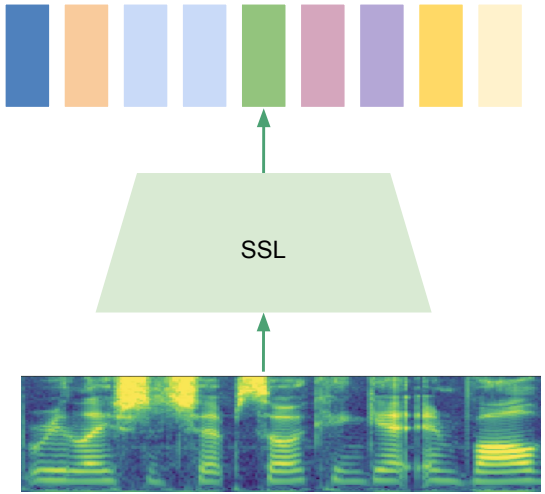
Advancing the Linguistic Capabilities of Speech Language Models

Ricard Marxer
Université de Toulon
ILLS, CNRS

Conversational AI reading group
March 12th 2026

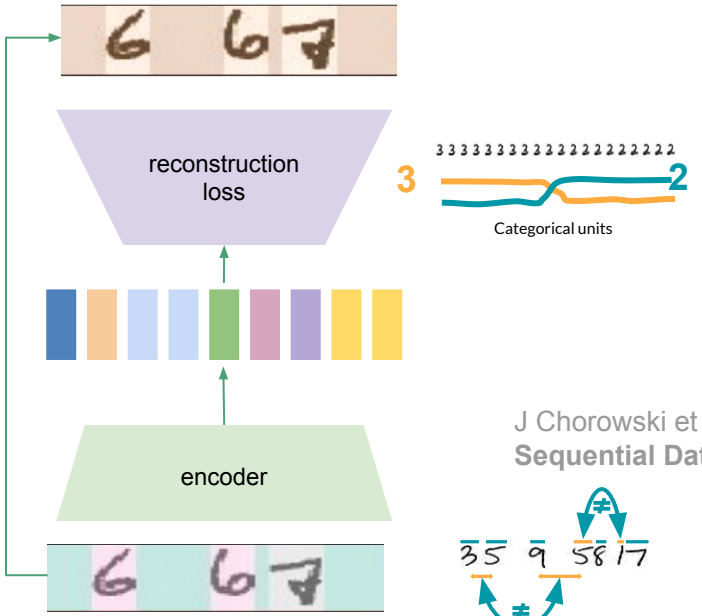
ILLS
International Laboratory on Learning Systems



Early work on self-supervised learning of speech

*with **Jan Chorowski** et al.*

Vector quantized autoencoder

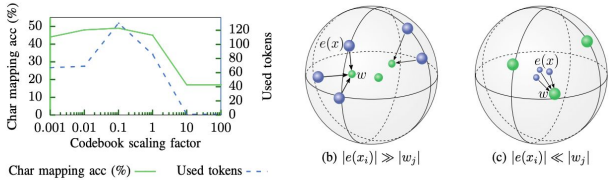


VQ-VAE, van der Oord et al. 2017

A Łańcucki et al.. **Robust Training of Vector Quantized Bottleneck Models**. 2020

Identify **difficulties with quantisation** bottlenecks

Solutions based on:
- normalisation,
- learning rate adaptation,
- codebook reinitialisation

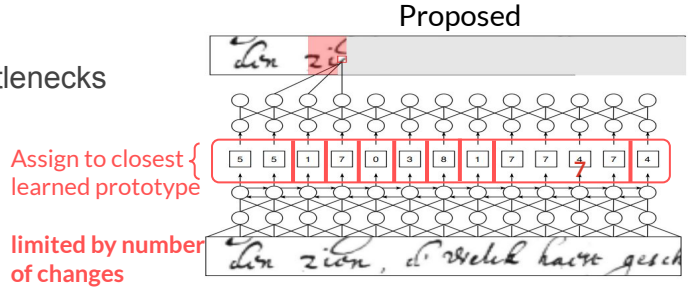


J Chorowski et al.. **Unsupervised Neural Segmentation and Clustering for Unit Discovery in Sequential Data**. 2019

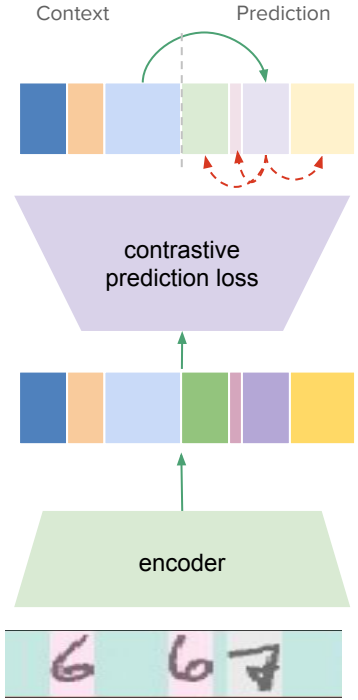
Segmental quantization bottlenecks

Equivalences in **HMM**

Exploit weak **duration supervision**

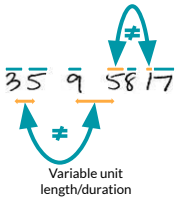


Contrastive predictive coding



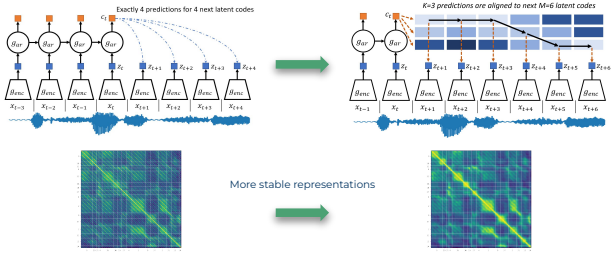
CPC, van der Oord et al. 2018

J Chorowski et al.. **Aligned Contrastive Predictive Coding**. 2021



Align fewer **predictions** to future timesteps

Piecewise linear representation ie **segmentation**



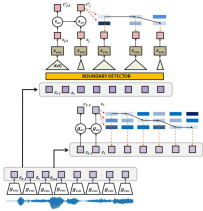
S Cuervo et al. **Contrastive prediction strategies for unsupervised segmentation and categorization of phonemes and words**. 2022

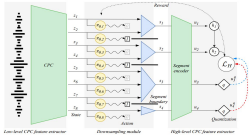
S Cuervo et al. **Variable-rate hierarchical CPC leads to acoustic unit discovery in speech**. 2022

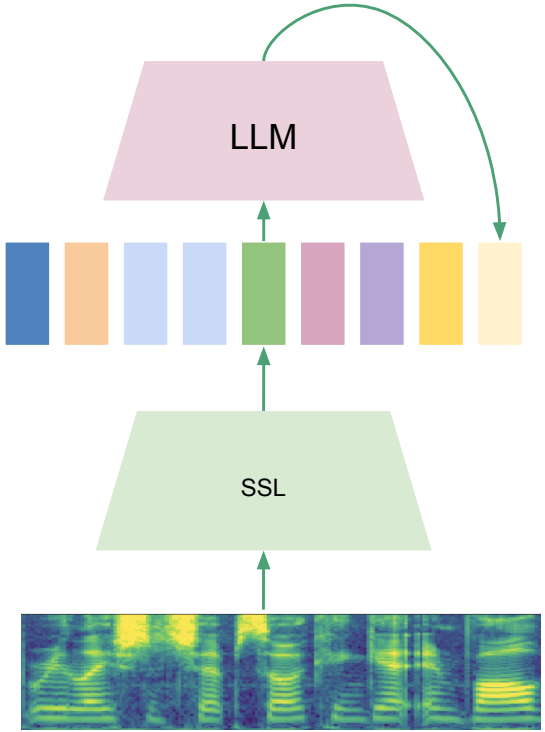
Contrastive prediction on segments
ie language modelling on phones

Top-down feedback and improve segmentation

RL for segmentation learning







Speech Language Models

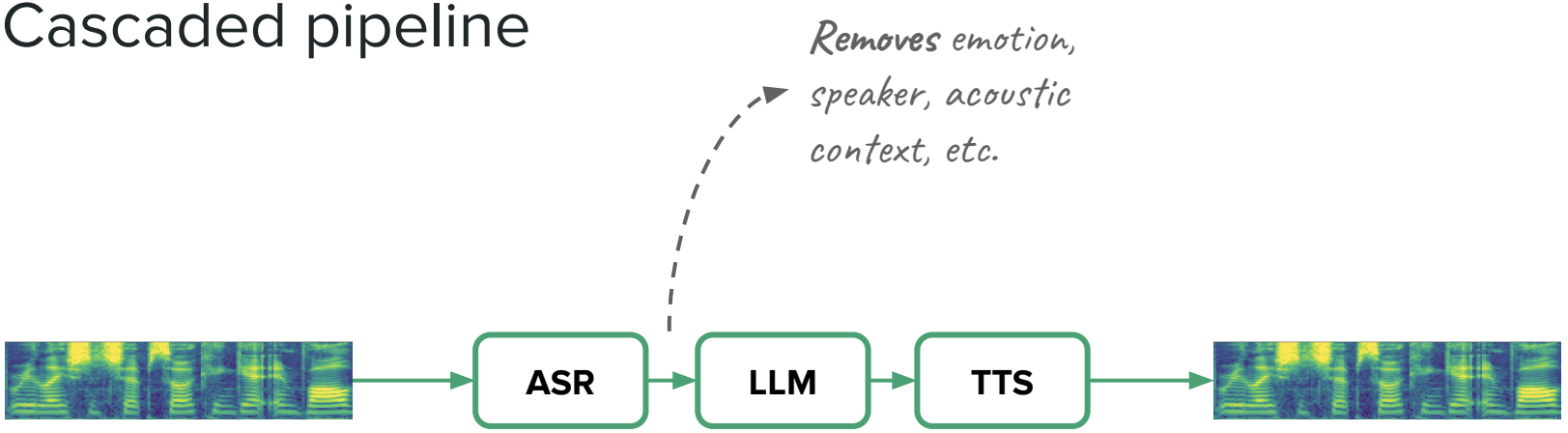
*with **Santiago Cuervo** et al.*

Text-based LLMs

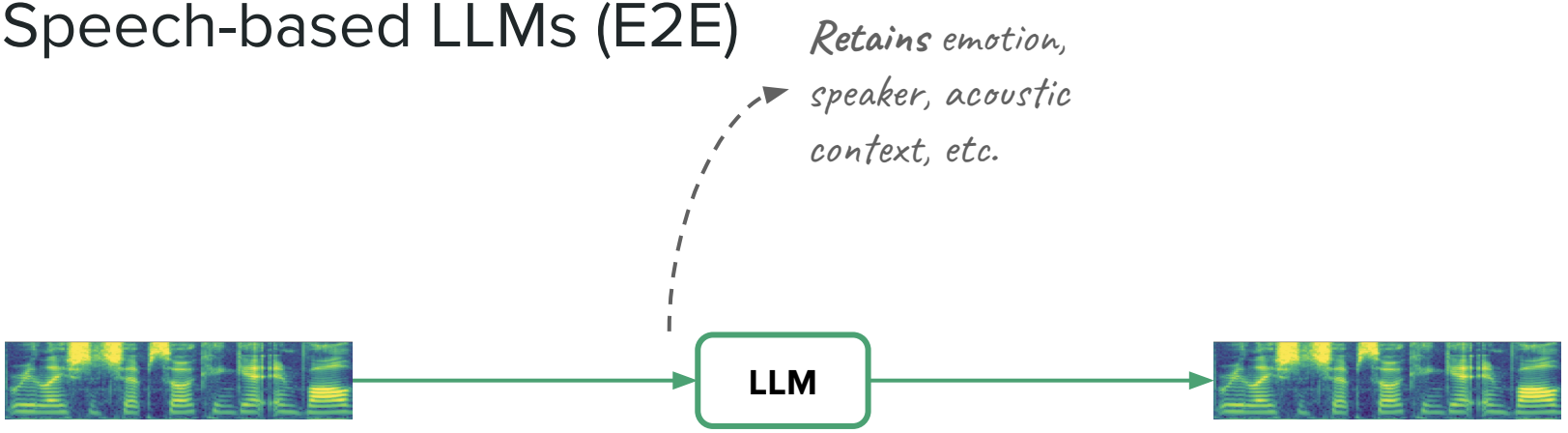
Speech-based LLMs

Cascaded pipeline

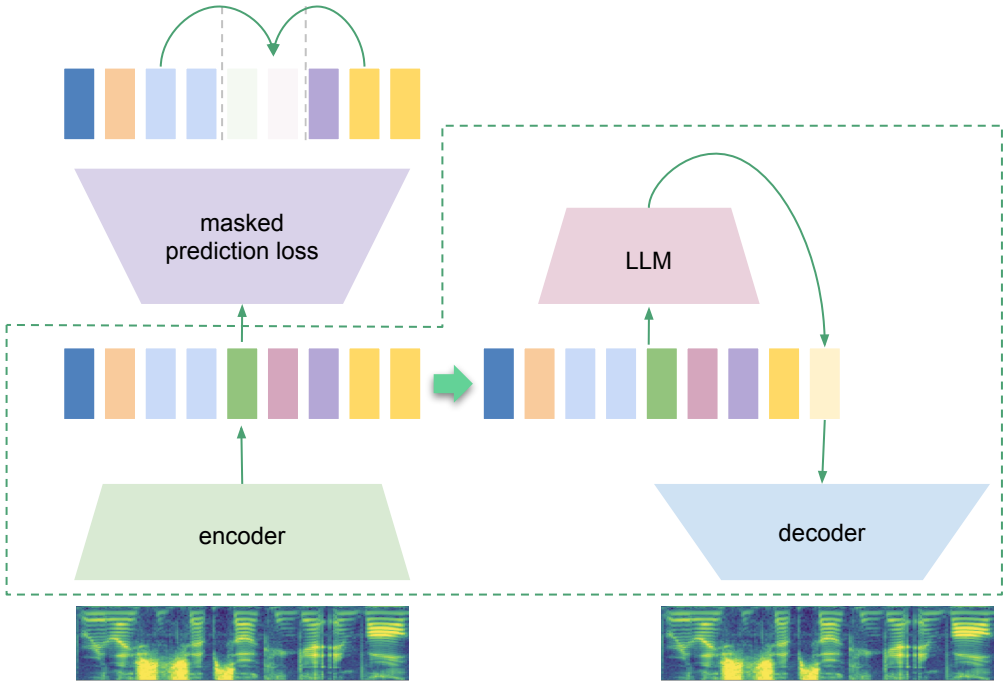
Cascaded pipeline



Speech-based LLMs (E2E)



Masked Prediction - Spoken language models



Leverage **self-supervised speech/audio representations** to perform tokenization

Speech token sequences processed using **same architectures as LLMs** (e.g. GPT)

Neural audio codecs (decoders) permit synthesis/generative modelling

NLP applications:

- Spoken language **understanding**
- **Expressive speech** synthesis
- Spoken **dialogue** management

GSLM, Lakhotia et al 2021
TWIST, Hassid et al 2023

Scaling of SLMs

Cuervo S, Marxer R. Scaling properties of speech language models. InProceedings of the 2024 Conference on Empirical Methods in Natural Language Processing 2024 Nov (pp. 351-361).

Empirical scaling laws have driven text LLM development

Test loss and downstream performance **scale as a power law** of compute (training tokens and parameter count)

Scaling Laws for Neural Language Models

Jared Kaplan *
Johns Hopkins University, OpenAI
jaredk@jhu.edu

Sam McCandlish*
OpenAI
sam@openai.com

Tom Henighan
OpenAI
henighan@openai.com

Tom B. Brown
OpenAI
tom@openai.com

Benjamin Chess
OpenAI
bchess@openai.com

Rewon Child
OpenAI
rewon@openai.com

Scott Gray
OpenAI
scott@openai.com

Alec Radford
OpenAI
alec@openai.com

Jeffrey Wu
OpenAI
jeffwu@openai.com

Dario Amodei
OpenAI
damodei@openai.com

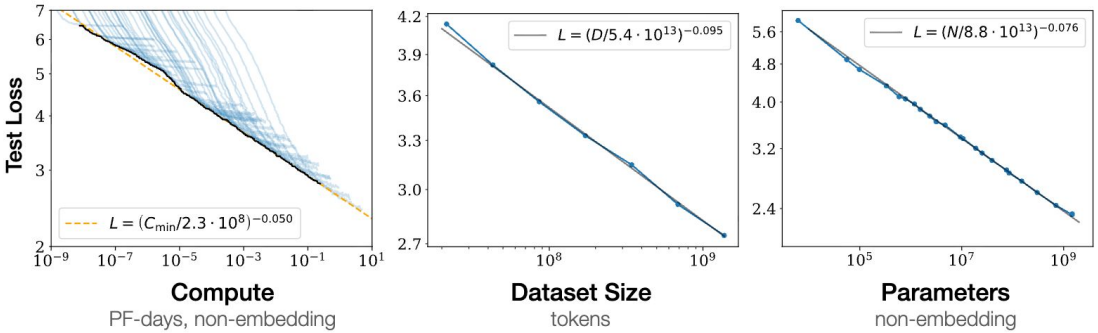
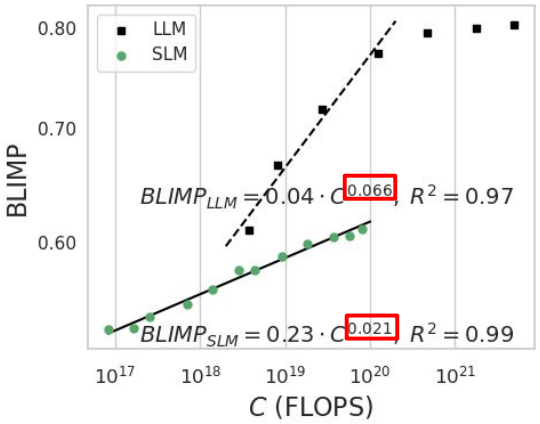
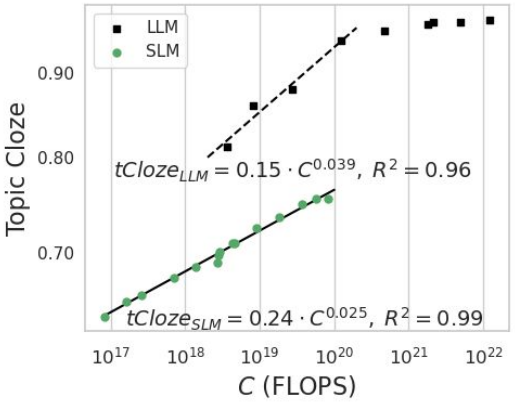


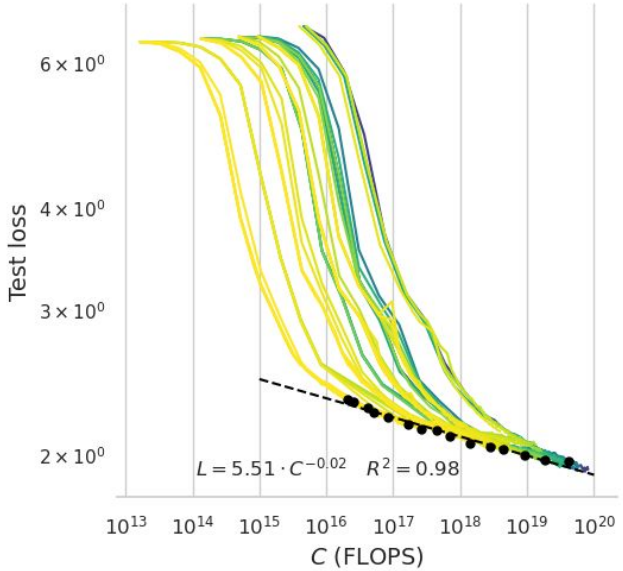
Figure 1 Language modeling performance improves smoothly as we increase the model size, datasetset size, and amount of compute[2] used for training. For optimal performance all three factors must be scaled up in tandem. Empirical performance has a power-law relationship with each individual factor when not bottlenecked by the other two.

Will current methods scale to match the performance of text-based NLP?

SLMs also **scale as a power law** of compute







However, **x1000 slower** than text-LLMs

Two hypotheses for the gap

Difference between modalities such as token rates (and context lengths)

alt → ***Unigram SentencePiece***

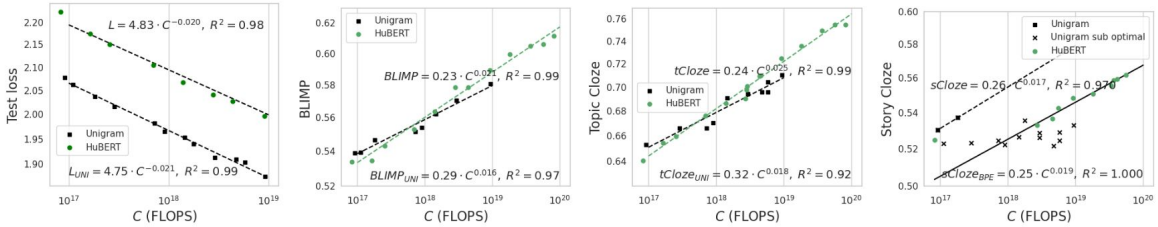
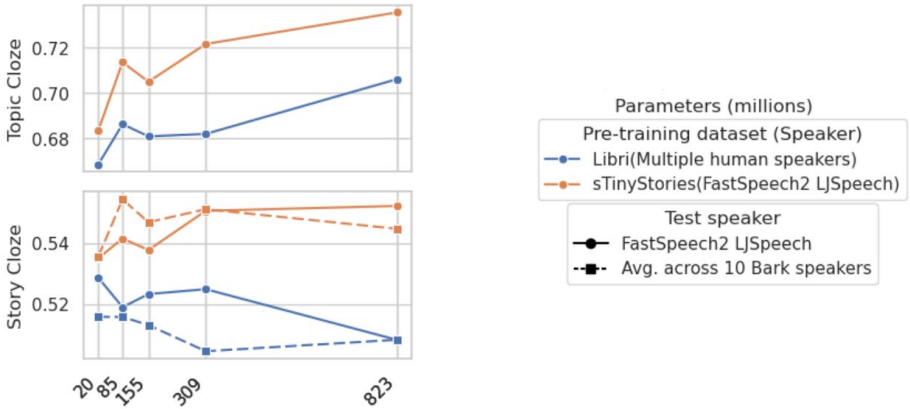


Figure 5: Comparison of the scaling behavior of SLMs trained on raw speech tokens and unigram compressed tokens. Axes are in logarithmic scale. The upstream loss of SLMs trained on unigram tokens scales better with compute, but downstream performance scales worse. Notably, the Story Cloze metric for SLMs trained on unigram tokens does not seem to improve with increased compute.



Data distribution mismatch: short utterances with little semantics (mostly used for ASR and TTS)

alt → ***sTinyStories***

Can we leverage knowledge from text LMs?

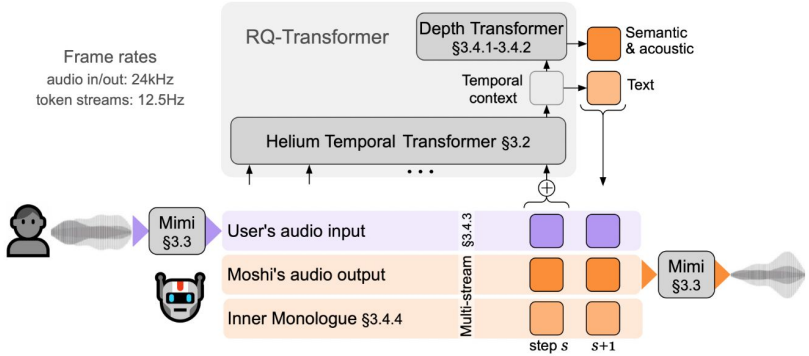


Figure 1: **Overview of Moshi**. Moshi is a speech-text foundation model which enables real-time spoken dialogue. The main components of Moshi's architecture are: a bespoke text language model backbone (Helium, see Section 3.2); a neural audio codec with residual vector quantization and with semantic knowledge distilled from a self-supervised speech model (Mimi, Section 3.3); the streaming, hierarchical generation of semantic and acoustic tokens for both the user and Moshi, along with time-aligned text tokens for Moshi when using Inner Monologue (Section 3.4).

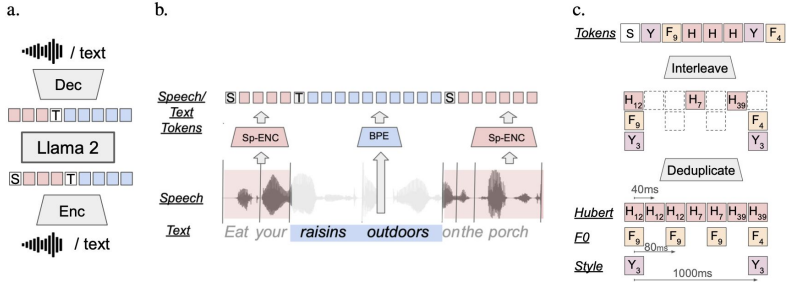
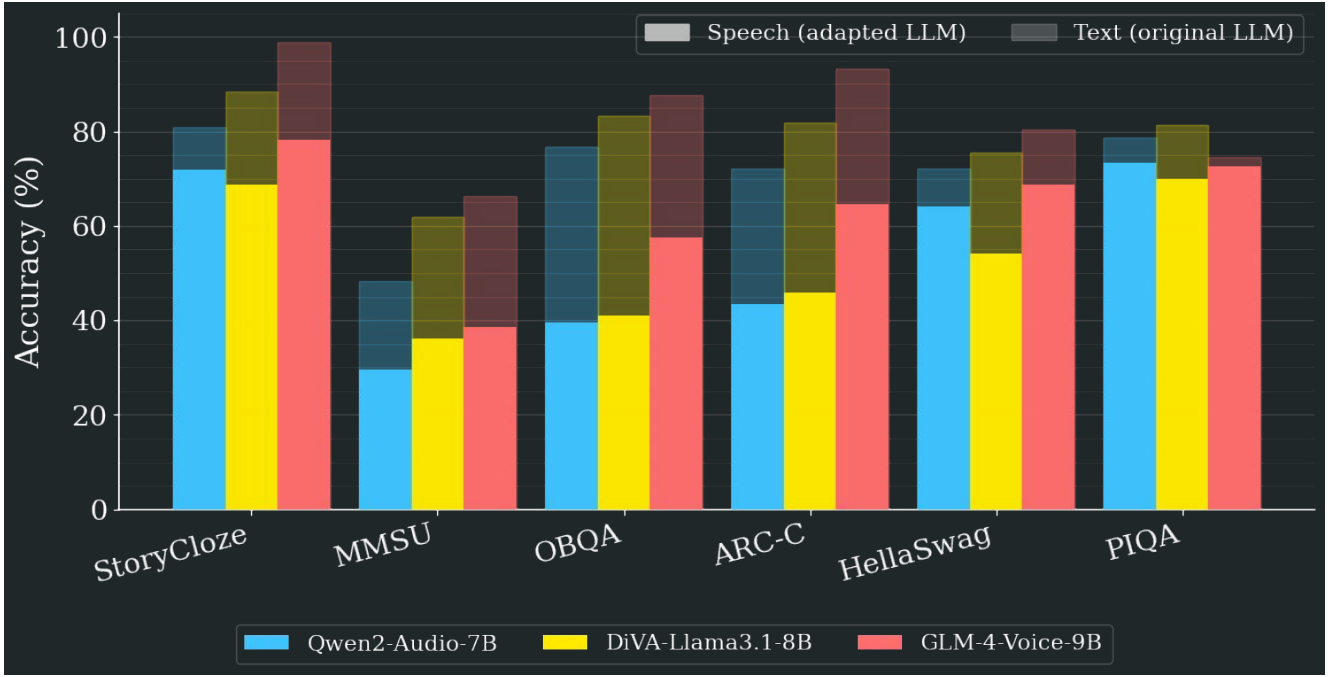
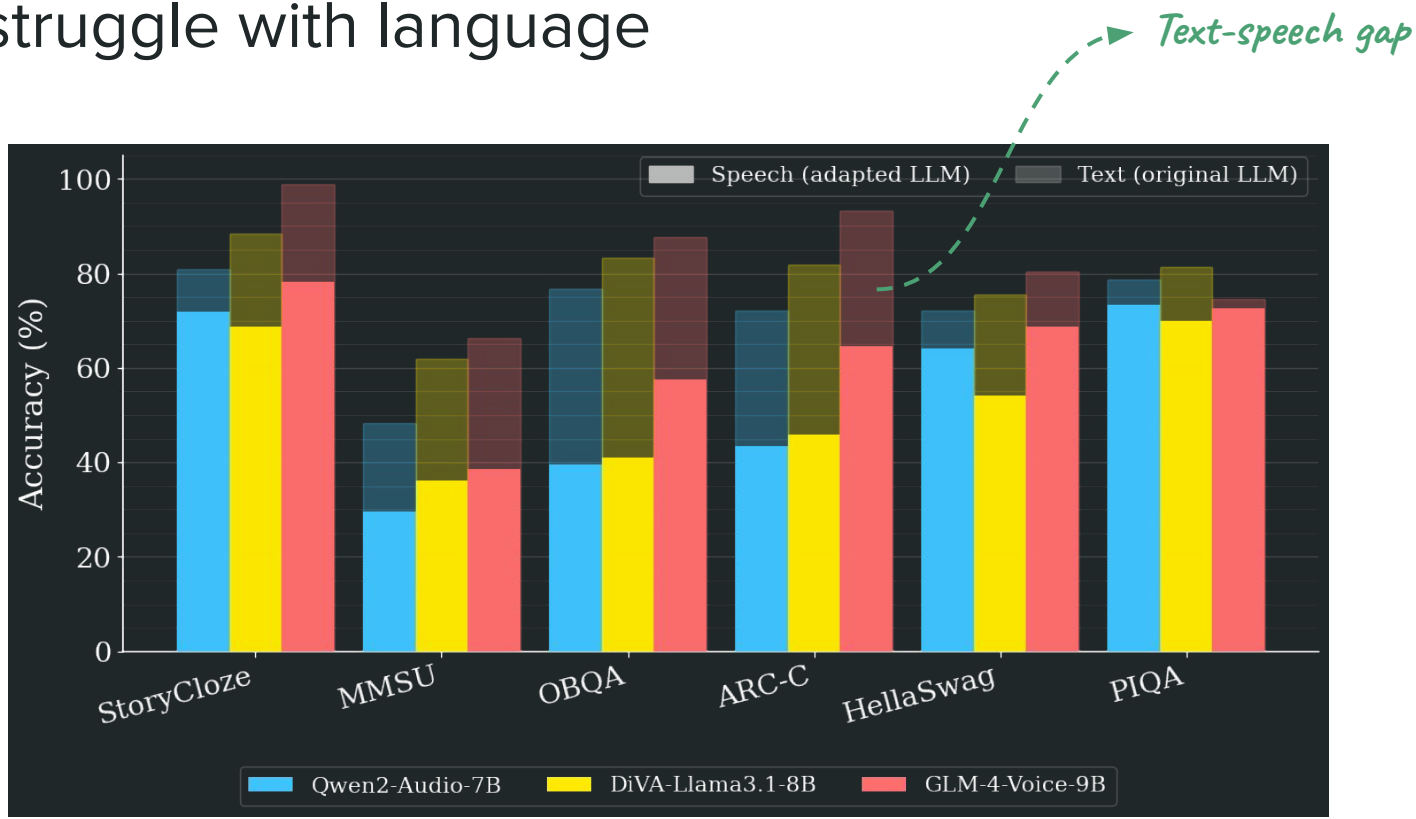


Figure 1: **a. The SPIRIT LM architecture.** A language model trained with next token prediction; tokens are derived from speech or text with an encoder, and rendered back in their original modality with a decoder. SPIRIT LM models are trained on a mix of text-only sequences, speech-only sequences, and *interleaved* speech-text sequences. **b. Speech-text interleaving scheme.** Speech is encoded into tokens (pink) using clusterized speech units (Hubert, Pitch, or Style tokens), and text (blue) using BPE. We use special tokens [TEXT] to prefix text and [SPEECH] for speech tokens. During training, a change of modality is randomly triggered at word boundaries in aligned speech-text corpora. Speech tokens are deduplicated and interleaved with text tokens at the modality change boundary. **c. Expressive Speech tokens.** For SPIRIT LM EXPRESSIVE, pitch tokens and style tokens are interleaved after deduplication.

Text-speech gap even with text pre-trained LLMs



SLMs struggle with language



Modality adaptation

Cuervo S, Moumen A, Labrak Y, Khurana S, Laurent A, Rouvier M, Woodland P, Marxer R. Late Fusion and Multi-Level Fission Amplify Cross-Modal Transfer in Text-Speech LMs. arXiv preprint arXiv:2503.06211. 2025 Mar 8.

How is information represented in LLMs?

LLMs exhibit two-phase abstraction dynamics

The geometry of hidden representations of large transformer models

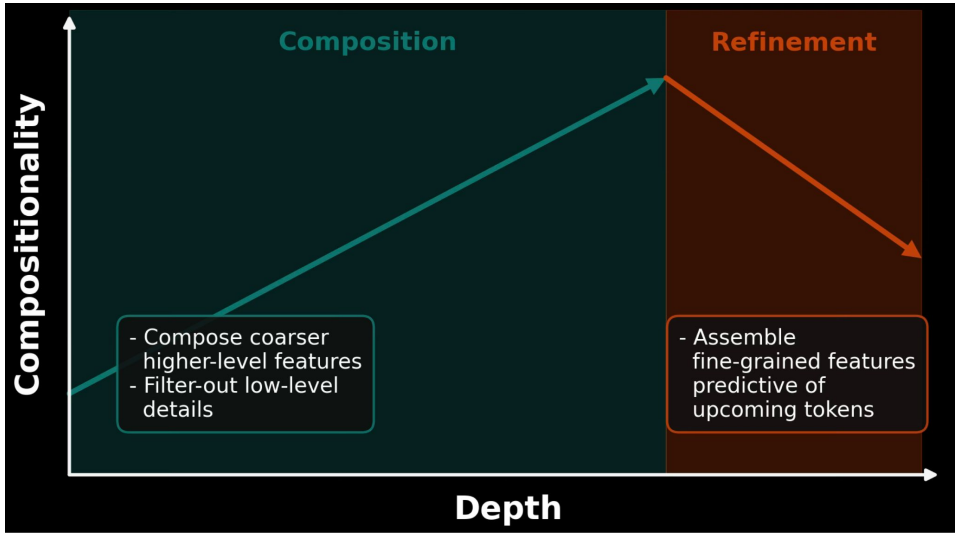
Lucrezia Valeriani[1,2*] Diego Doimo[1,3*] Francesca Cuturello[1]

Alessandro Laio[3,4] Alessio Ansuini[1†] Alberto Cazzaniga[1†]

[1] AREA Science Park, Trieste, Italy
[2] University of Trieste, Trieste, Italy
[3] SISSA, Trieste, Italy
[4] ICTP, Trieste, Italy

EMERGENCE OF A HIGH-DIMENSIONAL ABSTRACTION PHASE IN LANGUAGE TRANSFORMERS

Emily Cheng[1], Diego Doimo[2], Corentin Kervadec[1], Iuri Macocco[1], Jade Yu[3]
Alessandro Laio[4], Marco Baroni[1,5]
Universitat Pompeu Fabra[1], Area Science Park[2], University of Toronto[3], SISSA[4], ICREA[5]
emilyshana.cheng@upf.edu



How to improve modality transfer?

We aim to **re-utilize useful linguistic functions** from pre-trained text-LLMs to process non-textual modalities.

- **Abstraction dynamics across layers ➜** each layer functions have in/out spaces with a characteristic abstraction level.

  Hypothesis: **syncing abstraction dynamics** across modalities should lead to improved transfer

- **Perceptual modalities are more fine-grained** than text tokens

  Hypothesis: syncing **requires extra capacity for composition** and **refinement**.

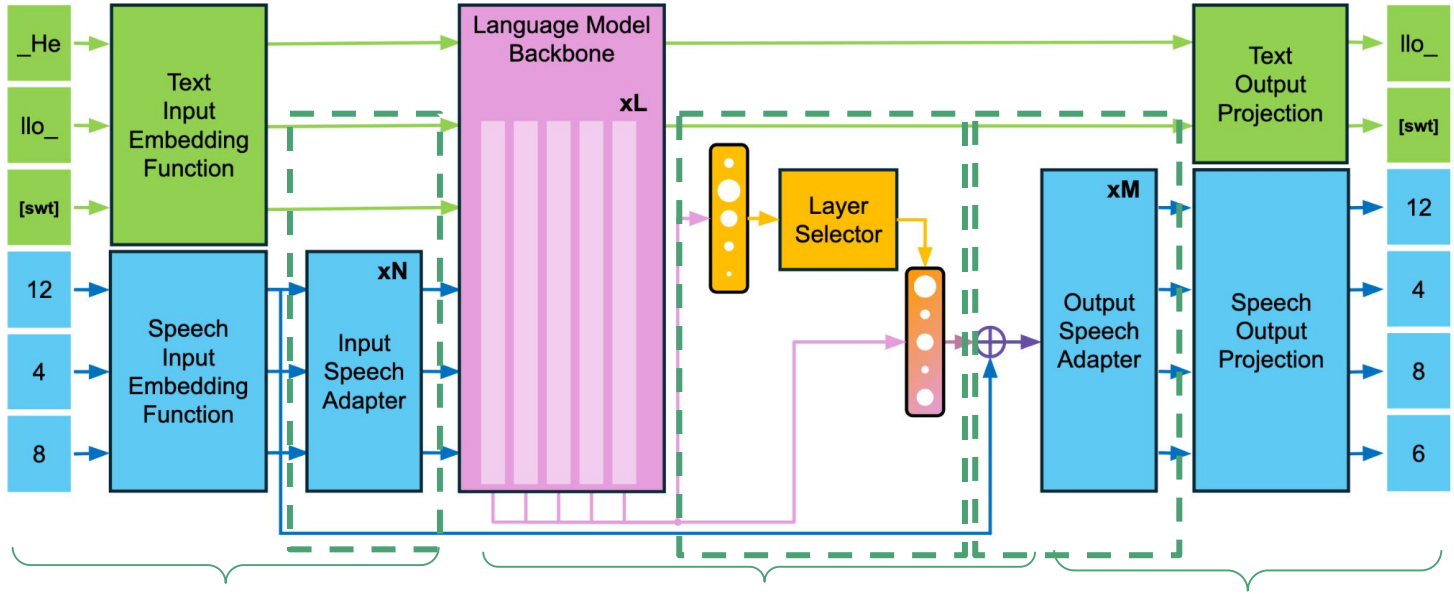
  Hypothesis: generating fine-grained signals requires both coarse-level features and fine details ➜ should **use representations at different abstraction levels**.

Our work

1. **Adapt a text LM** to generate **more fine-grained** language representations: **speech** tokens and **pixels**.
2. Add **additional capacity** (adapters) devoted to **composition** and **refinement**
3. Add the ability to use representations at **multiple abstraction levels**.
4. Analyze effects in terms of:
    a. **Compositionality** (intrinsic dimension of the dataset manifold).
    b. Cross-modal **representation space alignment**.
    c. **Performance** in knowledge and reasoning multimodal LM benchmarks.
5. Gain mechanistic insights through **informational probes**

Architectural changes to facilitate cross-modal transfer

Late Fusion and **Multi-Level Fission** Amplify Cross-Modal Transfer in Text-Speech LLMs



In adapter: Speech features are composed across additional speech-specific layers

Multi-level fission: The speech prediction attends to low and high-level features

Out adapter: additional layers are allocated to refine language-like features into speech features.

Experimental setup

Base model: SmolLM-360M (+ 135M and 1.7B for scaling experiments)

Modalities: speech and text images.

Data: ~110k hours of academic speech datasets, and rendered FineWiki-EN.

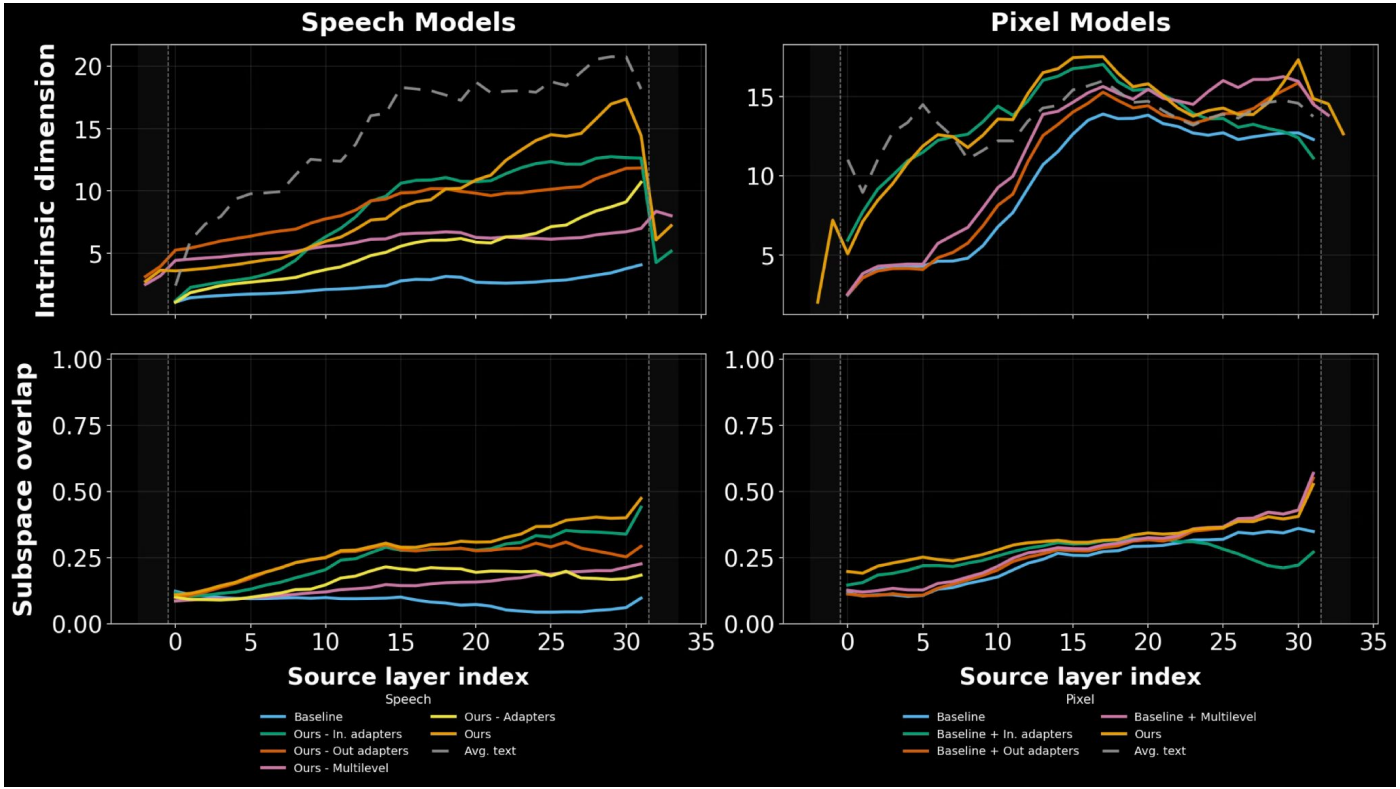
Baseline:

Standard LM fine tuning with:
early fusion (no input adapter) and early fission (no output adapter)

Our models:

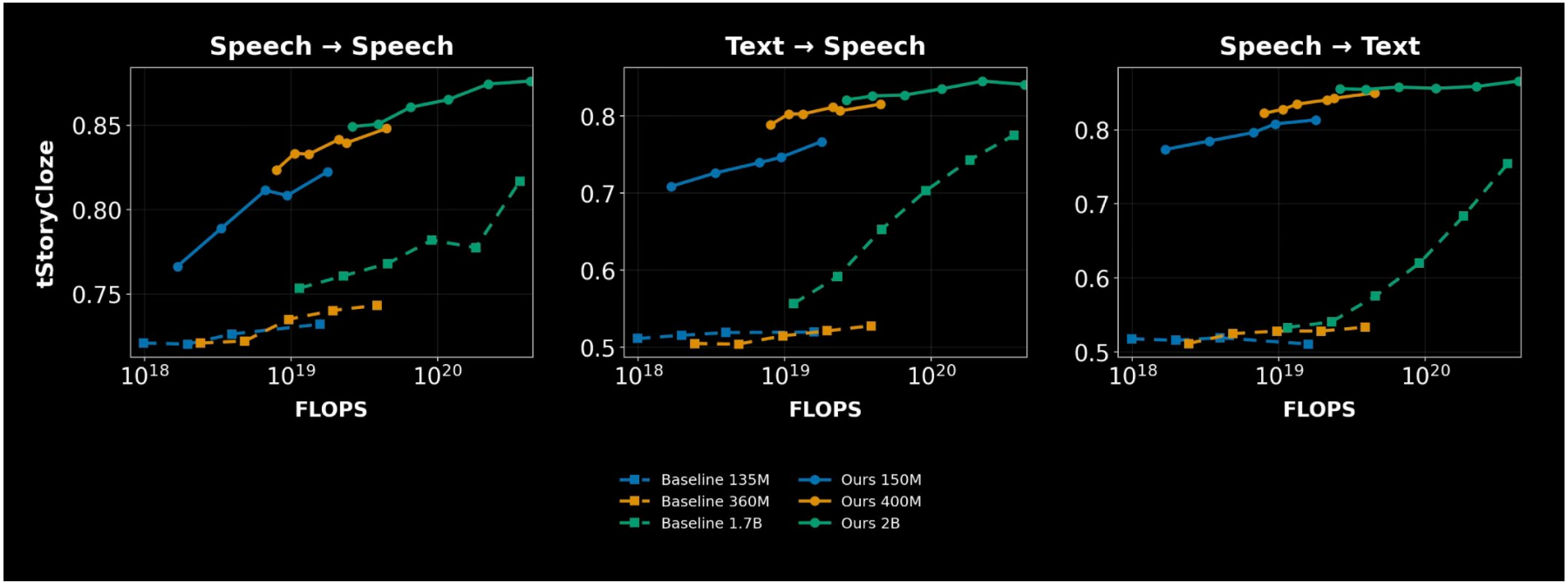
Baseline + **input adapter** + **output adapter** + multi-level representations

Compositionality and cross-modal alignment

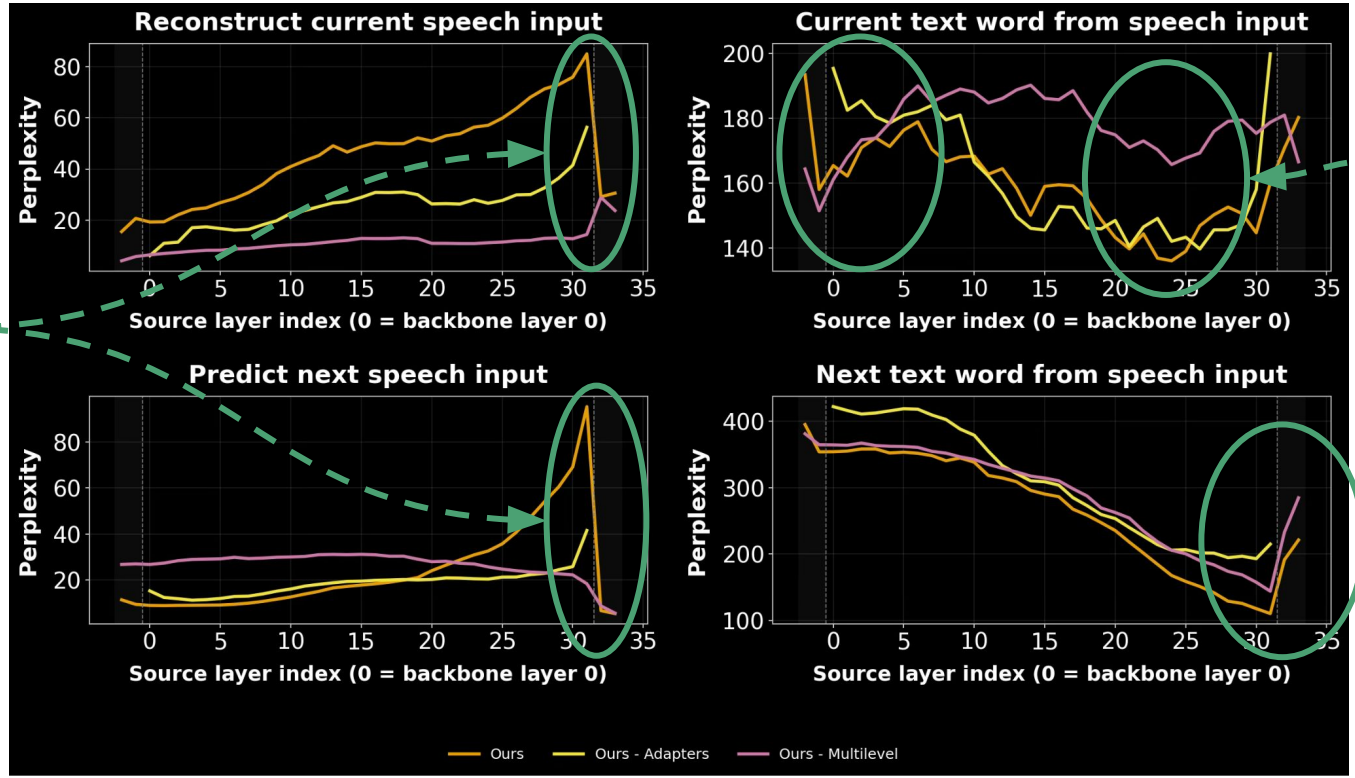


Our models outperform early fusion/fission baselines

Show cross-modal capabilities at smaller scales



Better usage of linguistic knowledge in backbone

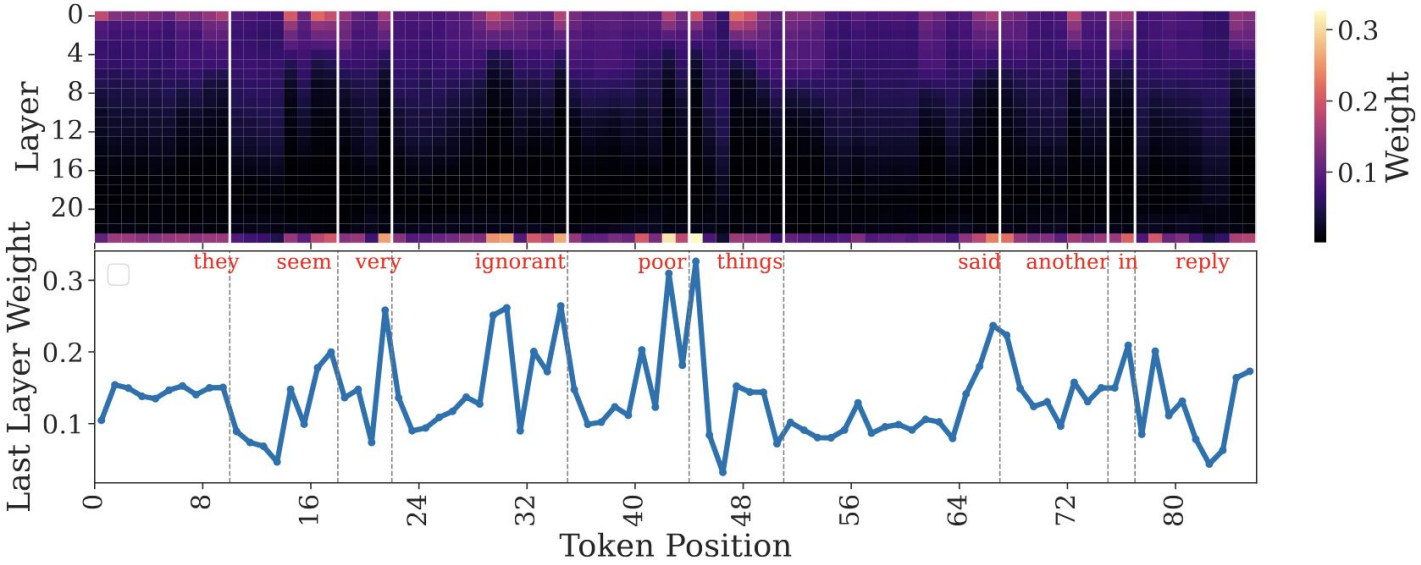


Low-level information **handled by adapters**

Word features **discovered early** *and* **better encoded**

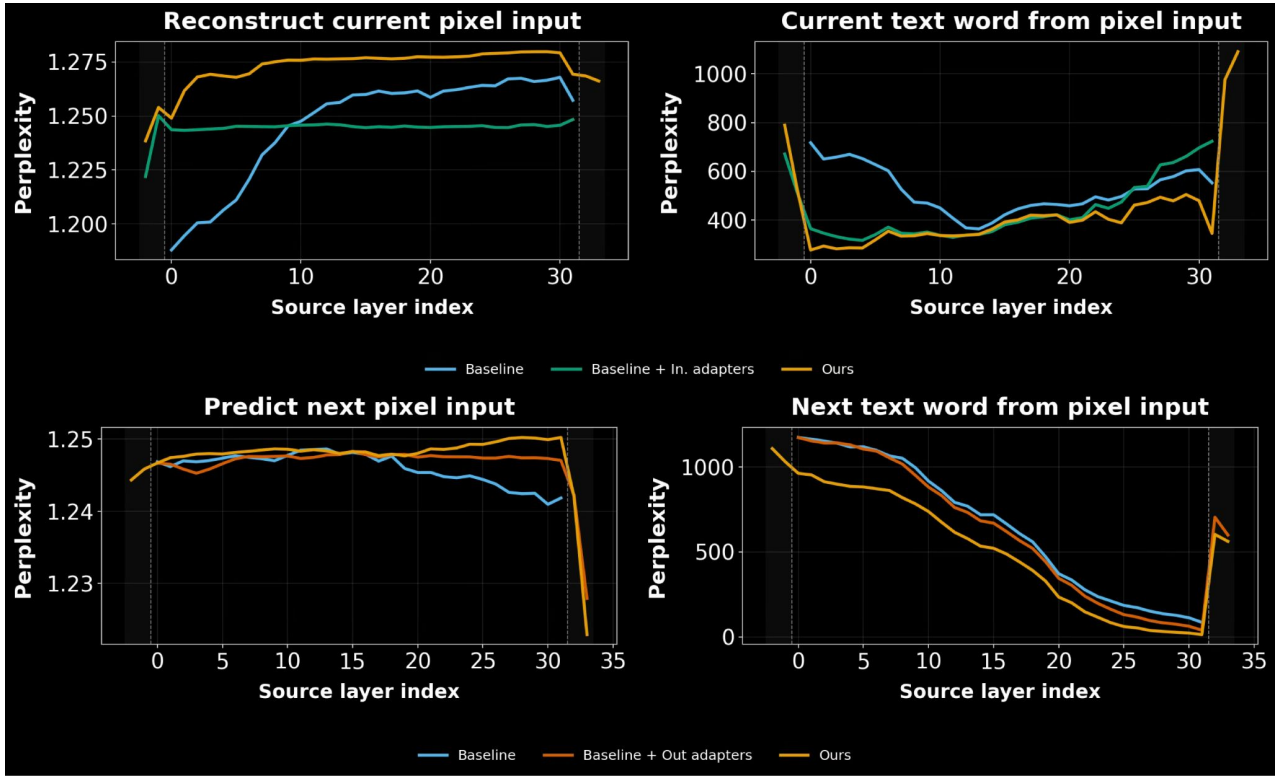
Final layer of backbone more **predictive of next word**

Attention patterns show the shift at word boundaries



Our model learns to use LLM's next-word predictive representations when most useful: close to word boundaries.

Same phenomena in rendered text



Our models achieve SotA-competitive performance...

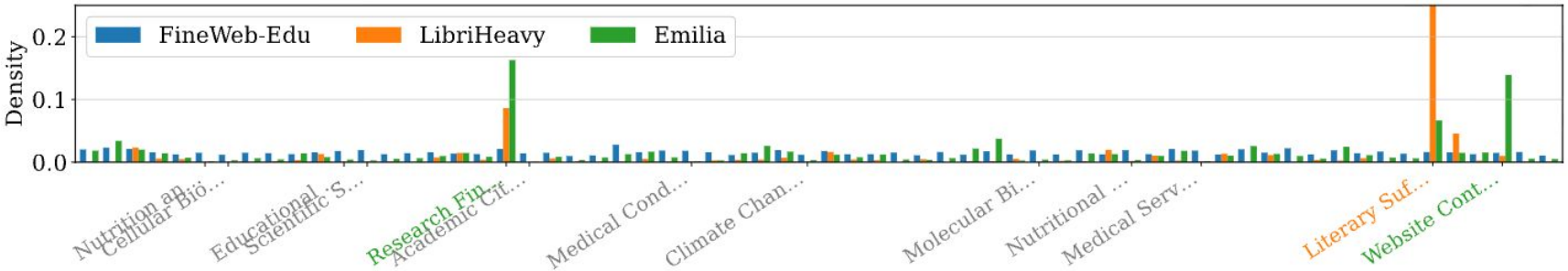
... with orders of magnitude less compute

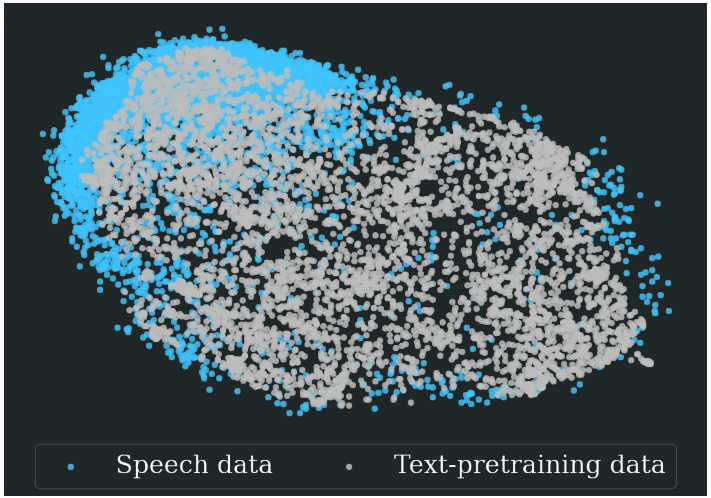
Model	Params.	Tokens	BLIMP		tStoryCloze				sStoryCloze				MMLU
			T	S	T	S	T→S	S→T	T	S	T→S	S→T	T (post/pre)
Previous TSLMs													
SPIRIT LM [Nguyen et al., 2024]	7B	∼175B	73.3	59.7	95.8	**90.5**	78.6	**94.3**	74.0	**66.3**	**64.7**	**71.7**	37.7 / 39.0
Moshi [Défossez et al., 2024]	7.7B	2.1T	—	58.8	—	83.0	—	—	—	60.8	—	—	49.8 / 54.3
GLM-4-Voice [Zeng et al., 2024] 1.5B	1.5B	1T	—	—	—	77.5	81.4	90.1	—	55.4	58.6	64.0	—
GLM-4-Voice [Zeng et al., 2024] 9B	9B	1T	—	—	—	83.0	**85.0**	**93.6**	—	**62.4**	63.2	**76.3**	—
Ours													
Baseline 1.7B	1.7B	16B	79.9	56.3	92.8	77.5	72.6	67.3	72.5	53.0	57.0	57.6	**40.0 / 40.0**
Baseline 1.7B	1.7B	32B	79.8	58.1	92.9	81.3	76.3	74.0	73.5	55.1	59.0	59.2	39.2 / 40.0
SMOLTOLK-150M	150M	16B	79.4	58.0	88.4	82.0	75.2	81.0	64.1	55.0	58.8	58.4	30.0 / 30.2
SMOLTOLK-400M	400M	16B	79.8	59.4	91.3	84.6	80.9	85.0	68.4	57.5	62.3	62.1	34.0 / 34.2
SMOLTOLK-2B	2B	16B	80.2	**61.4**	92.6	87.5	83.9	86.0	73.2	60.0	64.0	63.4	**40.0 / 40.0**
SMOLTOLK-2B	2B	32B	80.2	**61.9**	92.6	**87.6**	**84.3**	87.1	73.6	61.4	**64.2**	64.2	**40.1 / 40.0**

Cross-modal distillation

Cuervo S, Seto S, de Seyssel M, Bai RH, Gu Z, Likhomanenko T, Jaitly N, Aldeneh Z. Closing the gap between text and speech understanding in llms. arXiv preprint arXiv:2510.13632. 2025 Oct 15.

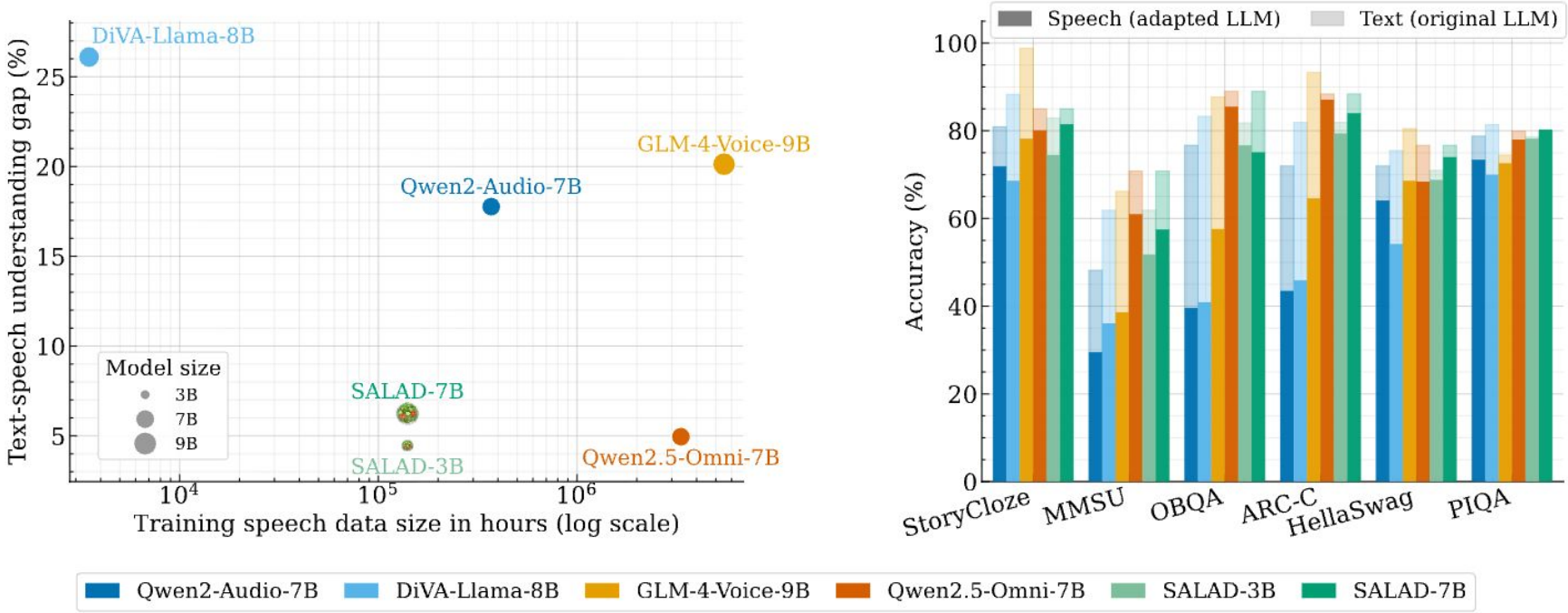
Why the gap? Domain mismatch





How to close the gap?

SALAD: **S**ample-efficient **A**lignment with **L**earning through **A**ctive selection and cross-modal **D**istillation



Check out Santiago's poster at ICLR 2026 !!

The Fourteenth International Conference on Learning Representations

ICLR 2026

Rio de Janeiro, Brazil Apr 23 2026 https://iclr.cc/Conferences/2026 program-chairs@iclr.cc

Please see the venue website for more information.
Submission Start: Sep 01 2025 11:59AM UTC-0, Abstract Registration: Sep 20 2025 11:59AM UTC-0, Submission Deadline: Sep 25 2025 11:59AM UTC-0

Accept (Oral) | Accept (Poster) | Conditional Accept (Oral) | Conditional Accept (Poster) | Reject | Withdrawn Submissions | Desk Rejected Submissions | Recent Activity

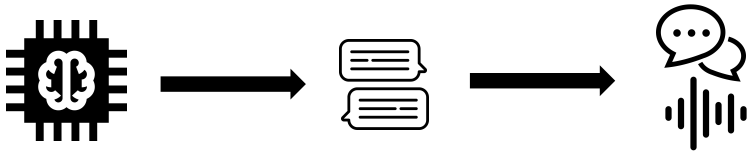
Closing the Gap Between Text and Speech Un

Closing the Gap Between Text and Speech Understanding in LLMs

Santiago Cuervo, Skyler Seto, Maureen de Seyssel, Richard He Bai, Zijin Gu, Tatiana Likhomanenko, Navdeep Jaitly, Zakaria Aldeneh

Published: 26 Jan 2026, Last Modified: 26 Feb 2026 ICLR 2026 Poster Readers: Everyone

Show details



What about conversations for domains with little to no data?

Tackled this question with the Play your Part team at JSALT 2025

Can we use role-playing LLMs to synthesize conversations in specific domains?

Realism

Domains or **scenarios**?

Authenticity linguistic, consistency, flow, clinical relevance

Acoustic environment noises, ambiance, room acoustics

Diversity

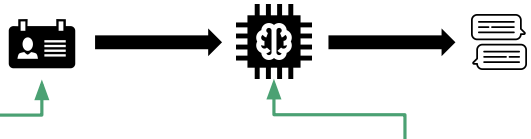
LLMs to sample scenarios/personas

Relation between **generations and persona** definitions

Controllability

Prompt design and **context** usage to condition generation

Activation steering to modulate persona



A Systematic Survey of Prompt Engineering in Large Language Models: Techniques and Applications

Pranab Sahoo[1], Ayush Kumar Singh[1], Sriparna Saha[1], Vinija Jain[2,3*], Samrat Mondal[1] and Aman Chadha[2,3*]

[1]Department of Computer Science And Engineering, Indian Institute of Technology Patna

[2]Stanford University, [3]Amazon AI

STEERING LANGUAGE MODELS WITH ACTIVATION ENGINEERING

Alexander Matt Turner, Independent researcher, alex@turntrout.com

Lisa Thiergart, MIRI

Gavin Leech, University of Bristol, g.leech@bristol.ac.uk

David Udell, Independent researcher

Juan J. Vazquez, Arb Research

Ulisse Mini, MATS

Monte MacDiarmid, Anthropic

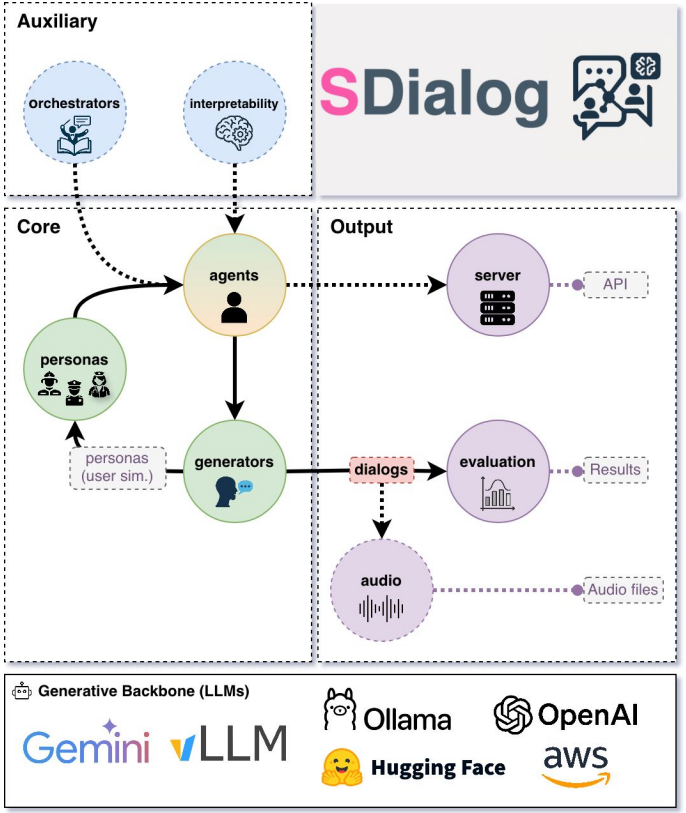
SDialog: agent-based dialog generation and analysis

Burdisso, S., Baroudi, S., Labrak, Y., Grunert, D., Cyrta, P., Chen, Y., Madikeri, S., Villatoro-Tello, E., Marxer, R. and Motlicek, P., 2025. SDialog: A Python Toolkit for End-to-End Agent Building, User Simulation, Dialog Generation, and Evaluation. (to appear)

EACL 2026
RABAT • MOROCCO
Mars • March 24-29, 2026 • مارس

Open-source library for generating and analysing agent-based dialogs

- **Standard dialog schema**
- **Persona-driven multi-agent simulation**
- **Composable orchestration**
- **Built-in evaluation**
- **Native mechanistic interpretability**
- **Easy creation of user-defined components**
- **Interoperability** across multiple LLMs and APIs
- **Audio generation**



Simple API for full pipelines

```python
from sdialog.agents import Agent
from sdialog.personas import SupportAgent

# Defining three tools
def verify_account(customer_id):
  ...

def update_address(customer_id, address):
  ...

def get_service_plans():
  ...

# Defining a persona for the agent
support_persona = SupportAgent(
  name="Michael",
  politeness="high",
  rules="Make sure to always verify the
     account when required"
)

# A function to get the agent given an LLM
def build_my_agent(llm_name) -> Agent:
  agent = Agent(
    persona=support_persona,
    think=True,
    tools=[verify_account,
           update_address,
           get_service_plans],
    context="Call center office",
    name="Support Agent",
    model=llm_name
  )
  return agent
```

```python
def generate_dialogs(llm_name, customer,
                     n, save_folder="."):

  agent = build_my_agent(llm_name)

  customer = Agent(
    persona=customer,
    name="Customer"

  )

  for ix in range(n):
    dialog = agent.talk_with(customer)
    dialog.to_file(
        f"{save_folder}/dialog_{ix}.json"
    )
```

Audio rendering

```
# 1. Init TTS engine
tts_engine = @{\className{KokoroTTS}}@()

# 2. Init voice database from HF dataset
voice_db = @{\className{
     HuggingfaceVoiceDatabase}}@(
    "sdialog/voices-kokoro"
)

# 3. Generate the audio dialogue
to_audio(
    dialog=my_dialog,
    tts_engine=tts_engine,
    voice_database=voice_db,
    dir_audio="./outputs_audio"
)
```

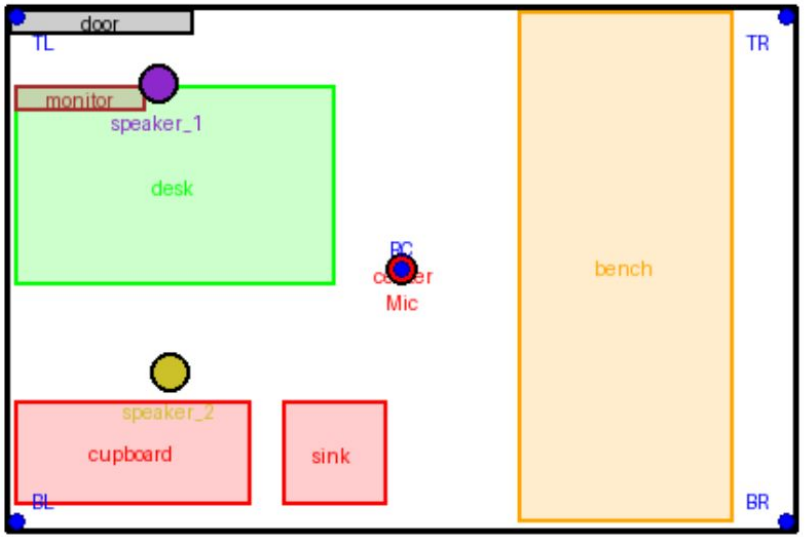


Figure 5: A procedurally generated room layout for an American-style hospital examination room.

Interpretability and activation steering

```python
# Harmful instructions loop
for harmful, harmless in requests :
    agent(harmful)
    x = inspector_x.input[0][
     post_instruct_idx]
    harmful_reps.append(x)
    # Same for harmless
    ...

mu = harmful_reps.mean(dim=0)
v = harmless_reps.mean(dim=0)
```

The refusal direction , defined as :

$$\mathbf{r}_i^{(l)} = \boldsymbol{\mu}_i^{(l)} - \boldsymbol{v}_i^{(l)} \tag{5}$$

can be translated, in the case of SDialog, to :

```python
# Get the direction
r = mu - v

# Optional : Save the direction
torch.save(r, "refusal_direction.pt")
```

Refusal in Language Models Is Mediated by a Single Direction

Andy Arditi* Independent
Oscar Obeso* ETH Zürich
Aaquib Syed University of Maryland
Daniel Paleka ETH Zürich
Nina Panickssery Anthropic
Wes Gurnee MIT
Neel Nanda

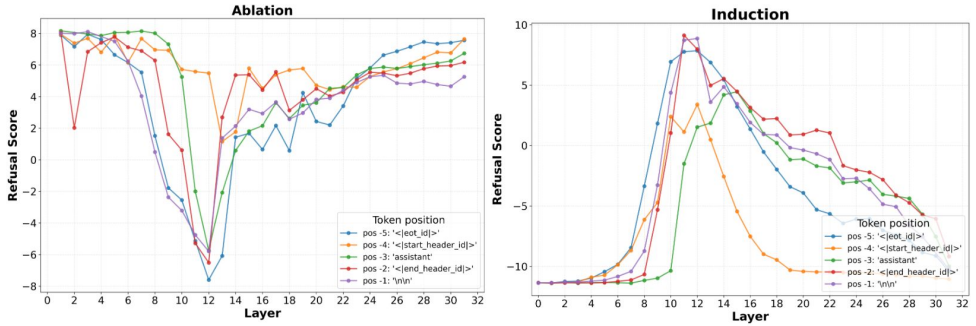


Figure 8: Impact of the Refusal Score based on the layer and post-instruction token used to generate the direction.

Applications to DoPaCo: doctor-patient conversations

Labrak, Y., Grünert, D., Baroudi, S., Chun, J., Cyrta, P., Burdisso, S., Hassoon, A., Liu, D., Rothschild, A., Van Deusen, R., Motlicek, P., Perrault, A., Marxer, R. and Schaaf, T. (2026) *Generating Synthetic Doctor-Patient Conversations for Long-form Audio Summarization*. Under submission.

Baroudi, S., Labrak, Y., Kumar, S., Kalda, J., Burdisso, S., Cyrta, P., Alvarez-Trejos, J.I., Motlicek, P., Bredin, H. and Marxer, R. (2026) Doctor or Patient? Synergizing Diarization and ASR for Code-Switched Hinglish Medical Conditions Extraction. arXiv preprint arXiv:2603.06373. Available at: https://arxiv.org/abs/2603.06373

Applications to DoPaCo: doctor-patient conversations

Baroudi, S., Labrak, Y., Kumar, S., Kalda, J., Burdisso, S., Cyrta, P., Alvarez-Trejos, J.I., Motlicek, P., Bredin, H. and Marxer, R. (2026) Doctor or Patient? Synergizing Diarization and ASR for Code-Switched Hinglish Medical Conditions Extraction. arXiv preprint arXiv:2603.06373. Available at: https://arxiv.org/abs/2603.06373

Synthesizing long-form persona-based doctor-patient conversations

"name": "Sharifa FIRESTONE"
"age": 44
"height": 175
“weight": 59
"race": "Honduran"
"gender": "Female”
"insurance": "Medicare"

"forgetfulness": "Perfect recall for details"
"formality": "Neutral"
"hurriedness": "Relaxed and not feeling pressured"
"openness": "Close-minded, judgmental and condescending",

"language": "Fluent English"

"occupation": "Tour and Travel Guides"
"marital_status": "Divorced"
"reason_for_visit": "Sudden Bilateral Leg Weakness After a Cold"

Basic *observable* information
+ Avoid bad surprises
+ Clinical diversity

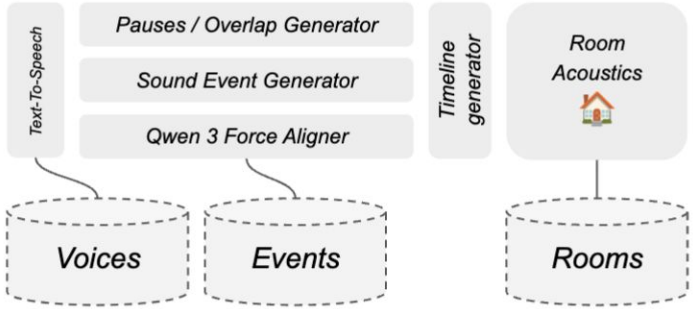
Personality
+ Linguistic diversity
+ Interpersonal challenge

Language
+ Linguistic diversity

Background and reason for visit
+ Clinical and linguistic diversity

Synthesizing long-form doctor-patient conversations and SOAP notes

Dataset	Doctor		Patient		Num Turns
	Turn Length	Fog Index	Turn Length	Fog Index	
Ours	49.9 ± 16.0	10.4	56.0 ± 15.2	6.9	28.4 ± 7.5
PriMock57	18.8 ± 4.7	6.6	12.3 ± 5.8	5.5	97.3 ± 17.7
Mocks	29.6 ± 6.3	7.0	13.6 ± 3.1	7.0	54 ± 6.2



	Train	Dev	Test	Total
Personas (Doc/Pat)	60/120	20/20	20/60	100/200
Dialogues	7,200	400	1,200	8,800
Hours	1,087	59	183	1,329
Words in dialogues	10.9M	582K	1.8M	13.3M
Turns/dialogue	28.4	28.7	28.0	28.4
Duration/dialogue (s)	544	530	548	544
Audio events/dialogue	37.7	36.7	36.5	37.5
Words per SOAP note	N/A [7]	325.1	324.2	N/A [7]

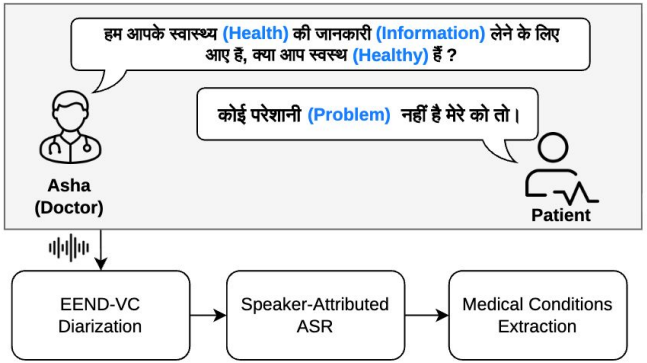
Data will be made available soon.

Contact Thomas Schaaf for more info!

DISPLACE 2026 CHALLENGE

DISPLACE-M

Diarization and Speech Processing for LAnguage understanding in Conversational Environments



Robust Diarization

- End-to-End Neural Diarization with Vector Clustering (EEND-VC)

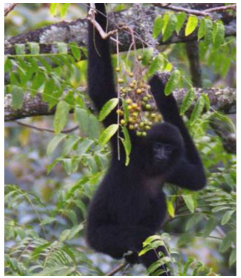
Targeted SA-ASR

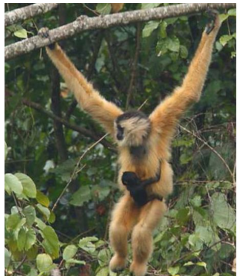
- Qwen3-ASR model → domain-specific fine-tuning → Devanagari script normalization → LLM post-correction

Medical Conditions Extraction

- Benchmarked open-source LLMs VS proprietary End-to-End audio models to establish the current performance ceiling.

Fully open-source cascaded pipeline that won **1st place out of 25** teams in the DISPLACE-M Challenge! 🏆





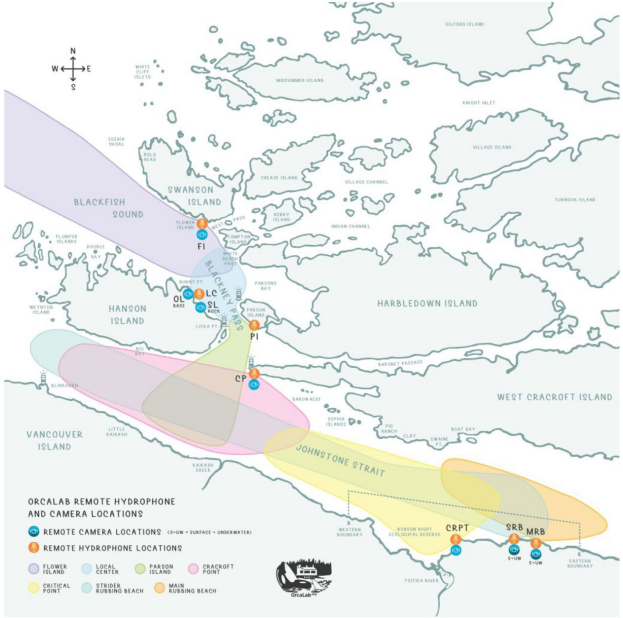
What about non-human “conversations”?

*with **Paul Best** et al.*

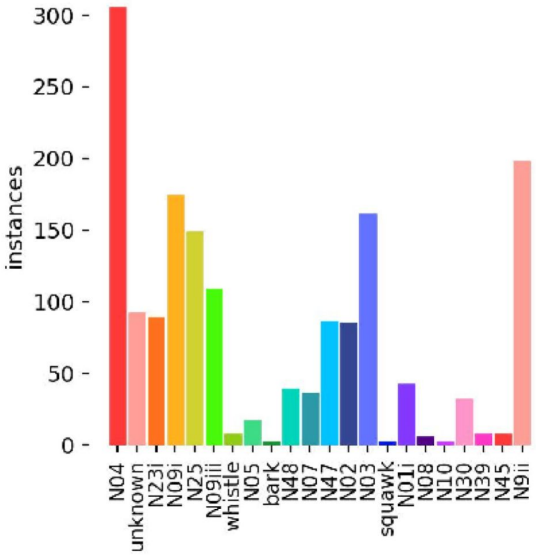
Orcas pod sizes and vocal complexity

Best, P., Poupard, M., Marxer, R., Spong, P., Symonds, H., & Glotin, H. (2025). Analysing vocal complexity in relation to sociality in orcas of British Columbia: An application of long-term computational passive acoustics. Ecological Informatics, 90, 103211.

5 years of recordings, annotations and observations



Map of the area and listening range for the **7 hydrophones** of the OrcaLab station



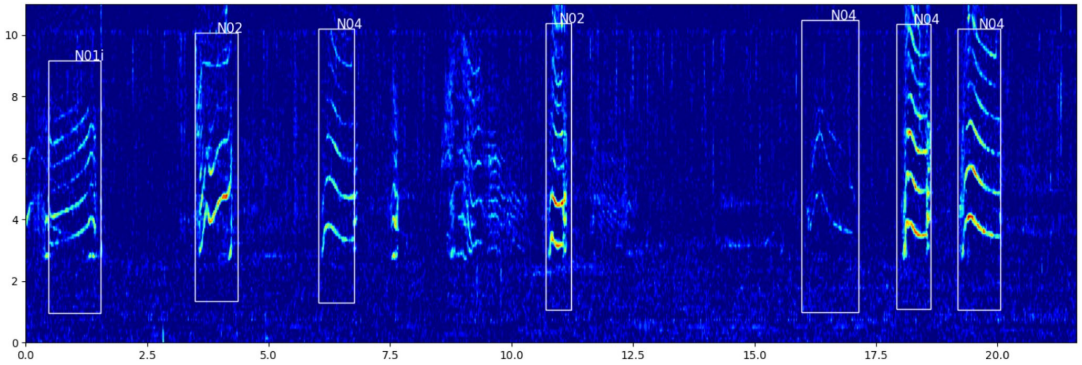
Distribution of annotations by **call type** in the training set

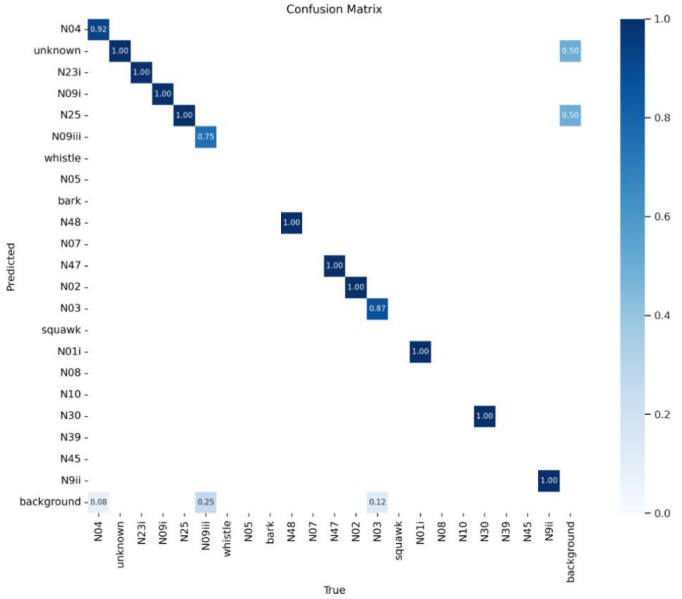
Table 2. Data volumes for acoustic detections that were attributed a pod from the visual observations of 2016.

	Sequences	Passages
A1	320	75
A4	15	4
A5	98	23
A5-A4	16	4

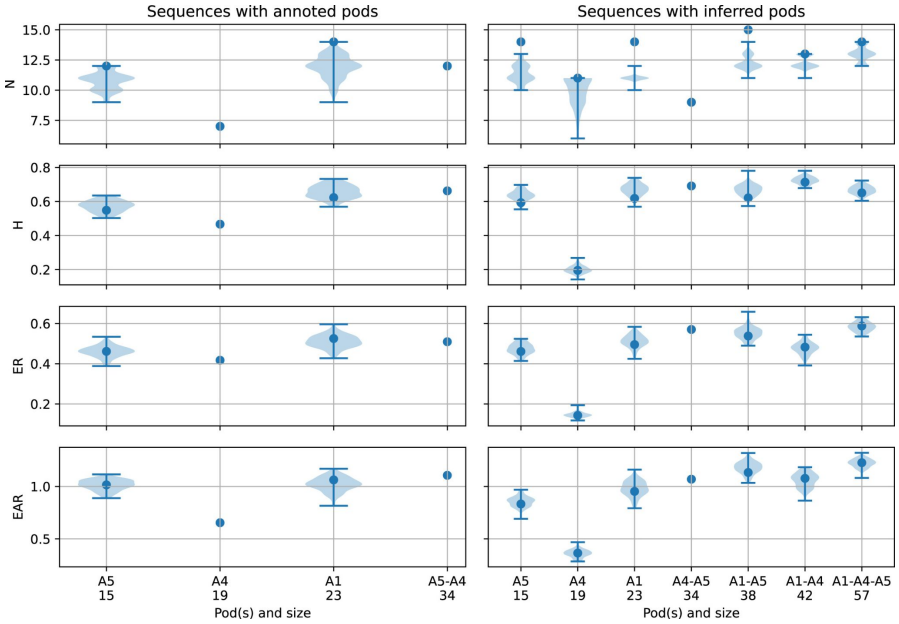
Attribution of sequences to **pods**

Classification of call types

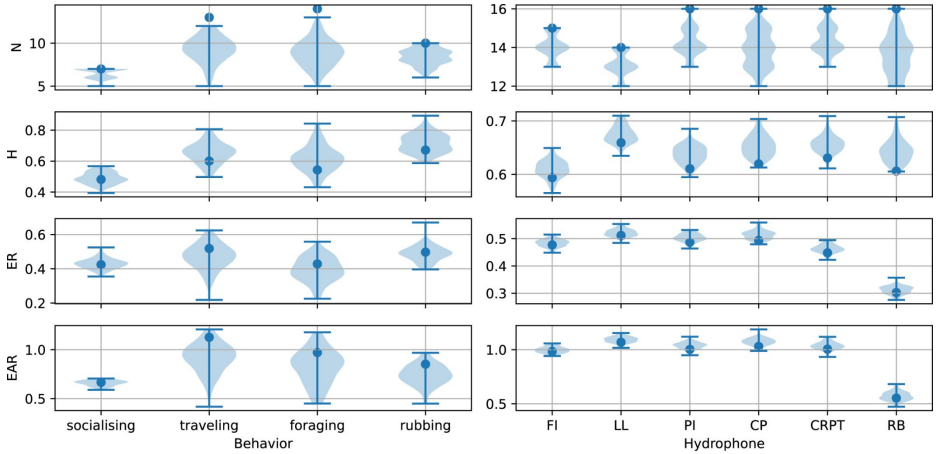




Vocal vs social complexity



Pod size

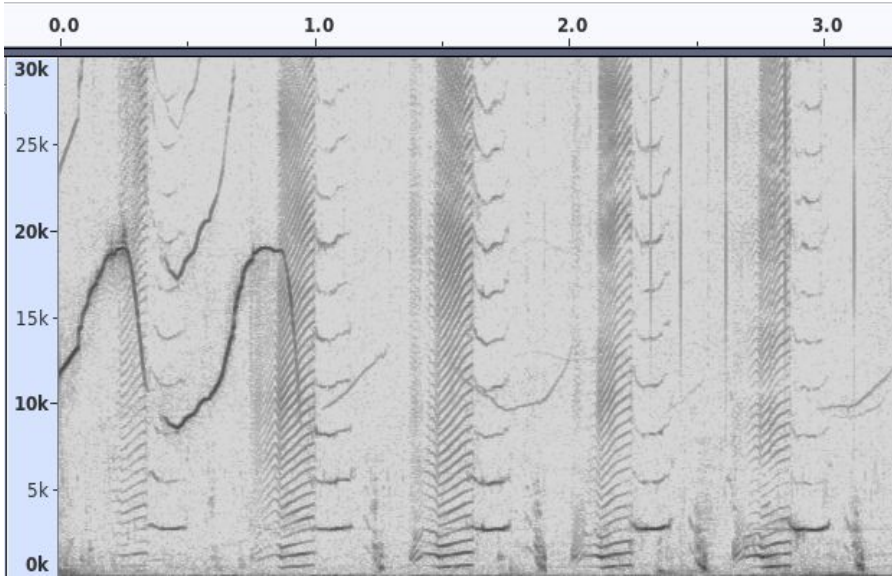


Behavior

Pilot whale diarization

Best, P., Habib-Dasetto, L., Cauzinille, J., Marxer, R., Delfour, F., Legout, T., & Montant, M. (2025, September). Underwater behaviour of Globicephala melas. In IBAC 2025.

Long-finned pilot whales (*Globicephala melas)*



Courts et al. 2020. Australian long-finned pilot whales (*Globicephala melas*) emit stereotypical, variable, biphonic, and sequenced vocalisations. *Scientific Reports*
Zwamborn and Whitehead 2017. Repeated call sequences and behavioural context in long-finned pilot whales off Cape Breton, Nova Scotia, Canada. *Bioacoustics*

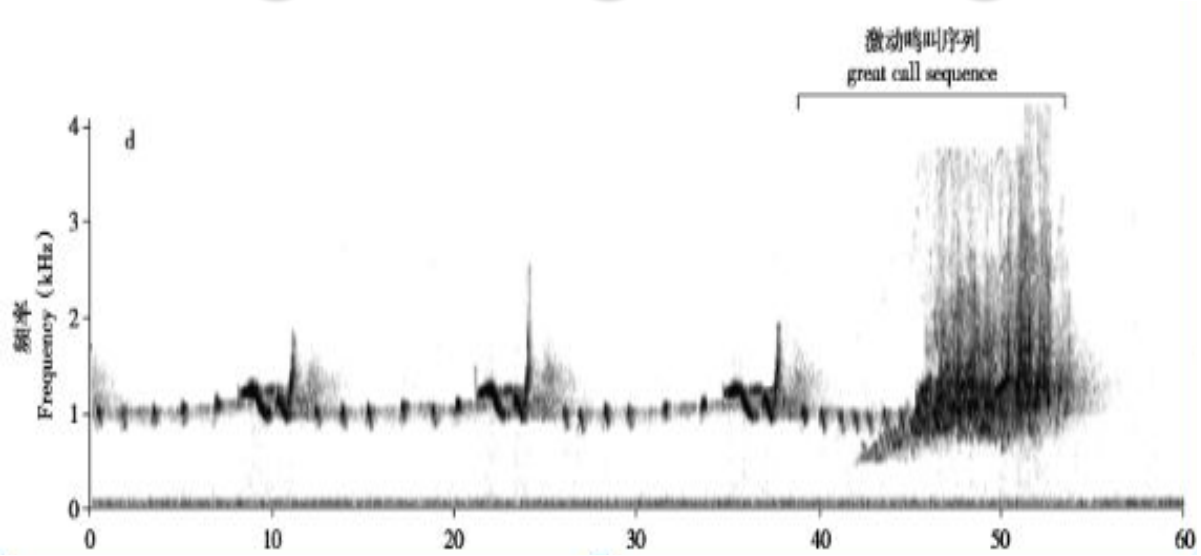
Spatialised bioacoustics - near field

Ferrari et al. 2020. 3D diarization of a sperm whale click cocktail party by an ultra high sampling rate portable hydrophone array for assessing individual cetacean growth curves. In: *Forum Acusticum, Lyon, France.*

Gibbons and territories

Best, P., Dassow, A., Kershenbaum, A., Nguyen, T. D., Pogson, M., Maheshwari, A., & Marxer, R. (2026). Spatial validation of acoustic individual identification models without ground truths: a case study with the cao-vit gibbon population. PeerJ, 14, e20655.

Eastern black-crested gibbons (*Nomascus nasutus)*



Considered extinct until 2002

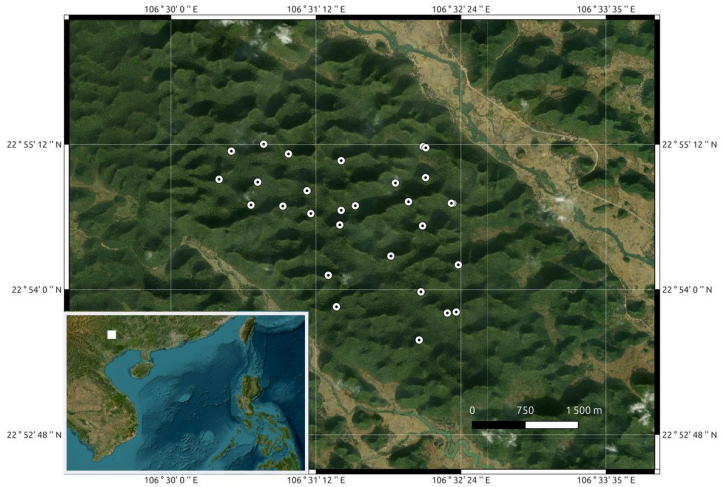
Territorial and vocal

Endemic to the China-Vietnam border region

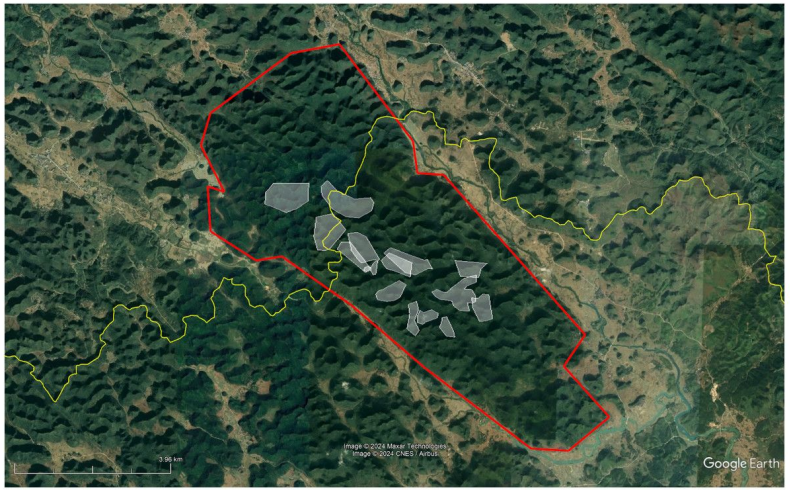
Junjuan et al. 2013. Call sonograms of eastern black crested gibbon (*Nomascus nasutus*). *Acta Theriologica Sinica*
Mootnick et al. 2011. A comparative study of crested gibbons (*Nomascus*). *American Journal of Primatology*

Spatialised bioacoustics - far field

System of PAM recorders

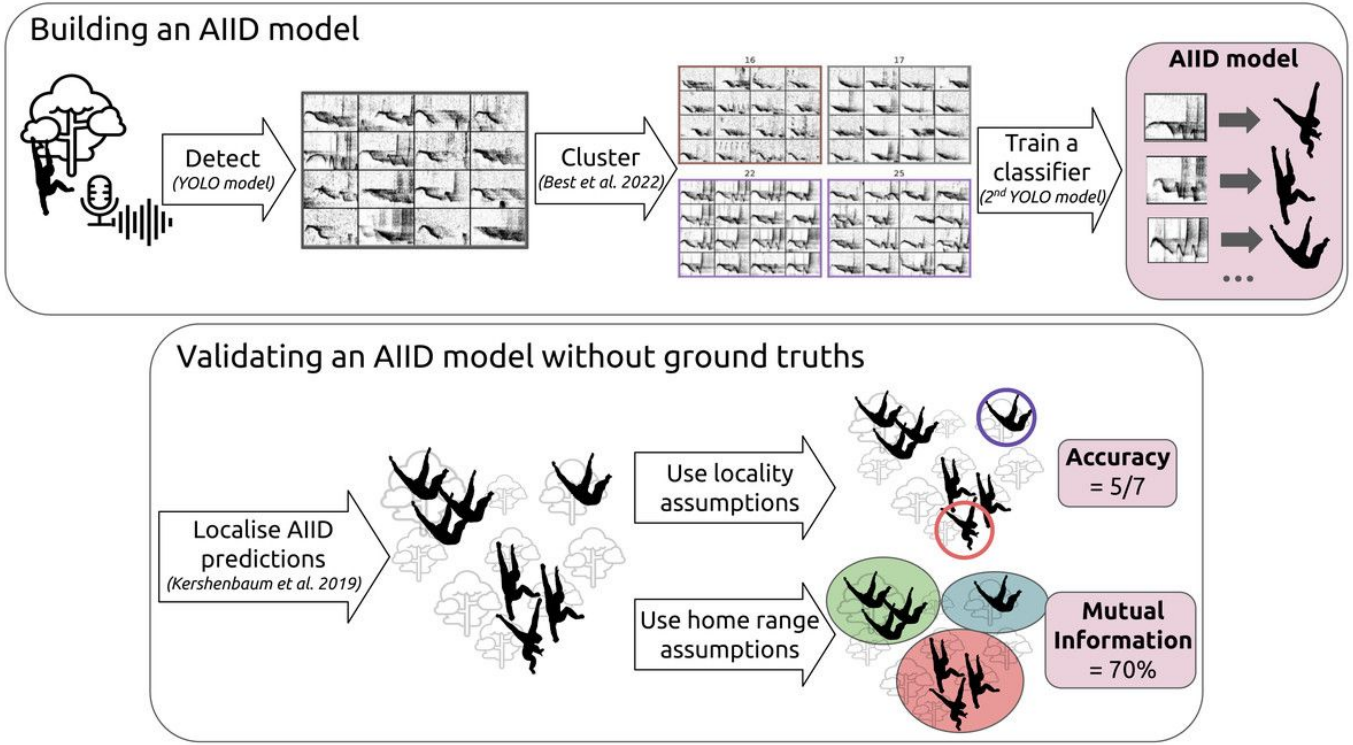


Home ranges from visual obs



Kershenbaum et al. 2019. Tracking cryptic animals using acoustic multilateration: A system for long-range wolf detection. *The Journal of the Acoustical Society of America*

How to validate that our acoustic individual ID is correct?



Kershenbaum et al. 2019. Tracking cryptic animals using acoustic multilateration: A system for long-range wolf detection. *The Journal of the Acoustical Society of America*