האוניברסיטה העברית בירושלים
THE HEBREW UNIVERSITY of JERUSALEM
الجامعة العبرية في القدس

ON THE **LANDSCAPE** OF **SPOKEN LANGUAGE MODELS**

Yossi Adi - The Hebrew University of Jerusalem, Israel

AGENDA

האוניברסיטה העברית בירושלים
THE HEBREW UNIVERSITY of JERUSALEM
الجامعة العبرية في القدس

- Speech language models
  - Motivation
  - Categorization and definitions
- Progress and scaling laws
- Going beyond spoken context in SLM evaluation
- Discussion & future directions

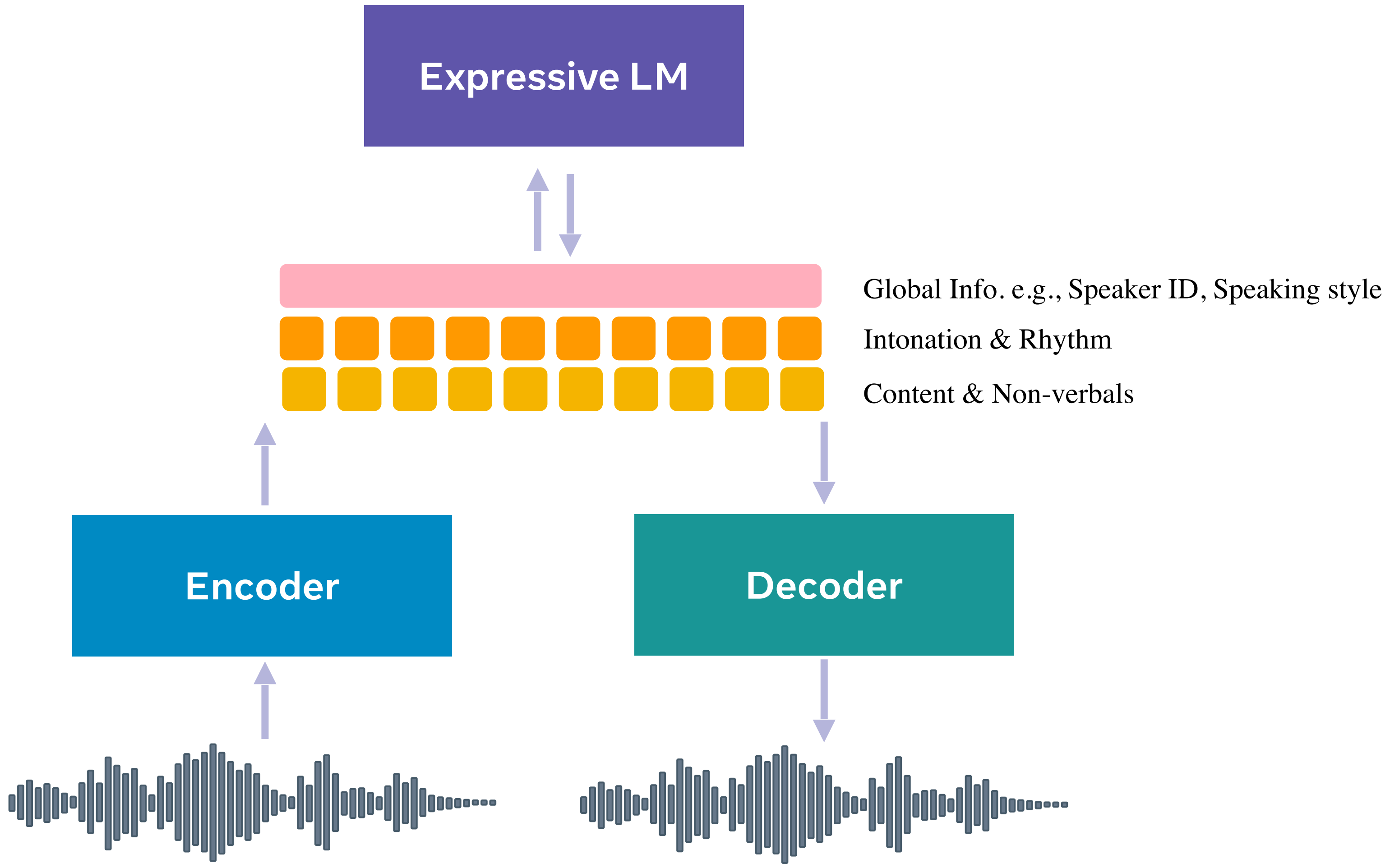
SPEECH LANGUAGE MODELS

MOTIVATION

- Speech and audio are the **primary** means of human **communication**.
- **Speaker-specific** properties beyond content (e.g., identity, style, emotion).
- Structured signals that are part of **natural human interaction** but are **not captured** in text (e.g., intonation, hesitation, laughter, smacking lips).
- Recording conditions / **non-speech** sounds.
- Can serve as both **generative** model and **universal** speech processor.

SPEECH LANGUAGE MODELS

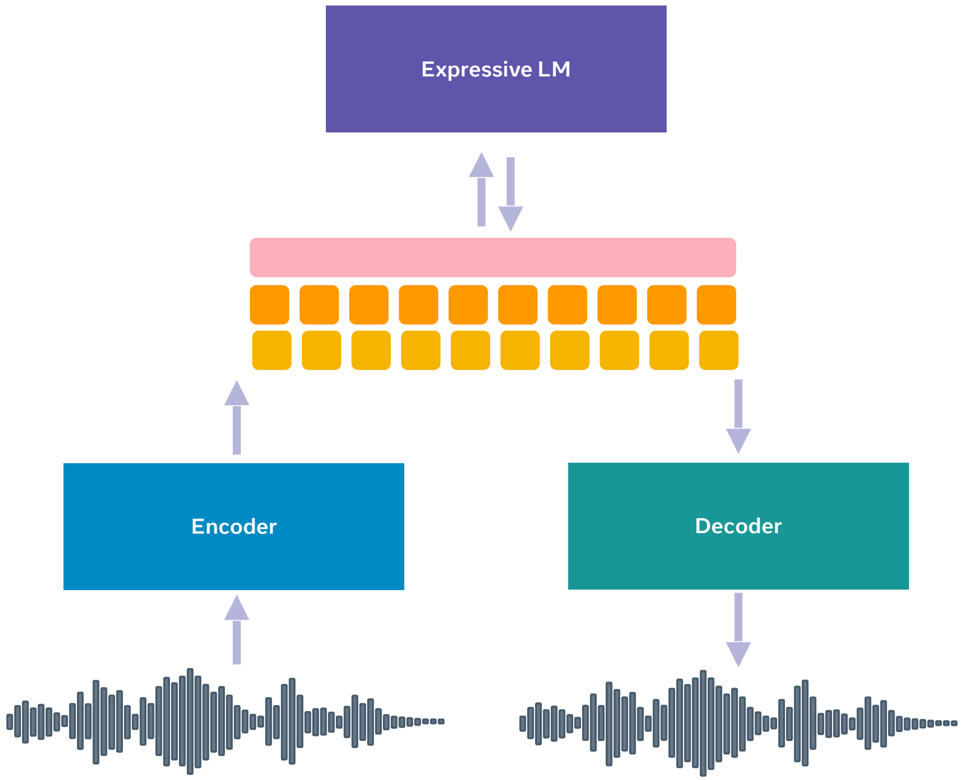
האוניברסיטה העברית בירושלים
THE HEBREW UNIVERSITY of JERUSALEM
الجامعة العبرية في القدس



SPEECH LANGUAGE MODELS

האוניברסיטה העברית בירושלים
THE HEBREW UNIVERSITY of JERUSALEM
الجامعة العبرية في القدس

- Different types of SLMs
  - e.g., textless-SLMs
  - Joint speech-text SLMs
  - Speech-aware LMs



- For more info about SLMs:

Arora, Siddhant, et al. "On the landscape of spoken language models: A comprehensive survey."

SPEECH LANGUAGE MODELS

Type of LM	Training Strat.	Distribution	Example
TextLM	pre-training	p(text)	GPT, Llama
TextLM	post-training	p(text I text)	ChatGPT, Llama-Inst.
textless-SpeechLM	pre-training	p(speech)	GSLM, AudioLM
textless-SpeechLM	post-training	p(speech I speech)	AlignSLM, Slamming
Joint speech-text LM	pre-training	p(text, speech)	SpiRitLM, Moshi
Joint speech-text LM	post-training	p(text, speech I text, speech)	Moshi, Mini-Omni
Speech-aware LM	post-training	p(text I speech, text)	Salmonn, Qwen-Audio-Chat

Arora, Siddhant, et al. "On the landscape of spoken language models: A comprehensive survey." arXiv preprint arXiv:2504.08528 (2025).

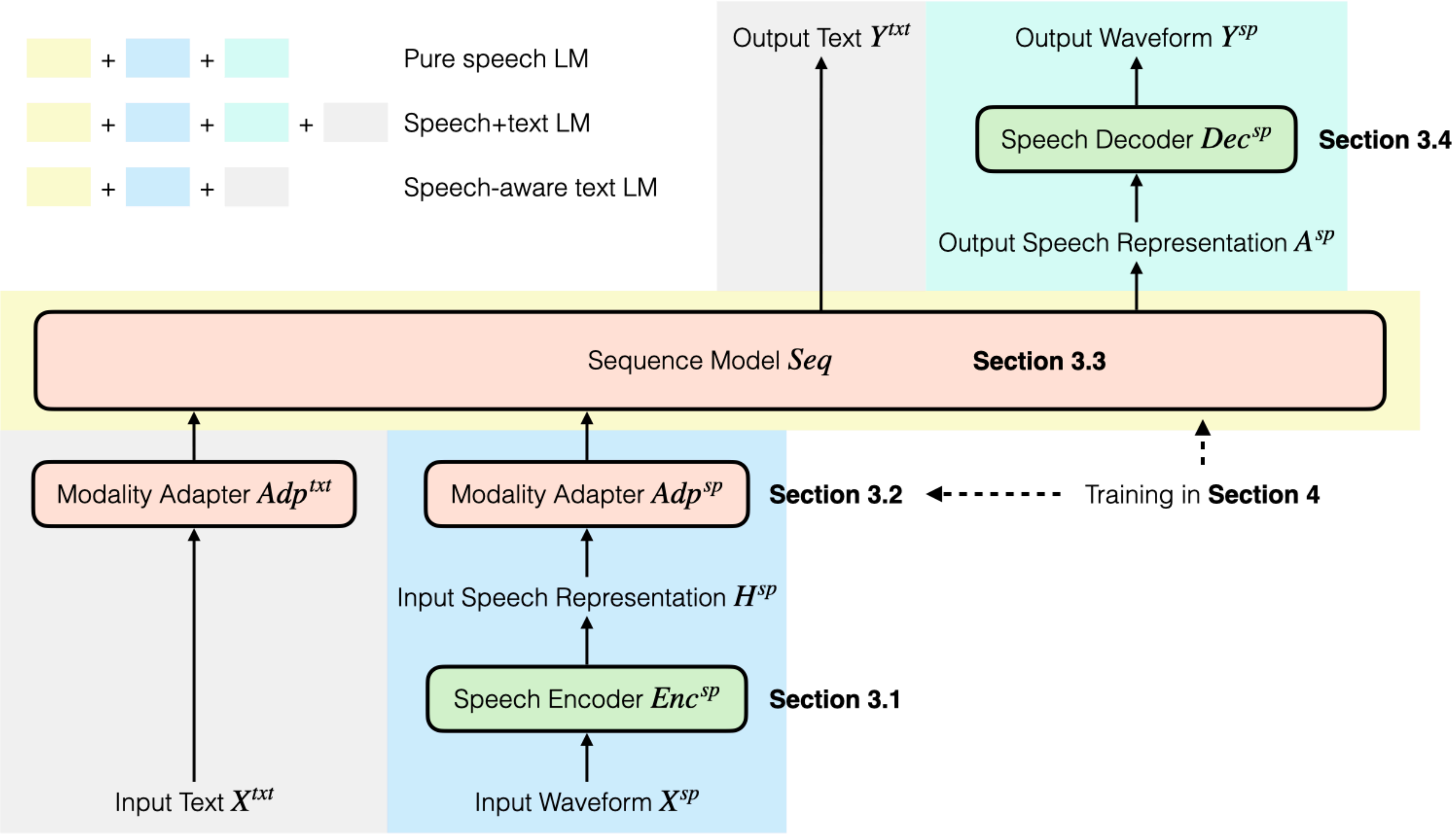
SPEECH LANGUAGE MODELS

- *Universal Speech Processing Systems*
- Definition:
  - It has both **spoken input** and **spoken output** with optional text input and/or output. The spoken input may serve as either an instruction or a context.
  - It is intended to be **"universal"**; that is, it should in principle be able to address **arbitrary spoken language tasks**, including both traditional tasks and more complex reasoning about spoken data.
  - It takes **instructions** or **prompts** in the form of natural language (either speech or text), and not, for example, task specifiers or soft prompts.

Arora, Siddhant, et al. "On the landscape of spoken language models: A comprehensive survey." arXiv preprint arXiv:2504.08528 (2025).

SPEECH LANGUAGE MODELS

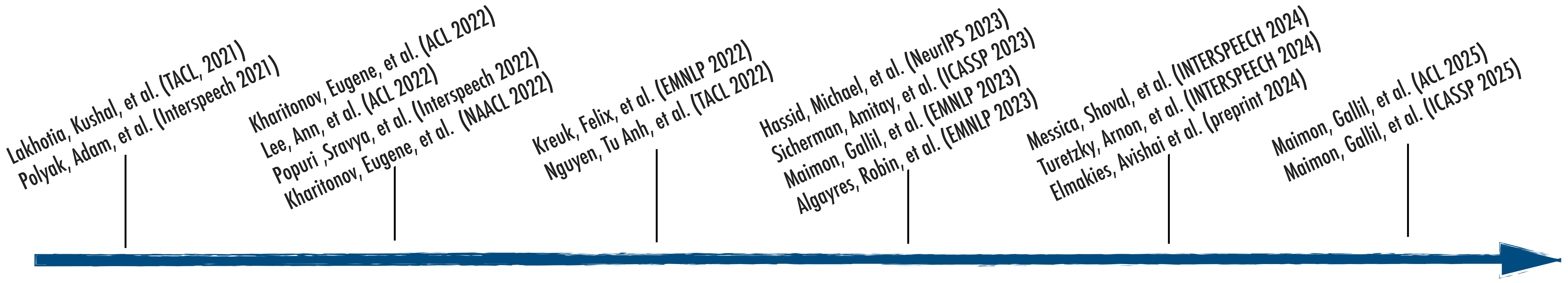
האוניברסיטה העברית בירושלים
THE HEBREW UNIVERSITY of JERUSALEM
الجامعة العبرية في القدس



Arora, Siddhant, et al. "On the landscape of spoken language models: A comprehensive survey." arXiv preprint arXiv:2504.08528 (2025).

TEXTLESS-SLMS

- In textless-SLMs we consider speech-only tokens (no textual tokens)
- We've made a lot of progress!



- Current textless-SLMs can be consistent with:
  - Speaker id, acoustics, etc.
  - Syntax is mostly ok
  - Semantics is not good :(

EXAMPLE

האוניברסיטה העברית בירושלים
THE HEBREW UNIVERSITY of JERUSALEM
الجامعة العبرية في القدس

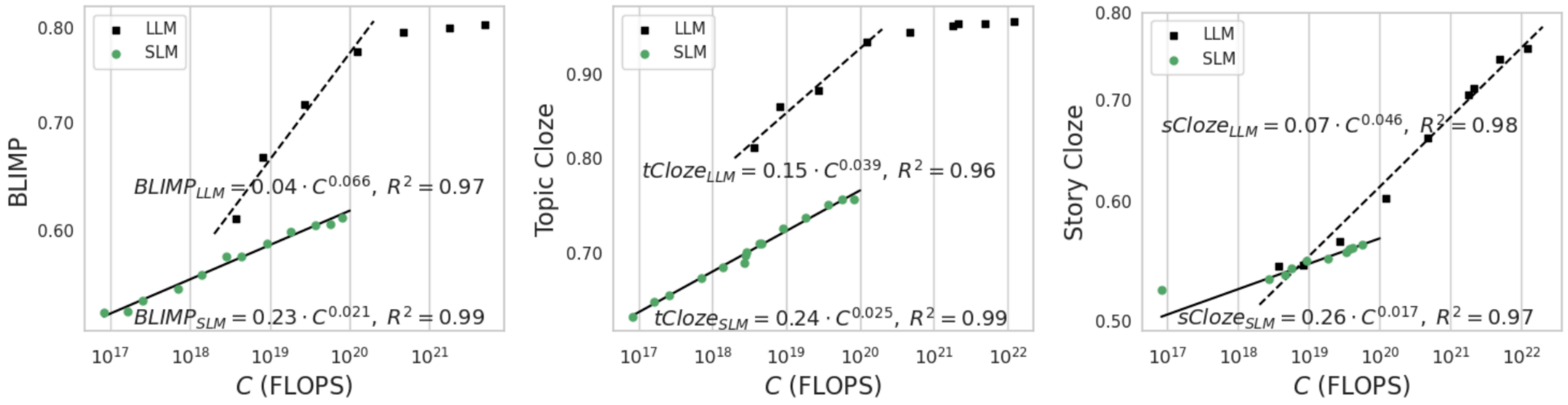
<u>Prompt</u>: *I really enjoy playing outside…*

<u>Continuation</u>: I really like to play outside because it forces me to really learn and really like, stay in my zone, I like that.

האוניברסיטה העברית בירושלים
THE HEBREW UNIVERSITY of JERUSALEM
الجامعة العبرية في القدس

TEXTLESS-SLMS

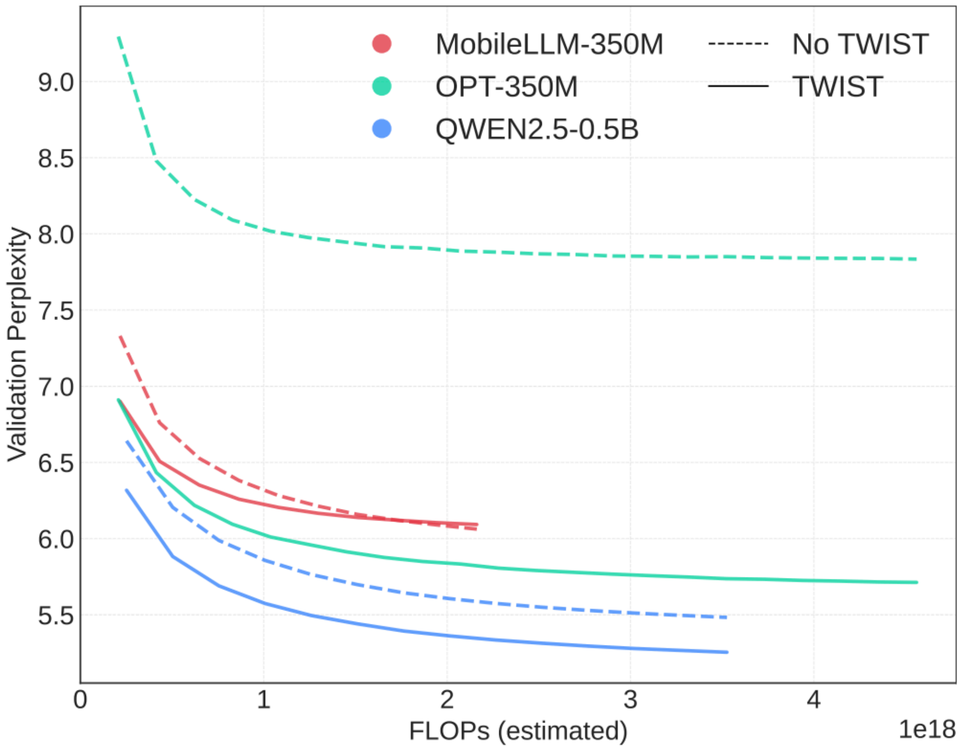
- Encouraging results!
- However, current SLM scaling laws show a somewhat pessimistic view.
  - Cuervo, Santiago, and Ricard Marxer. "Scaling properties of speech language models." *EMNLP* (2024).
- *tl;dr*: according to their study, we need 3x more data than text LMs!!!



TEXTLESS-SLMS

האוניברסיטה העברית בירושלים
THE HEBREW UNIVERSITY of JERUSALEM
الجامعة العبرية في القدس

- But how does TWIST initialization effect these scaling laws?
- What about model architecture? Some architectures might benefit more / less?
- Should we expect to get better performance following such an approach?

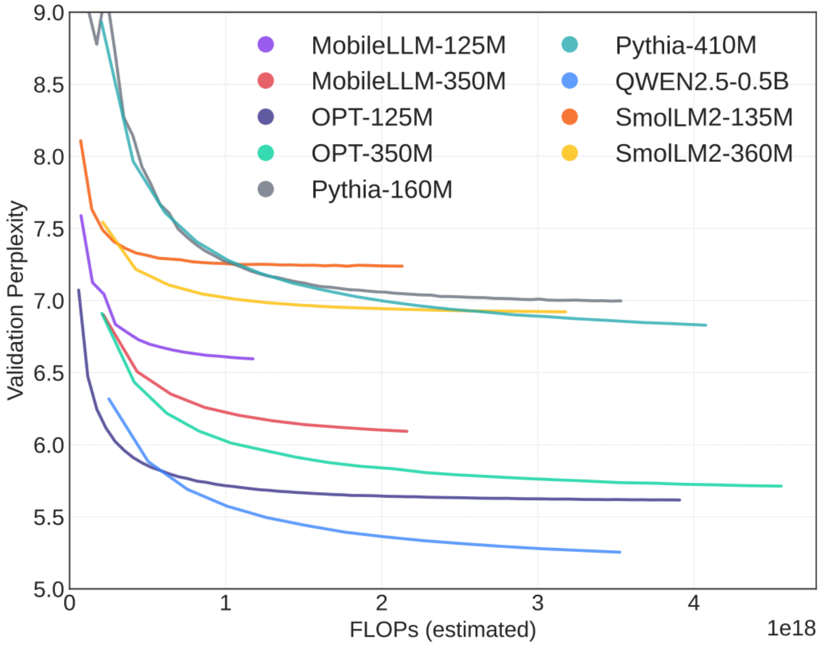


Maimon, Gallil, Avishai Elmakies, and Yossi Adi. "Slamming: Training a speech language model on one gpu in a day." ACL (2025).

TEXTLESS-SLMS

האוניברסיטה העברית בירושלים
THE HEBREW UNIVERSITY of JERUSALEM
الجامعة العبرية في القدس

- SO TWIST is good! Lets deep dive into the effect of model initialization?
- Bigger models does not always perform better!



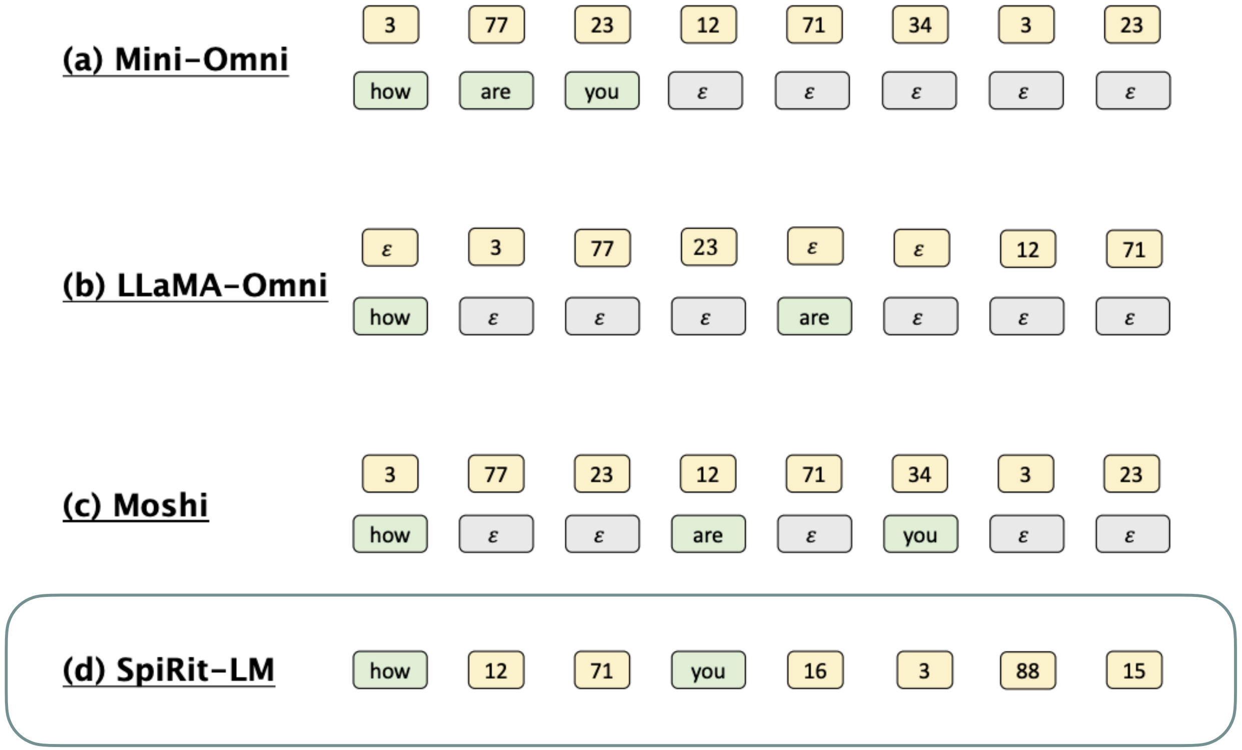
Maimon, Gallil, Avishai Elmakies, and Yossi Adi. "Slamming: Training a speech language model on one gpu in a day." ACL (2025).

JOINT SPEECH-TEXT SLMS

האוניברסיטה העברית בירושלים
THE HEBREW UNIVERSITY of JERUSALEM
الجامعة العبرية في القدس

- Will it be different when we incorporate text?

- How should we incorporate text?



Arora, Siddhant, et al. "On the landscape of spoken language models: A comprehensive survey." arXiv preprint arXiv:2504.08528 (2025).

SPEECH-TEXT INTERLEAVING

	Web Questions		Llama Questions		TriviaQA	
	S	S->T	S	S->T	S	S->T
Text LMs (text only)	-	32.3	-	75.0	-	56.4
Moshi	9.2	26.6	21.0	62.3	7.3	22.8
Zeng et al. w. syn. data	15.9	32.2	50.7	64.7	26.5	39.1

Zeng, Aohan, et al. "Scaling speech-text pre-training with synthetic interleaved data." *arXiv preprint arXiv:2411.17607* (2024).

SPEECH-TEXT INTERLEAVING

האוניברסיטה העברית בירושלים
THE HEBREW UNIVERSITY of JERUSALEM
الجامعة العبرية في القدس

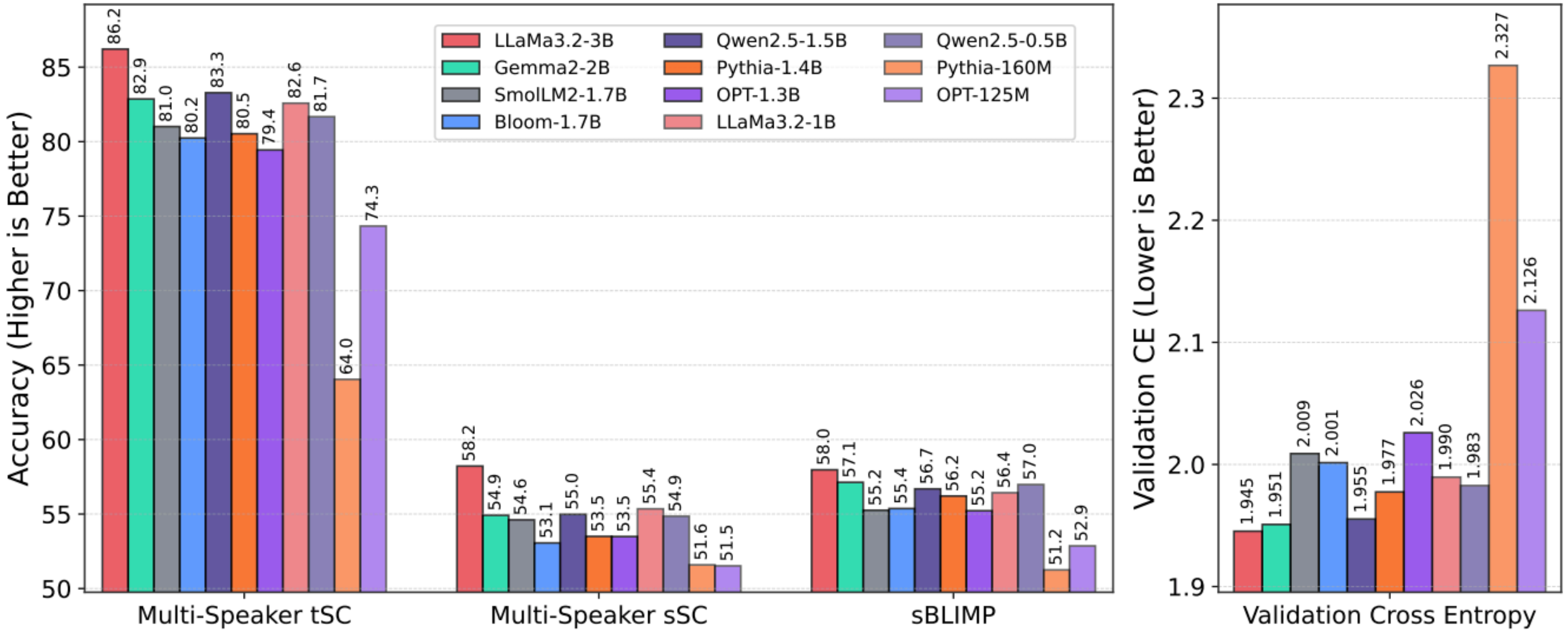
Training Data		Metric					
Real	Syn.	sBLIMP↑	sSC↑	MS_sSC↑	tSC↑	MS_tSC↑	Val. CE↓
✓	✗	56.77	54.94	52.66	72.15	78.93	**1.96129**
✗	✓	52.98	**61.30**	54.84	81.24	72.35	3.44569
✓	✓	**56.98**	59.81	**54.85**	**81.51**	**81.67**	1.98267

Maimon, Gallil, et al. "Scaling Analysis of Interleaved Speech-Text Language Models." *arXiv preprint arXiv:2504.02398* (2025).

SPEECH-TEXT INTERLEAVING

האוניברסיטה העברית בירושלים
THE HEBREW UNIVERSITY of JERUSALEM
الجامعة العبرية في القدس

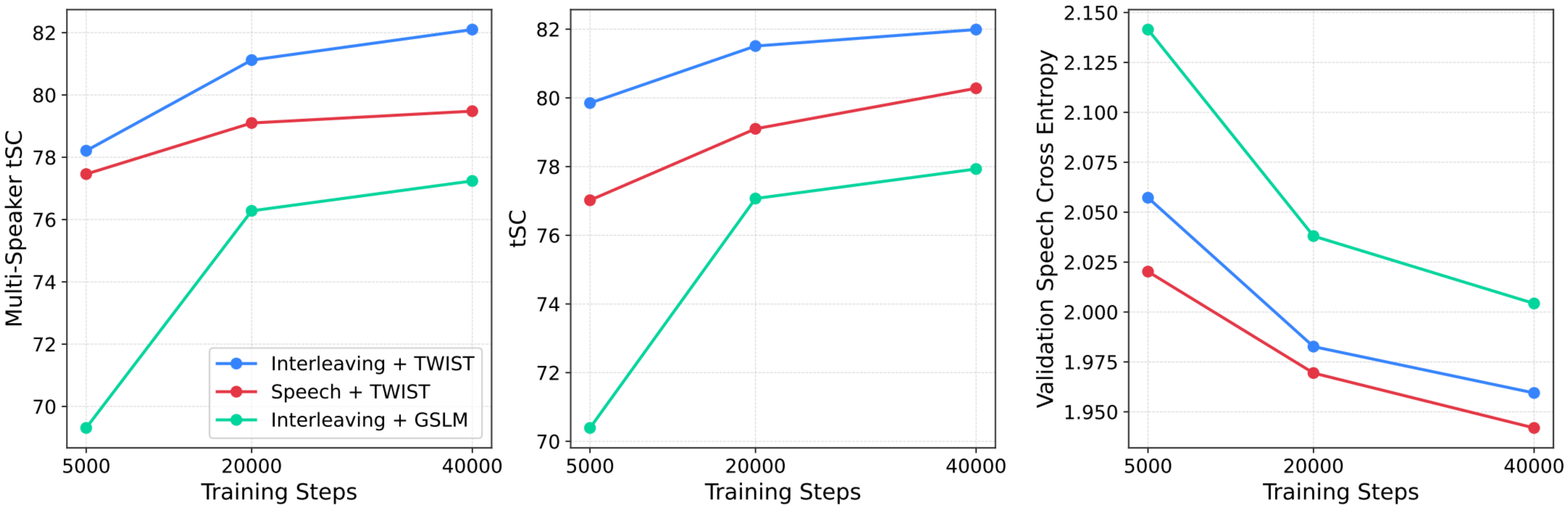
- So interleaving is great, and synthetic data too
- What is the effect of model type / family under the interleaving setup



Maimon, Gallil, et al. "Scaling Analysis of Interleaved Speech-Text Language Models." *arXiv preprint arXiv:2504.02398* (2025).

SPEECH-TEXT INTERLEAVING

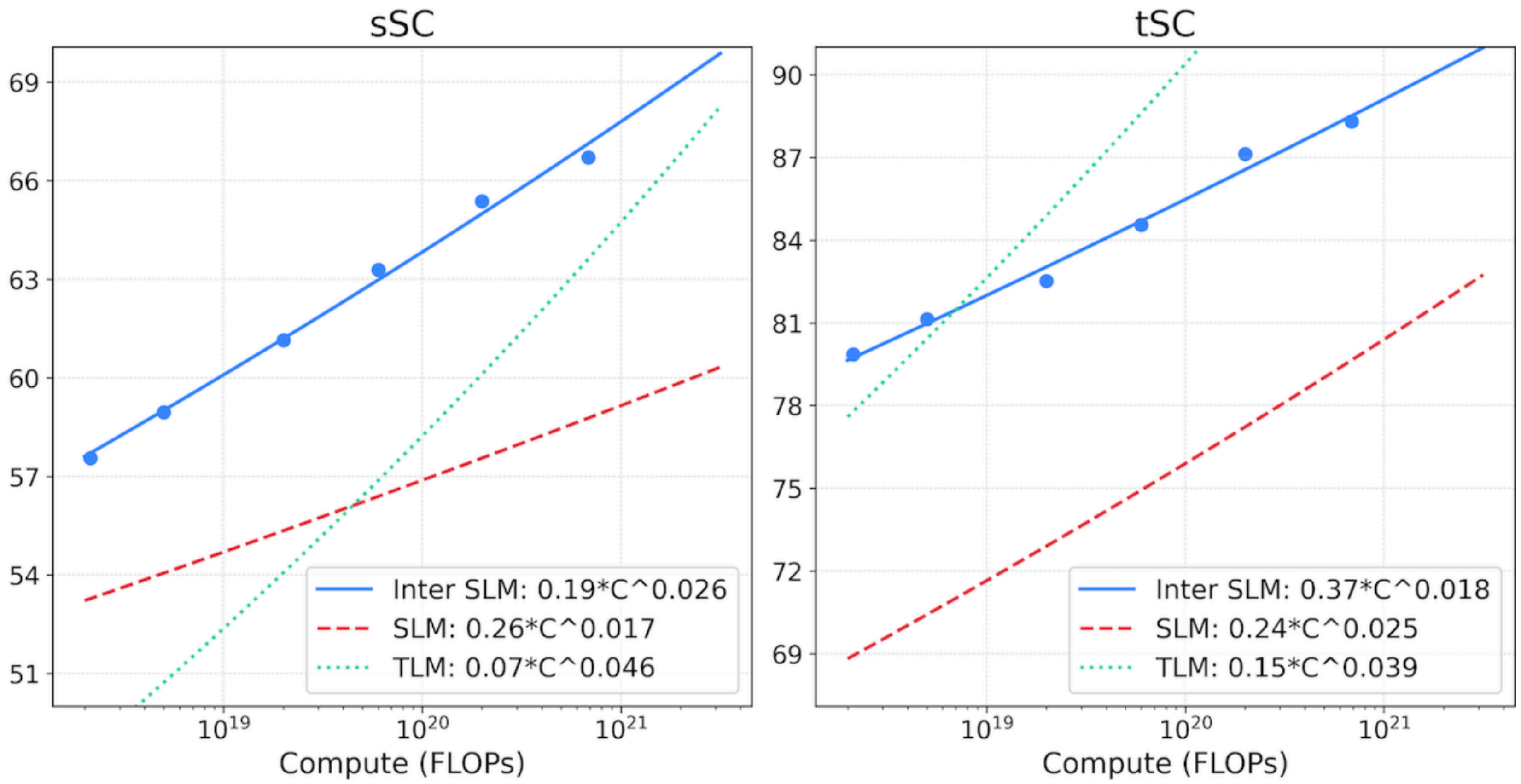
- Do we need both TWIST and speech-text interleaving?



Maimon, Gallil, et al. "Scaling Analysis of Interleaved Speech-Text Language Models." *arXiv preprint arXiv:2504.02398* (2025).

SPEECH-TEXT INTERLEAVING

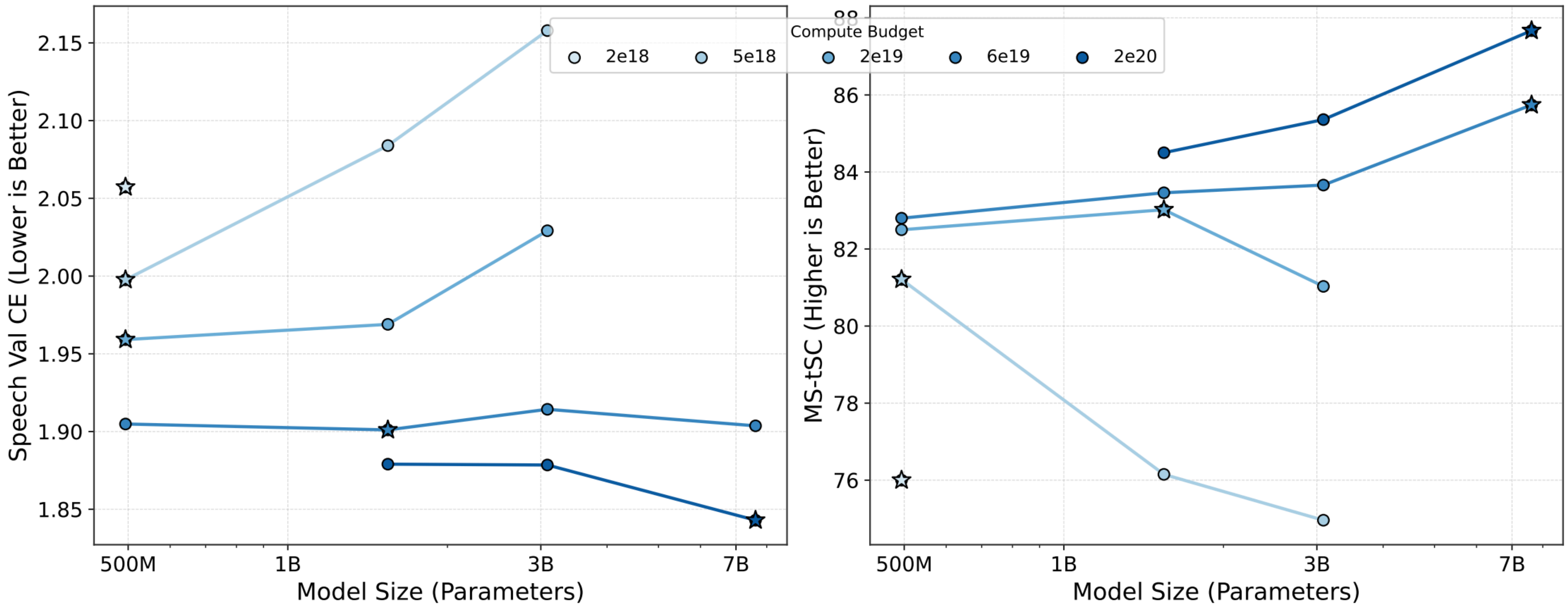
האוניברסיטה העברית בירושלים
THE HEBREW UNIVERSITY of JERUSALEM
الجامعة العبرية في القدس



Maimon, Gallil, et al. "Scaling Analysis of Interleaved Speech-Text Language Models." *arXiv preprint arXiv:2504.02398* (2025).

SPEECH-TEXT INTERLEAVING

האוניברסיטה העברית בירושלים
THE HEBREW UNIVERSITY of JERUSALEM
الجامعة العبرية في القدس



Maimon, Gallil, et al. "Scaling Analysis of Interleaved Speech-Text Language Models." *arXiv preprint arXiv:2504.02398* (2025).

EXAMPLE

Prompt: *The capital of France is*

Continuation: is Paris, that’s the city where we live. Paris is famous for it’s culture and glory, and the culture is….

Maimon, Gallil, et al. "Scaling Analysis of Interleaved Speech-Text Language Models." *arXiv preprint arXiv:2504.02398* (2025).

SPEECH-TEXT SLMS

האוניברסיטה העברית בירושלים
THE HEBREW UNIVERSITY of JERUSALEM
الجامعة العبرية في القدس

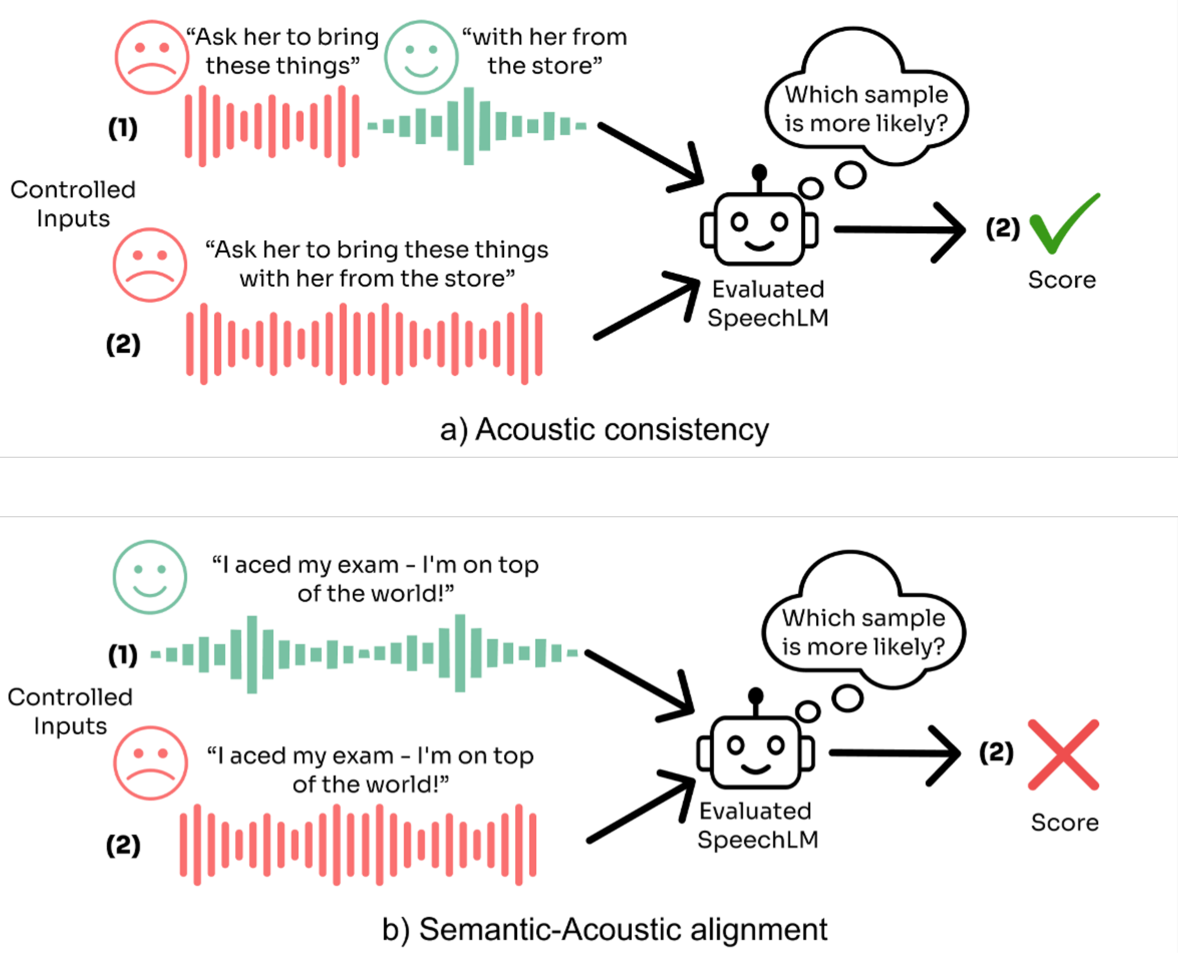
- Great progress!
- But all evaluations are based on textual content :(
- Topline: ASR -> LLM -> TTS

- Can we design a benchmark / evaluation task that will leverage the richness of spoken data?
- Benchmark in which text based LLMs would perform at a chance level?
- *SALMON: Maimon, Gallil, Amit Roth, and Yossi Adi. "A suite for acoustic language model evaluation." ICASSP, 2025.*

SALMON

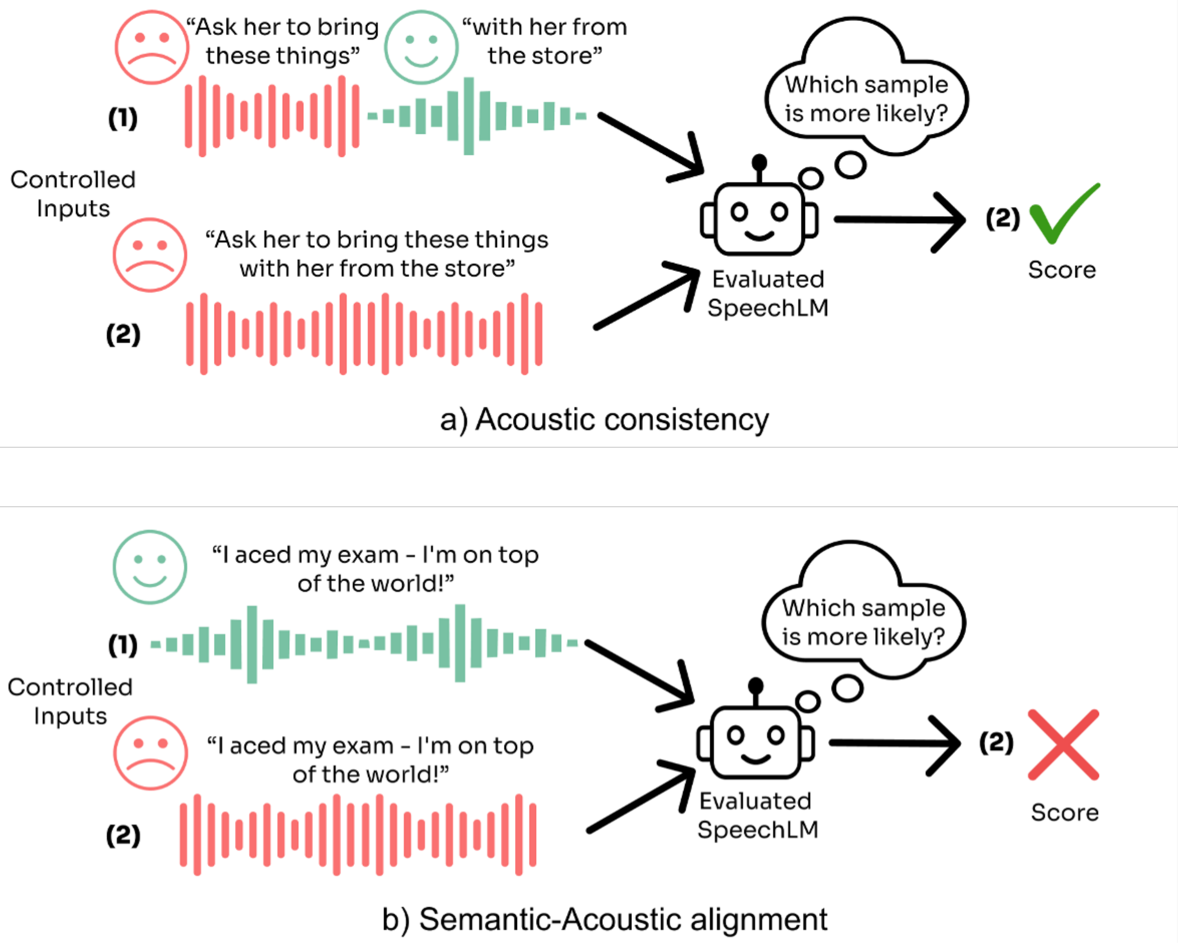
האוניברסיטה העברית בירושלים
THE HEBREW UNIVERSITY of JERUSALEM
الجامعة العبرية في القدس

- Draw inspiration from early NLP benchmarks
- Create pairs of audio samples: positive and negative
- Negative has same content, but unlikely acoustics
- Ask SpeechLM *"which sample is more likely?"*
- Each sub-set of the metric checks one aspect



SALMON

- Speech data:
  - VCTK, LJ
  - Expresso
  - AzureTTS
- Background noise:
  - FSD50K
- RIR:
  - EchoThief
- GPT4 - generate sentences for alignment tasks



SALMON - EXAMPLES

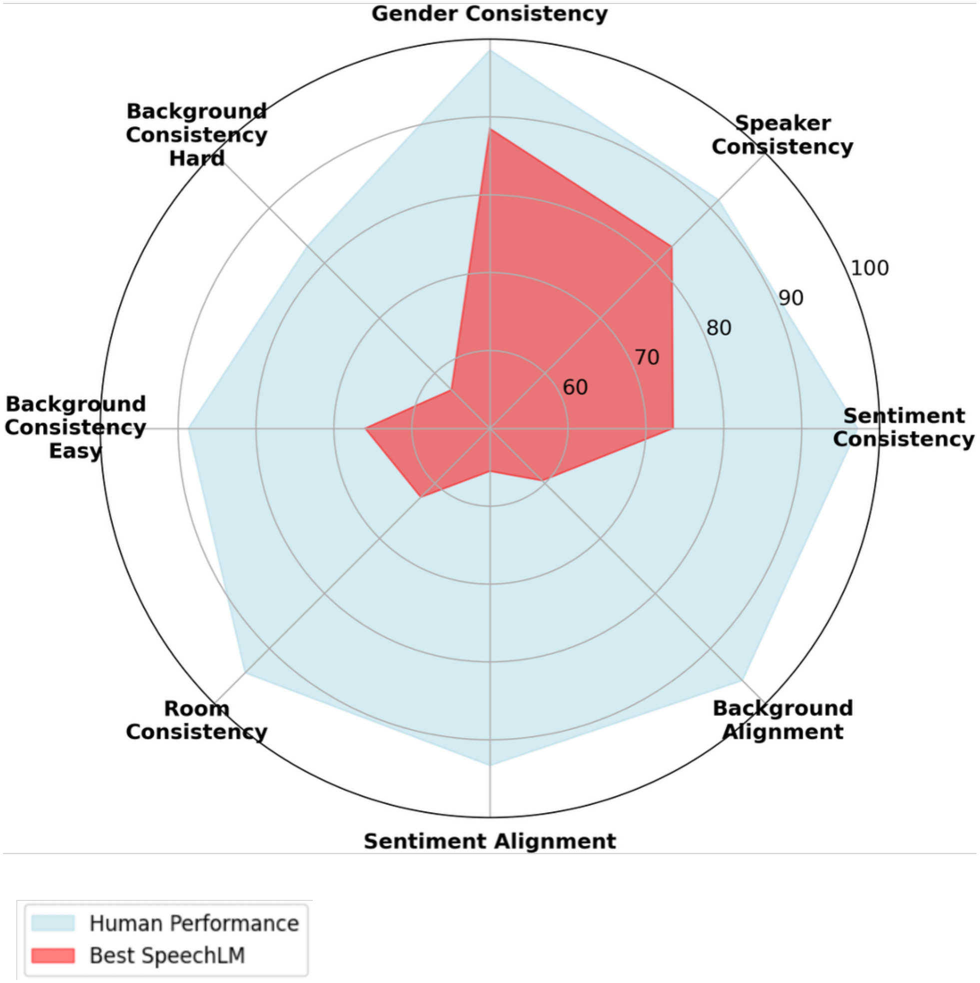
Speaker Consistency

Emotion Consistency

Background Alignment

SALMON

- The task is trivial to humans
- Even expressive SpeechLMs struggle to detect basic inconsistencies
- In detecting mis-match between acoustic elements and content - results are random!



MULTI-MODAL SLMS

האוניברסיטה העברית בירושלים
THE HEBREW UNIVERSITY of JERUSALEM
الجامعة العبرية في القدس

- Great progress!
- But all evaluations are based on textual content :(
- Topline: ASR -> LLM -> TTS

- Can we design a benchmark / evaluation task that will leverage the richness of spoken data?
- Benchmark in which text based LLMs would perform at a chance level?
- *SALMON: Maimon, Gallil, Amit Roth, and Yossi Adi. "A suite for acoustic language model evaluation." ICASSP, 2025.*
- ***Sentence Stress!***

SENTENCE STRESS

האוניברסיטה העברית בירושלים
THE HEBREW UNIVERSITY of JERUSALEM
الجامعة العبرية في القدس

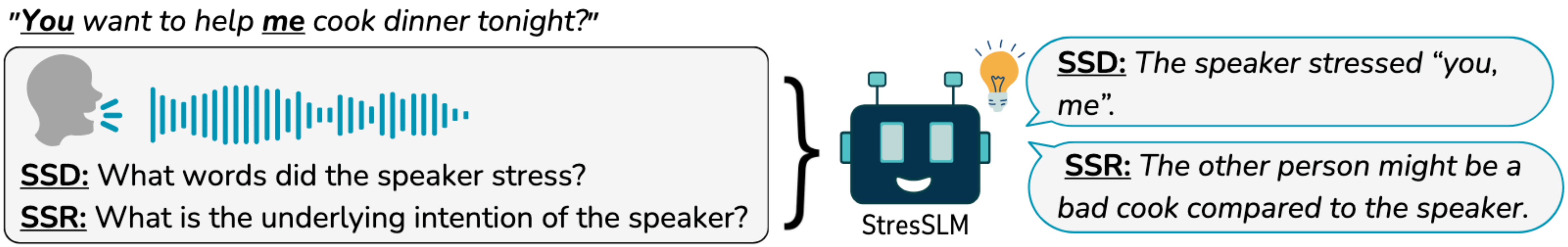
- Sentence stress refers to emphasis, placed on specific words within a spoken utterance to highlight or contrast an idea, or to introduce new information.

Stress Type	Description	Stressed speech (intention)
Contrastive	Demonstrates **contrast** with another option	"***I*** *didn't take your book.*" (vs. someone else) "*I didn't take your* ***book***." (vs. something else)
Emphatic	Amplifies or diminishes the **intensity** of a concept.	"*They* ***loved*** *how you treated her dog.*" (You really exceeded their expectations)
New-Information	Marks a **surprising** or **novel** content in the discourse.	"*He's actually moving to* ***New York.***" (surprising since its far from his current home)
Focus	General-purpose mechanism for **highlighting** key elements.	"*I enjoy the taste of espresso* ***at sunrise.***" (It's about that particular time)

STRESS-TEST

האוניברסיטה העברית בירושלים
THE HEBREW UNIVERSITY of JERUSALEM
الجامعة العبرية في القدس

- Yosha, Ido, et al. *StressTest*: Can YOUR Speech LM Handle the Stress?
- A **single-speaker** dataset (recorded by a professional actor) comprising 101 **manually curated** unique texts, each recorded with at least two distinct sentence stress pattern.



EXAMPLE

They **never** answer my calls

Highlighting that it absolutely never happened.

They never answer **my** calls

They might answer someone else's calls.

STRESS-TEST

- Model performance is measured in accuracy
- LLM-as-a-Judge
  - GPT-4o

$$\mathrm{SSR}_{acc}(\mathcal{M}, \mathcal{D}) = \frac{1}{|\mathcal{D}|} \sum_{(a,t,s,A,l)\in\mathcal{D}} \mathbb{I}\{\mathcal{J}(\mathcal{M}(a, \mathcal{P}(A))) = l\}.$$

- Where:
  - M - model, D - dataset, P - prompt, J - Judge
  - a - audio signal, A - possible set of answers, l - target answer

STRESS-TEST - AUDIO ONLY

Model	SSR
Qwen2Audio-7B-Instruct [Chu et al., 2024]	56.4
SALMONN [Tang et al., 2024]	56.8
LLaMA-Omni [Fang et al., 2025]	53.6
Phi-4-multimodal-instruct [Microsoft et al., 2025]	53.2
gpt-4o-audio [Hurst et al., 2024]	58.7
Human	**96.1**

STRESS-TEST - TEXT W. STRESS

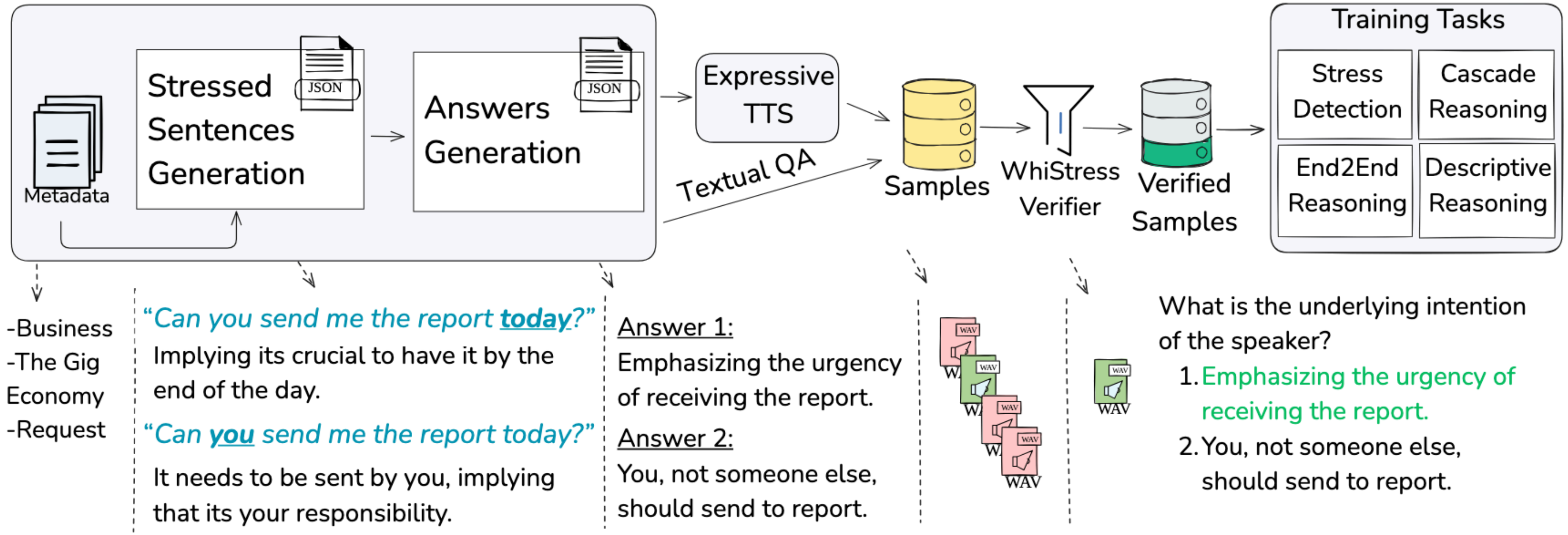
Model	SSR
TextLM w. Oracle transcription + sentence stress	
Llama-3.1-8B-Instruct [Grattafiori et al., 2024]	73.3
Qwen2-7B-Instruct [Yang et al., 2024a]	67.8
Qwen-7B-Chat [Bai and et al., 2023]	61.4
Cascade models	
Whiper+WhiStress → Llama-3.1-8B-Instruct	66.9
Whiper+WhiStress → Qwen2-7B-Instruct	63.7
Whiper+WhiStress → Qwen-7B-Chat	55.5

Yosha, Iddo, et al. "WhiStress: Enriching Transcriptions with Sentence Stress Detection", *arXiv preprint arXiv:2505.19103* (2025).

SYNTHETIC DATA GENERATION

האוניברסיטה העברית בירושלים
THE HEBREW UNIVERSITY of JERUSALEM
الجامعة العبرية في القدس

- ~17k hours of synthetic data. ~4.5k hours were verified.
- 4 type types of tasks



STRESS-SLM (SSR)

- Fine-tune Qwen2Audio-7B-Instruct model using our data
  - Including rehearsal data
  - Curriculum learning

Model	SSR
Qwen2Audio-7B-Instruct [Chu et al., 2024]	56.4
SALMONN [Tang et al., 2024]	56.8
LLaMA-Omni [Fang et al., 2025]	53.6
Phi-4-multimodal-instruct [Microsoft et al., 2025]	53.2
gpt-4o-audio [Hurst et al., 2024]	58.7
StressSLM (ours)	**81.6**

STRESS-SLM (SSD)

Model	Precision	Recall	F1
gpt-4o-audio [Hurst et al., 2024]	33.1	52.1	40.5
SALMONN [Tang et al., 2024]	19.1	29.5	23.2
LLaMA-Omni [Fang et al., 2025]	24.1	47.6	32.0
Phi-4-multimodal-instruct [Microsoft et al., 2025]	19.9	32.8	24.7
Qwen2Audio-7B-Instruct [Chu et al., 2024]	24.6	46.2	32.1
StressSLM (ours)	**89.6**	**83.3**	**86.4**
WhiStress (verifier)	**88.8**	**88.1**	**88.5**

Yosha, Iddo, et al. “WhiStress: Enriching Transcriptions with Sentence Stress Detection”, *arXiv preprint arXiv:2505.19103* (2025).

STRESS-SLM (SSD) - EXPRESSO

Model	Precision	Recall	F1
gpt-4o-audio [Hurst et al., 2024]	23.6	66.1	34.7
SALMONN [Tang et al., 2024]	13.2	45.5	20.5
LLaMA-Omni [Fang et al., 2025]	18.7	58.2	28.3
Phi-4-multimodal-instruct [Microsoft et al., 2025]	22.5	37.5	28.2
Qwen2Audio-7B-Instruct [Chu et al., 2024]	34.2	30.6	32.3
StressSLM (ours)	**51.8**	**68.6**	**59.1**
WhiStress (verifier)	**57.3**	**86.3**	**68.9**

Nguyen, Tu Anh, et al. "Expresso: A benchmark and analysis of discrete expressive speech resynthesis." *Interspeech* (2023).
Yosha, Iddo, et al. “WhiStress: Enriching Transcriptions with Sentence Stress Detection”, *arXiv preprint arXiv:2505.19103* (2025).

STRESS-SLM

Verifier	Training Samples	SSD			SSR
		Precision	*Recall*	*F1*	*Acc.*
✓	~4K	87.3	76.3	81.4	79.3
✗	~17K	87.4	81.9	84.5	76.6
✗→ ✓	~17K → ~4K	88.3	83.7	85.9	78.4

STRESS-SLM

Model	ASR (WER)				SER
	dev-clean	*dev-other*	*test-clean*	*test-other*	*MELD*
Qwen2Audio-7B-Instruct	2.30	4.64	2.31	4.92	54.6
StressSLM (ours)	2.70	4.60	2.46	5.50	57.2

SUMMERY AND DISCUSSION

האוניברסיטה העברית בירושלים
THE HEBREW UNIVERSITY of JERUSALEM
الجامعة العبرية في القدس

- Discussed a bit about SLMs
- Categorization and definitions
- SLMs scaling laws

- Joint Speech-Text Interleaving
- Multi-modality improves scaling properties!

- Benchmarking SLMs beyond spoken content

SUMMERY AND DISCUSSION

האוניברסיטה העברית בירושלים
THE HEBREW UNIVERSITY of JERUSALEM
الجامعة العبرية في القدس

- What do we really want to get from these models?
  - An interface to textLLMs vs. universal speech processing systems
- What do we want SLMs to support?
  - ASR, TTS, reason over audio, source-separation, denoising, etc.
  - Each will affect the modeling and benchmark choices

- For instance, in textlessSLMs we should not expect to text on factual knowledge (e.g., QA)
- Goal should be to mimic a 3-5 years old, not a knowledge resource
- For speech-aware SLM, we should focus on speech properties! basically a speech interface to LLMs

האוניברסיטה העברית בירושלים
THE HEBREW UNIVERSITY of JERUSALEM
الجامعة العبرية في القدس

THANKS!

ADIYOSS@MAIL.HUJI.AC.IL