 97
ALS	14%	1,594 · 70
Down syndrome	3%	358 · 21

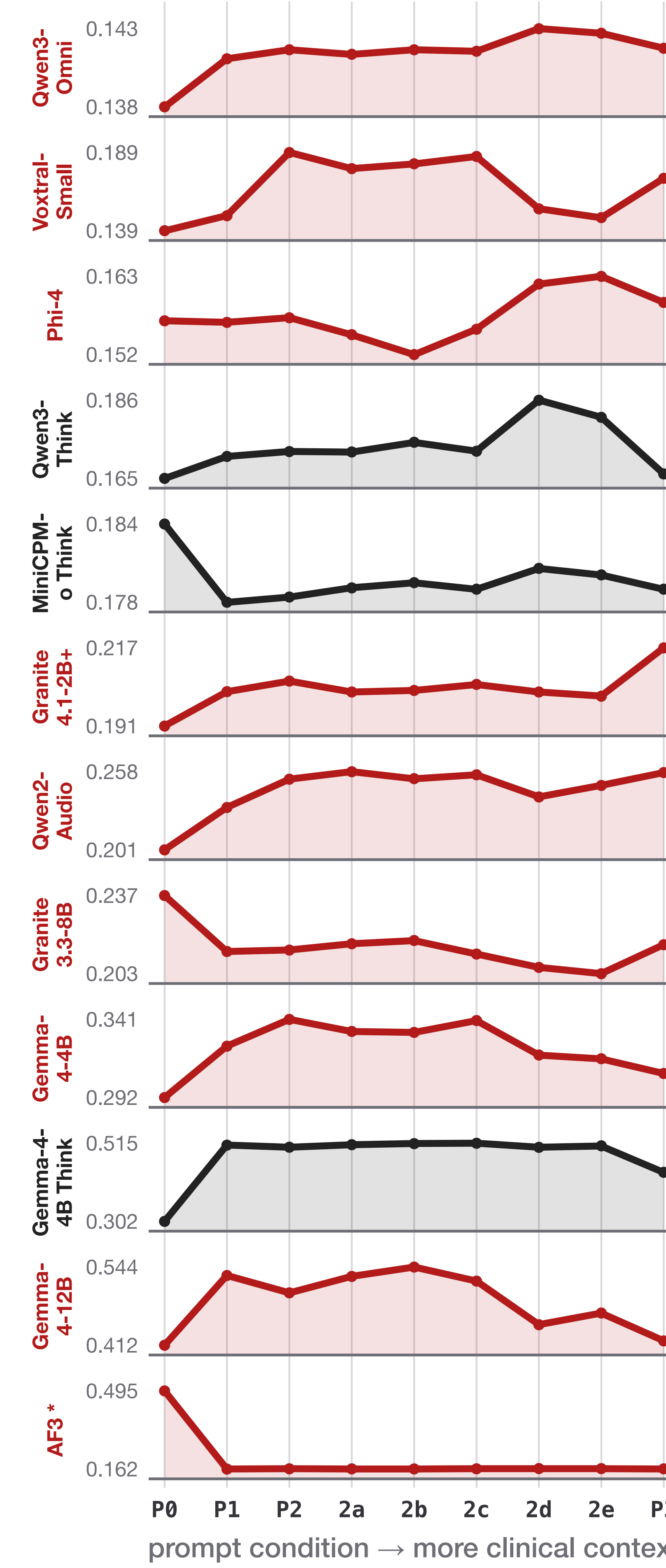
Across Audio-LMs, Clinical Context Rarely Helps

Zero-Shot WER (25 Systems)



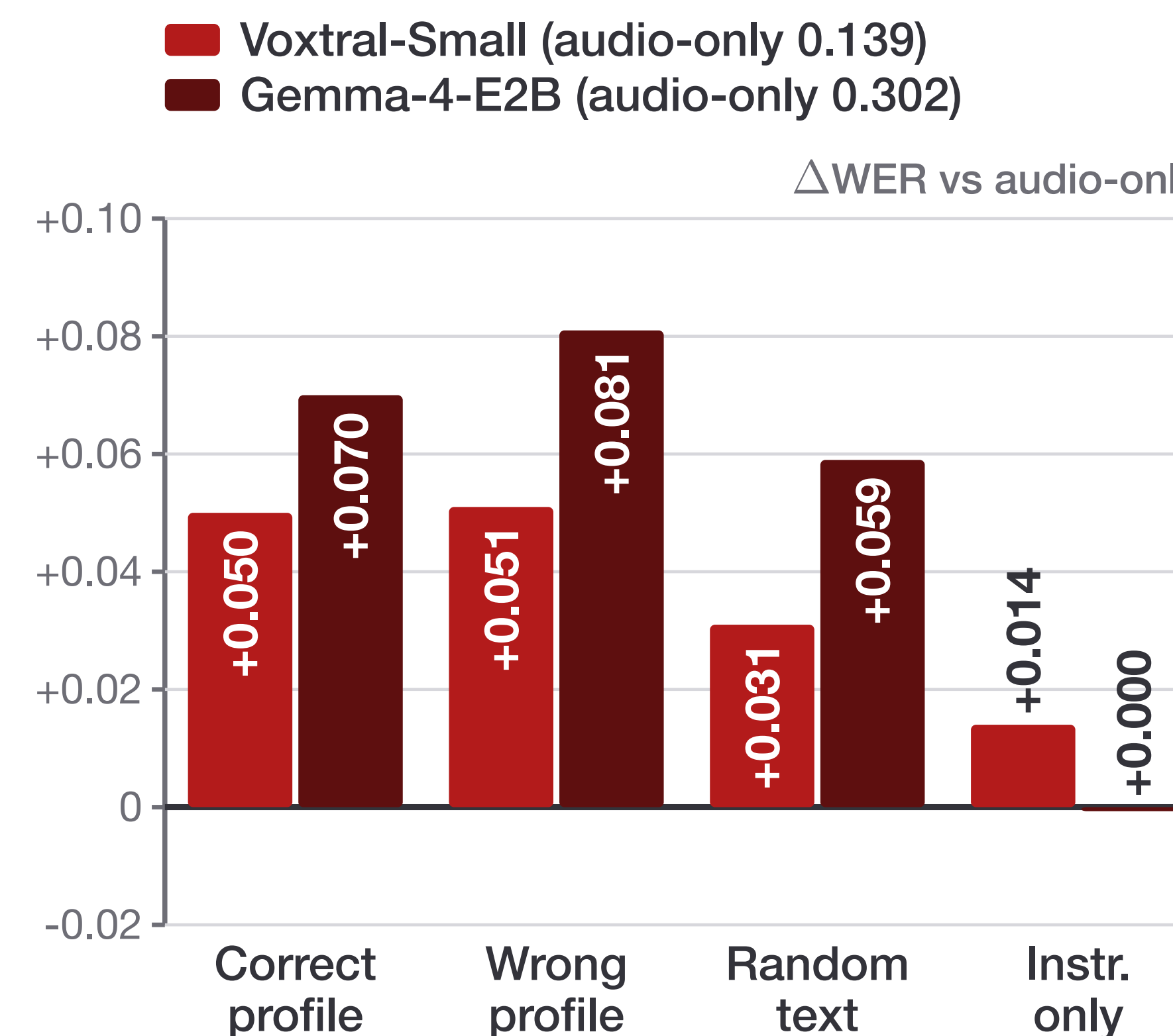
Zero-shot ranking; black markers: each model's WER range across prompts. Size is a weak predictor ($R^2 = 0.07$).

WER by Prompt Condition



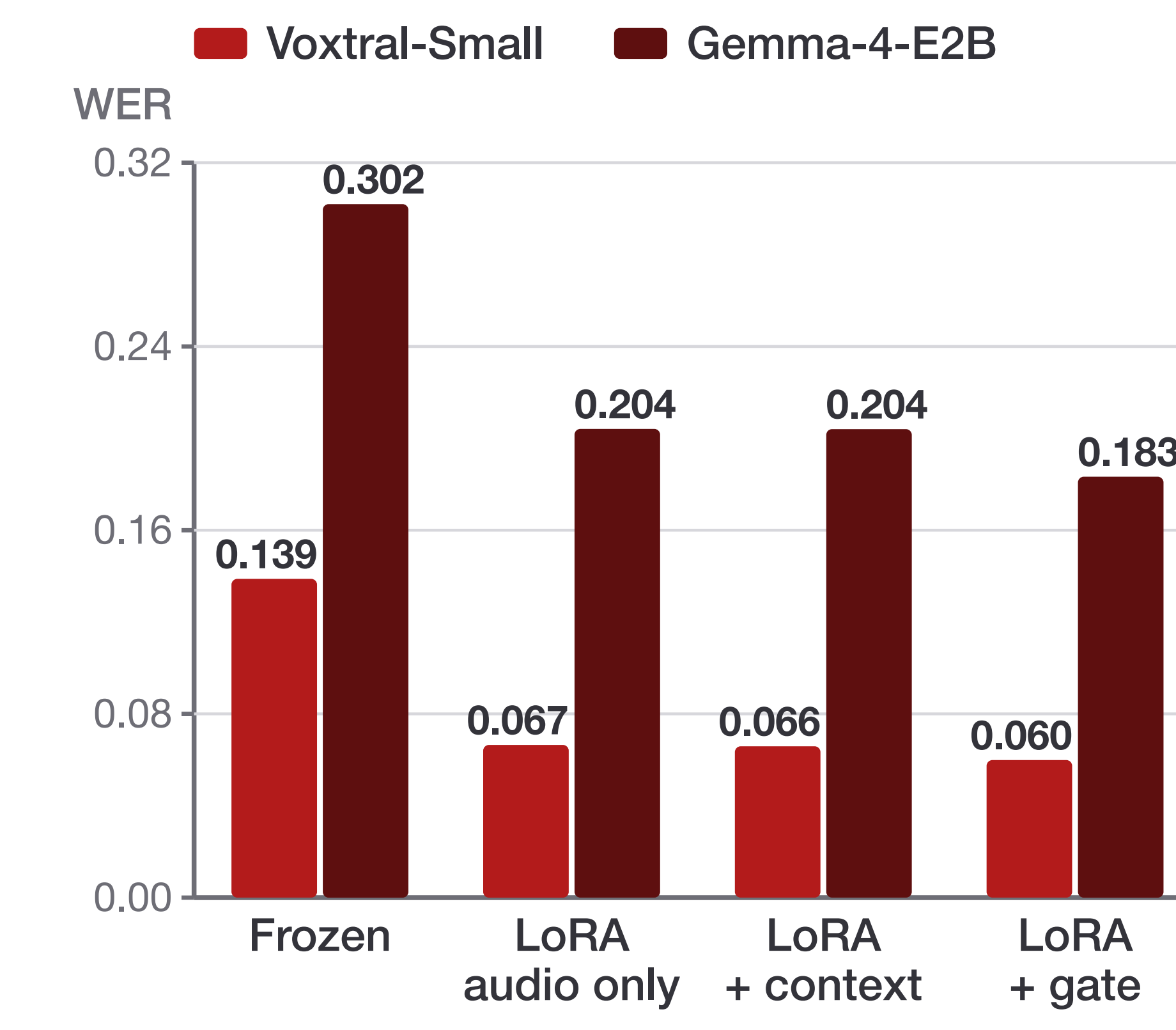
Own scale per model, one shared axis: almost all stay flat or degrade.

Random $\approx$ Clinical Content



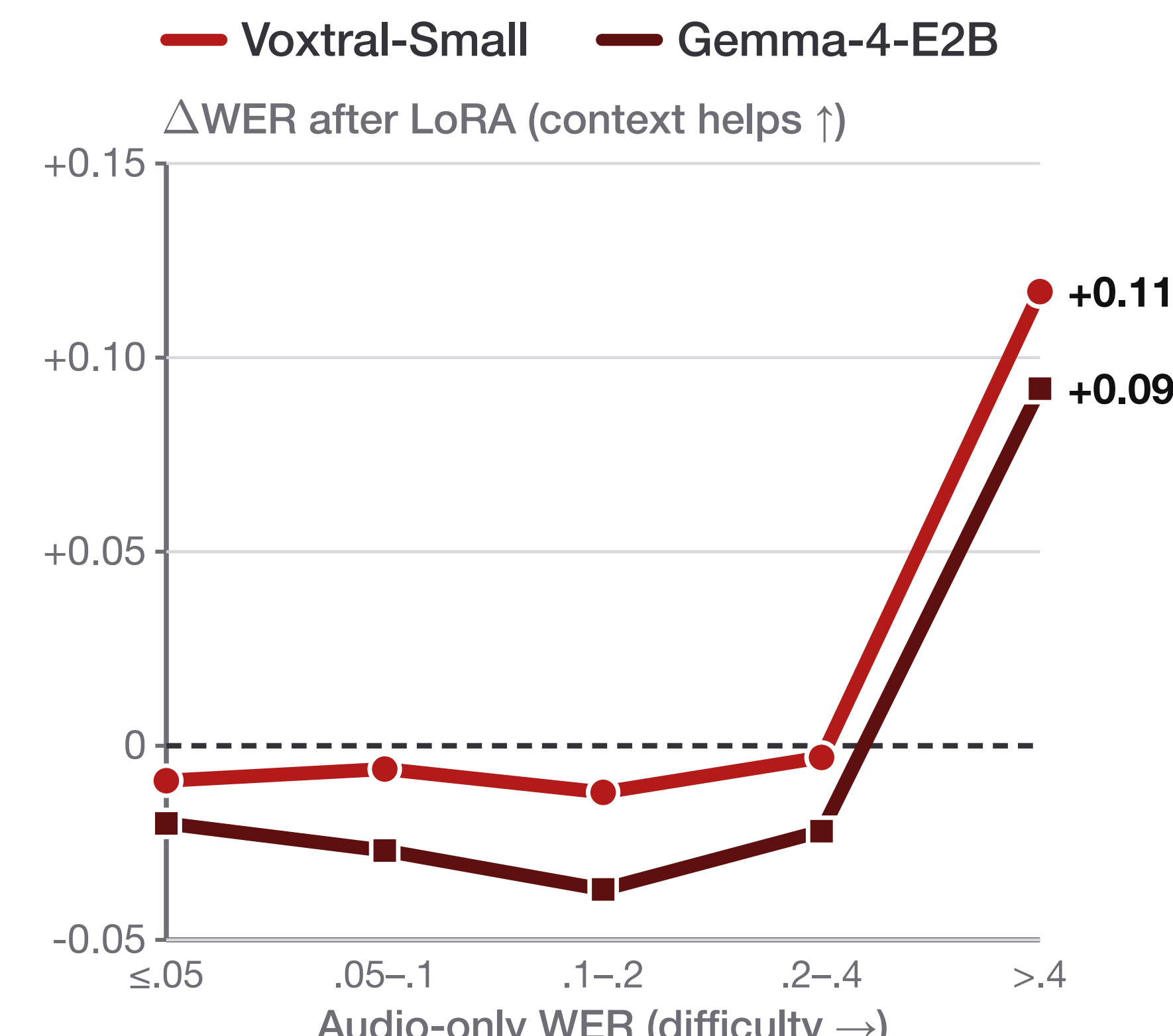
- Wrong profile $\approx$ correct profile
- Random text hurts almost as much
- Instruction-only $\approx$ audio-only

LoRA Removes the Failure



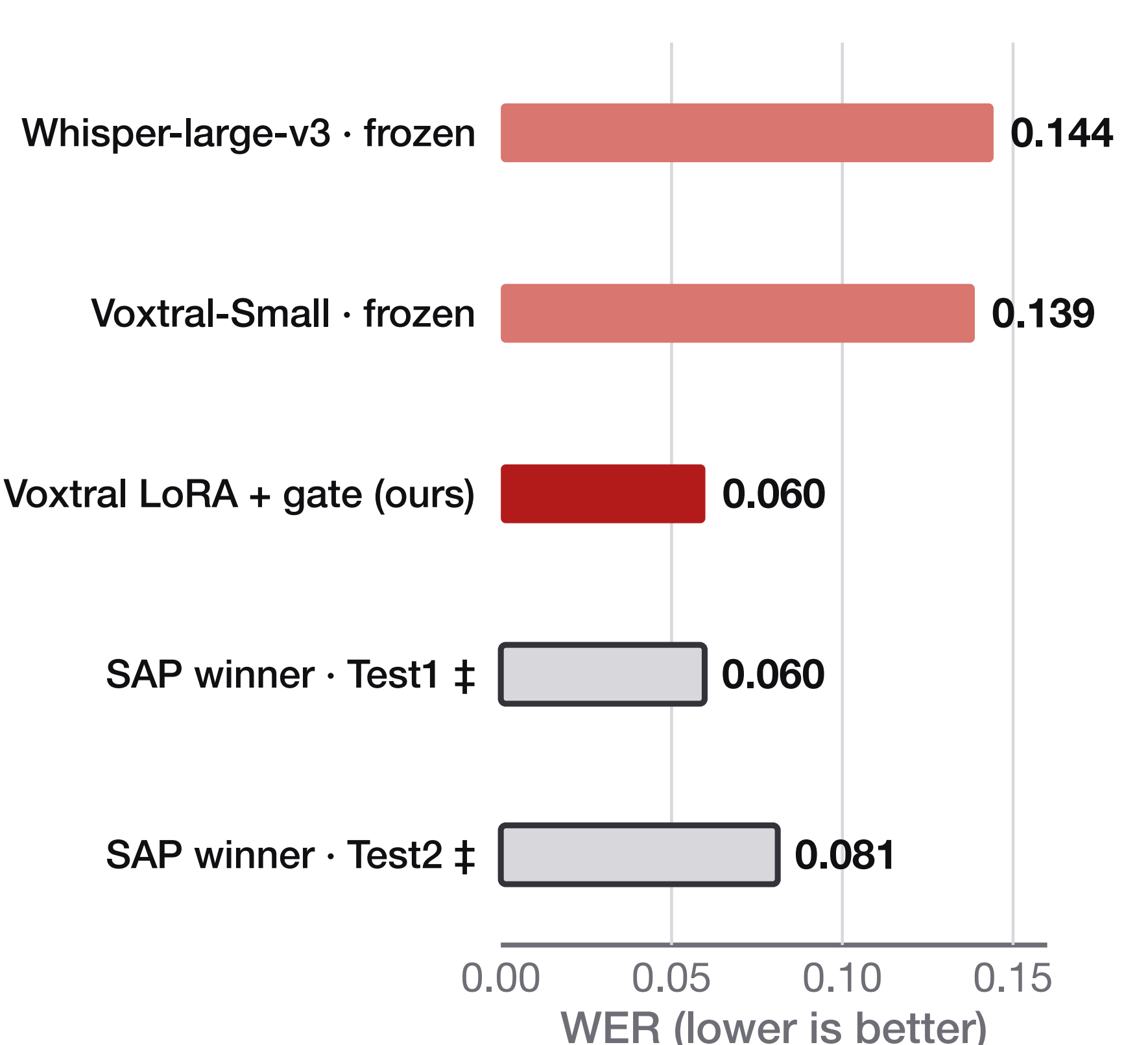
- Voxtral: 0.139 $\rightarrow$ 0.066 WER (–52%)
- Gemma-4-E2B too: 0.302 $\rightarrow$ 0.204
- Difficulty gate: another –10%

Hardest Speech Gains Most



- After LoRA, gain grows with difficulty
- +0.117 (Voxtral), +0.092 (Gemma)
- $\rightarrow$ give context only to hard speech

On Par with the SAP Challenge



- Rivals the Challenge winner (‡)
- With LoRA, not full fine-tuning
- Context: presence matters, not content

• Prompt structure, not clinical content, breaks frozen audio-LMs

• Yet context helps most on the hardest speech (difficulty gate)

• LoRA repairs it — rivals the Speech Accessibility Challenge winner