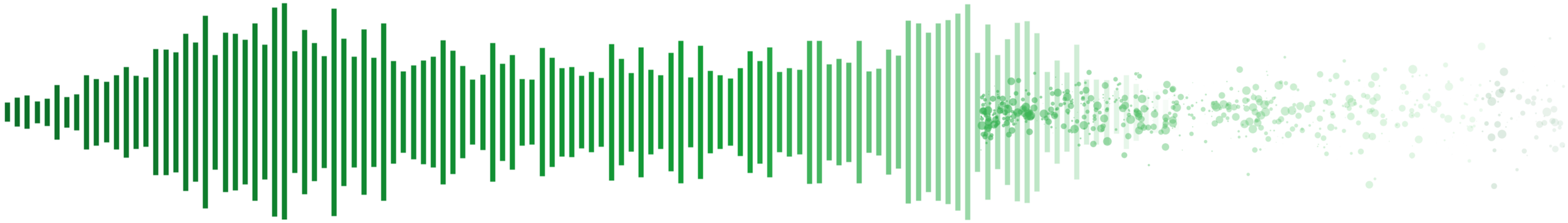
AI for Humans

Making conversation with AI as natural as talking to a person



BosonAI

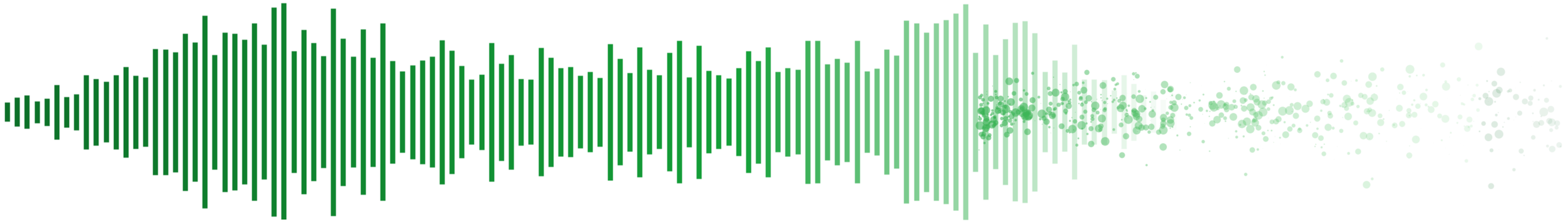
ALEX SMOLA · KDD 2026

Show and tell

Building it
Evaluation
Picking good benchmarks

1

Voice and conversation with Higgs Audio



BosonAI

Voice Cloning and Generation

BosonAI

Audio Understanding

Where could the loud and deep purring be originating from?

Choices:

- A zoo
- A playground
- A veterinary clinic
- A forest

Correct Answer:
A veterinary clinic

Chain of Thought:

PLANNING

The user wants to determine the source of a loud and deep purring sound heard in the audio clip. To answer this, I need to carefully listen to the audio and identify the characteristics of the purring sound. I will then consider the possible environments where such a sound could originate from and match it with the given choices.

CAPTIONING

After listening to the audio clip, I hear a continuous, deep, and resonant purring sound. The purring is consistent and has a low frequency, which gives it a rumbling quality. There are no other sounds present in the background, making the purring the dominant and only audible element in the clip.

REASONING

Okay, so the question is asking where the loud and deep purring sound could be coming from. The choices are a zoo, a playground, a veterinary clinic, or a forest. Let me think through this step by step.

First, I need to consider what kind of animal could produce a loud and deep purring sound. From what I know, cats are known for purring, but typically, their purring is not described as loud or deep. It's usually a softer, more consistent sound. However, big cats like lions or tigers are said to have deeper purring sounds. But are these sounds typically described as loud? Maybe in a zoo setting, where such large cats might be present, but it's not common to hear them purring loudly all the time.

Then there's the veterinary clinic. If there's a large cat, like a tiger or lion, in a clinic, especially one that deals with exotic animals, they might produce a deep purring sound, especially if they're relaxed or content. A veterinary clinic that handles big cats would be a plausible place. But would the purring be described as loud? It could be, depending on the size of the cat and the environment. A playground is an unlikely place. There are usually children, equipment like swings and slides, and maybe some birds or squirrels, but not big cats. So that can be ruled out.

A forest... Big cats like lions and tigers are not typically found in forests. They have specific habitats like savannas or jungles. So a forest is not a likely source.

So between a zoo and a veterinary clinic, which one is more likely? Zoos do have big cats, but they are often spread out, and the sounds might not be as prominent unless you're very close. A veterinary clinic that specializes in exotic animals would have these cats in more controlled environments, possibly closer to the recording device, making the purring more audible and deeper. Also, in a clinic, the focus is on the animals, so the sound might be more noticeable.

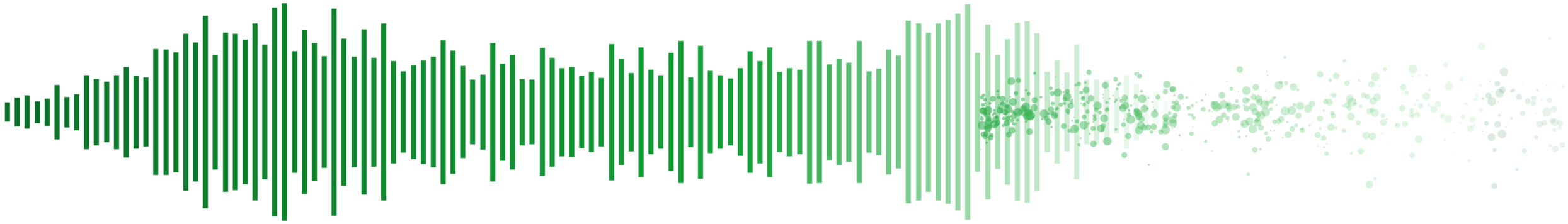
Therefore, the most logical answer is a veterinary clinic.

SUMMARY

The loud and deep purring sound in the audio clip is most likely originating from a veterinary clinic. This is because such a sound could be produced by a large cat like a tiger or lion, which is more likely to be found in a controlled environment like a clinic that specializes in exotic animals. The purring is consistent and deep, suggesting a relaxed and content state of the animal, which is more probable in a veterinary setting where the animal is being cared for. In contrast, a zoo or a forest is less likely to produce such a prominent and deep purring sound, and a playground is an unlikely environment for such an animal to be present.

BosonAI

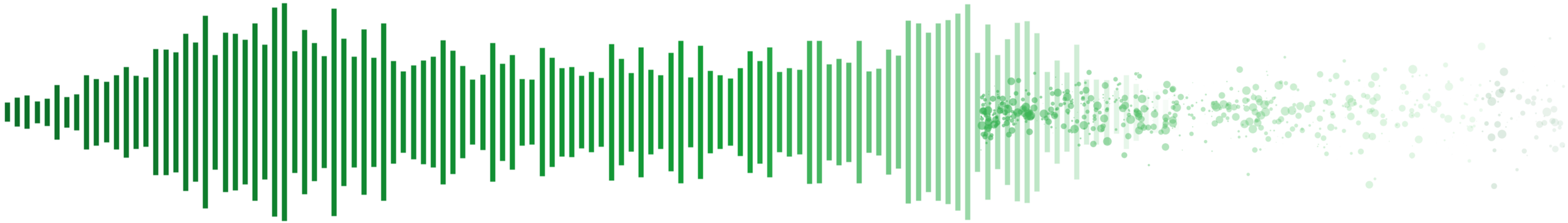
Outbound Sales Agent (2025)



BosonAI

Live Demo (2026)

www.boson.ai/workspace/voice-studio



BosonAI

Demo Features

SERVING

First audio in under a second

Everything streams — STT, generation, TTS. 750 ms median in production; 1.16 s at p90.

REALTIME

Recovery from interruption

Direct speech-to-speech, no transcription step. IHBench measures workflow recovery.

INSTRUCTABILITY

On topic under provocation

The agent holds its interview rubric and compliance constraints when pushed off-script.

TOOL USE

Web search mid-conversation

The model issues tool calls during dialogue and folds results into its next spoken turn.

MULTILINGUAL

Switched language mid-interview

TTS covers 100+ languages, STT covers 94 — switching happens inside one conversation.

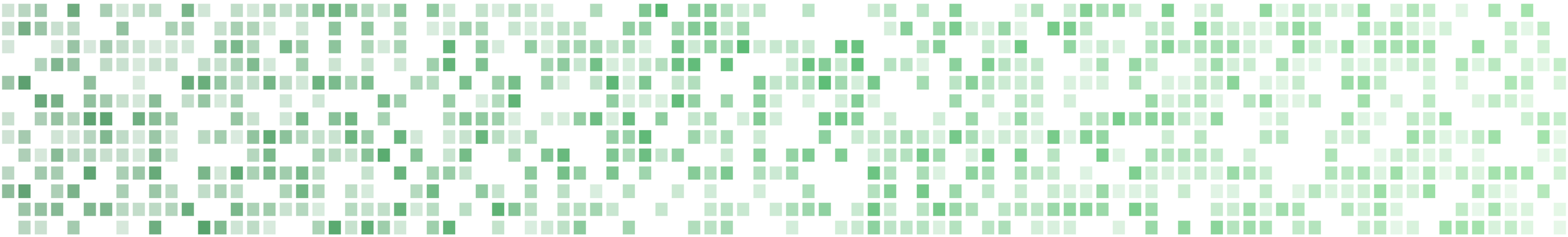
Show and tell

Building it

Evaluation

Picking good benchmarks

2



BosonAI

Pipeline

DATACENTER

Store everything
Compute for processing

- 30 PB disks
- 1000 GPUs
- 175TB RAM
- 750Tbit Switching

DATA

Lifetimes of Audio

- 100M hours trainable audio
- Small models built on 10M
- Extract & annotate (emotions, speakers, language)

TOKENIZER

Converting audio

- Acoustic fidelity
- Semantic information
- Low token frequency (for cost)

MODEL

LLM plus Audio

- LLM backbone
- Audio model pretraining
- Rehearsal data

ALIGNMENT

Instruction Following

- Cross language understanding
- Tool use
- Audio + text tasks
- Understanding / ASR / TTS

SERVING

Low Latency

- VAD
- Tokenizer causes delay
- Efficient inference
- Latency hiding


Boson AI Datacenter


Boson AI Datacenter
• 1 MW of power
• Canada for stability
• 520 H100 (training) and 400 A100 (inference) GPUs
• 30PB+ storage
• Local DC management team
• 750 Tbit/s switching capacity

Audio Data

- **Quantity**
  - 100M+ hours of unannotated audio (crawled)
    (YouTube, podcasts, social, news, audiobooks, etc.)
  - Public datasets (1M hours or less), often unannotated
- **Quality**
  - Clean segments from voice actor collections
    (emotion, high quality recording, accent, identification)
  - Manual annotation for spot checking quality
  - 100+ languages

See also - Koluguri et al, 2025; Granary: Speech Recognition and Translation Dataset in 25 European Languages

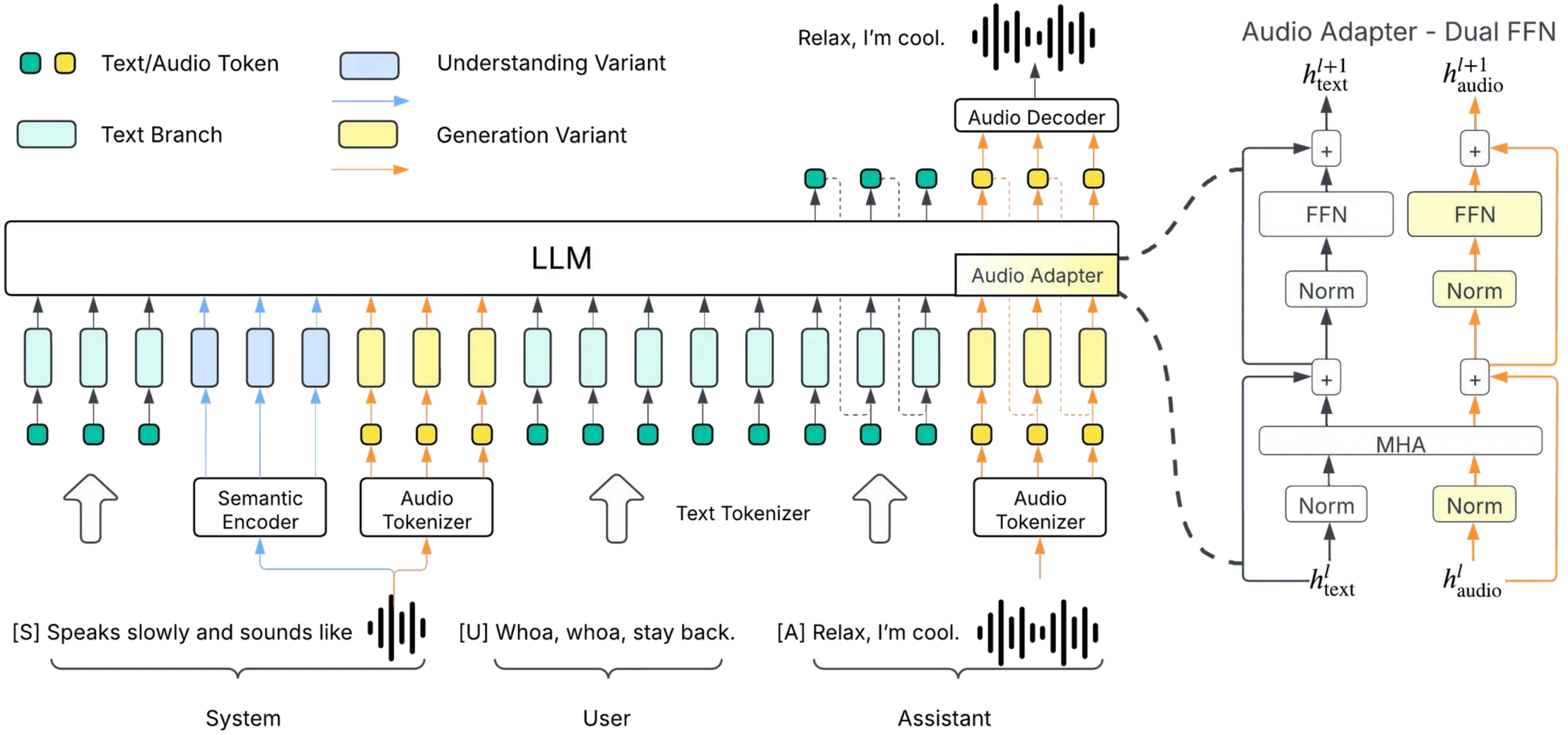
Audio Data

Large public datasets for tail languages are missing!

- **Preprocessing**
  - Careful segmentation (use ASR or generative model)
  - Open source ASR as a first step (also SSL)
  - LLM cleanup
- **Artificial Judge**
  - Use understanding model to judge
  - 3P judge for QC/calibration/benchmark
  - 10M hours of clean transcribed data for English (human lifetime is around 500,000 hours).

See also - Manku et al, 2024; EmergentTTS Eval

Higgs Audio (v2) Architecture

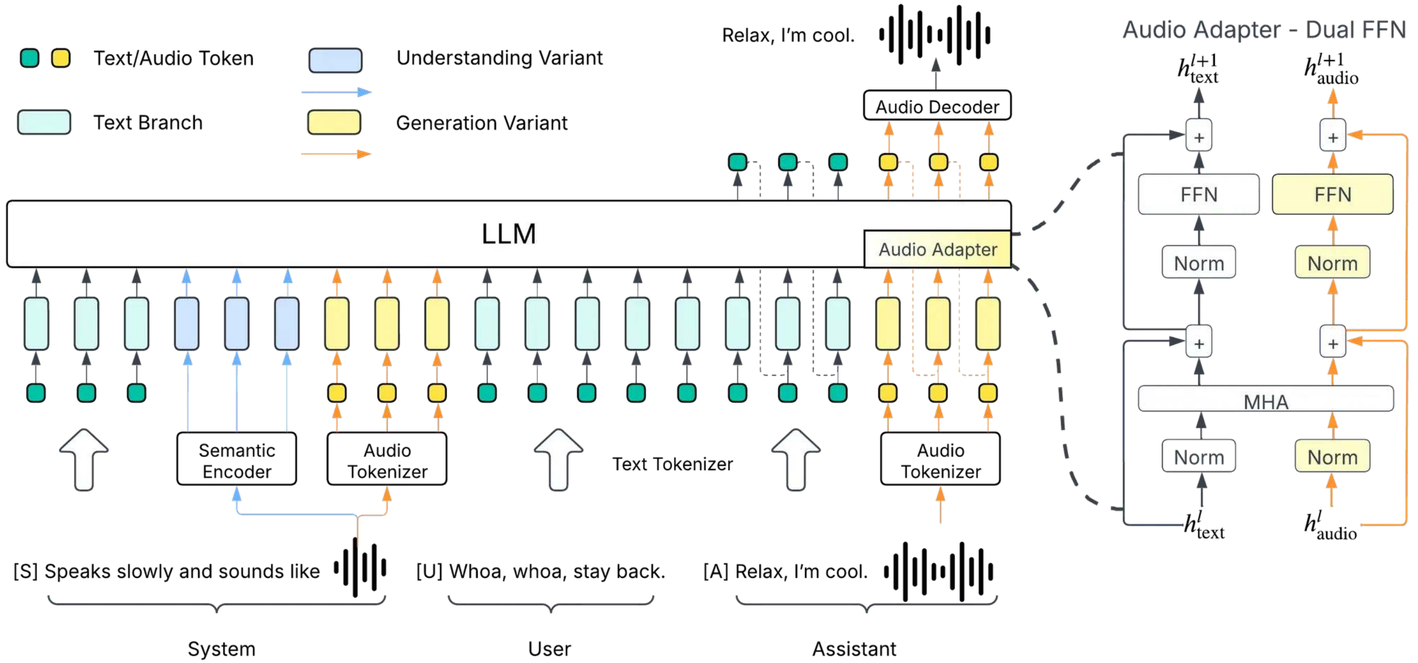


BosonAI

Very similar designs in Boson Higgs Audio 2, Xiaomi MiMo Audio, Microsoft VibeVoice

Training it

- **Input**
  - Text (instructions, transcription, system message, user)
  - Audio (reference voice, context, transcription task)
- **LLM backbone**
  - Tool use
  - Reasoning
- **Output**
  - Text (transcription, answer)
  - Audio (voice, music)



BosonAI

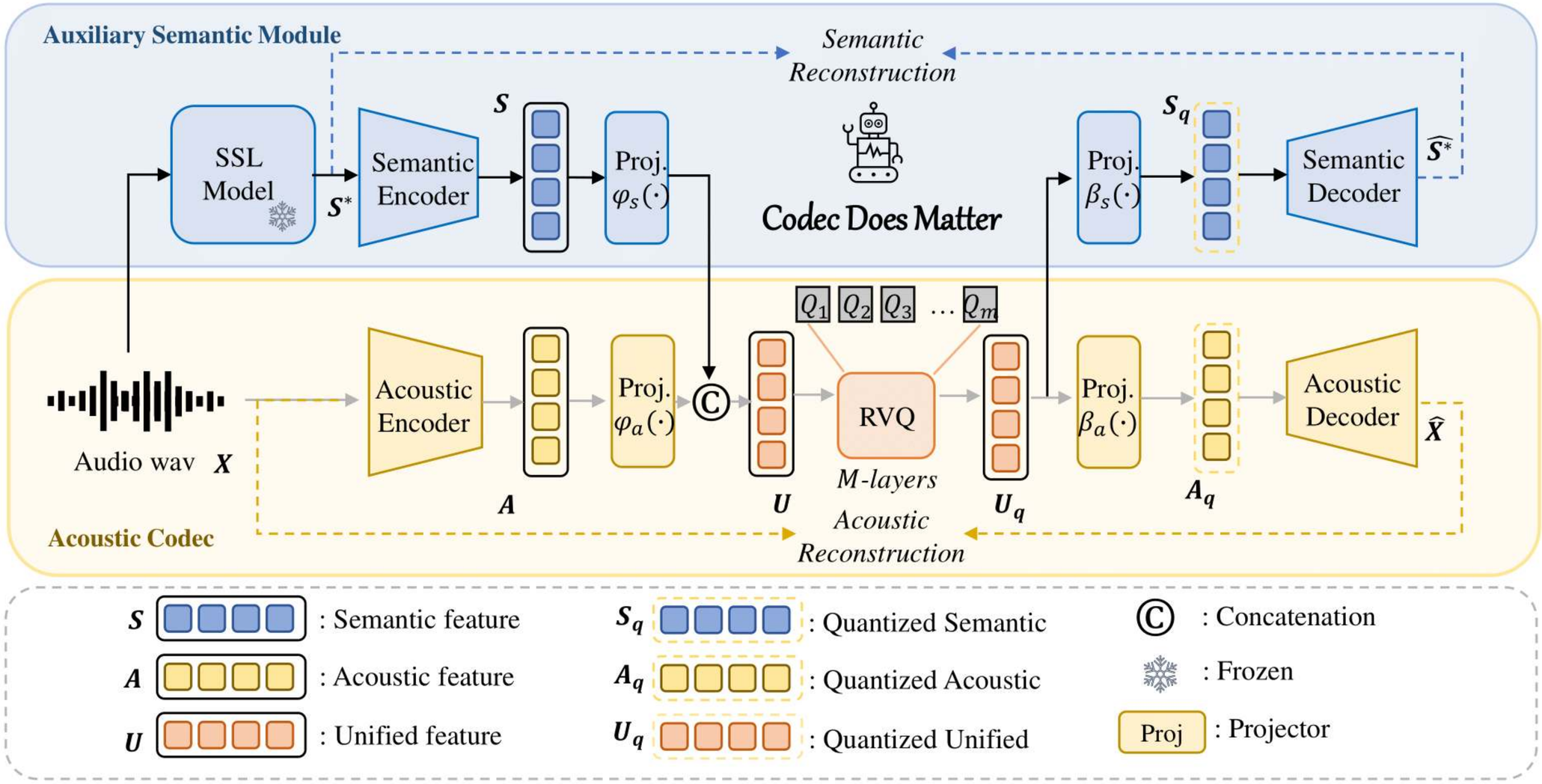
LLM Backbone

- **Plethora of great open weights LLMs, e.g.**
  - Llama (e.g. 3.3 70B, **3.2 1B/3B**, 11B)
  - Qwen 3.x (e.g. 235B, 30B, **4B**)
- **Features we care about**
  - Language (obviously) & language coverage (by now ~100)
  - Context window (audio has 5-10x larger tokens/s)
  - Tool use & reasoning
  - Speed vs. intelligence (smarter models are slower)

Tokenizer

- **Conflicting goals**
  - Understand any audio (even noisy telephony grade)
  - Generate beautiful audio
- **Solution**
  - Train one tokenizer for acoustic reconstruction
  - Train one tokenizer for understanding (e.g. via Whisper ASR model)
  - Use both …

Example - Xcodec 2.0

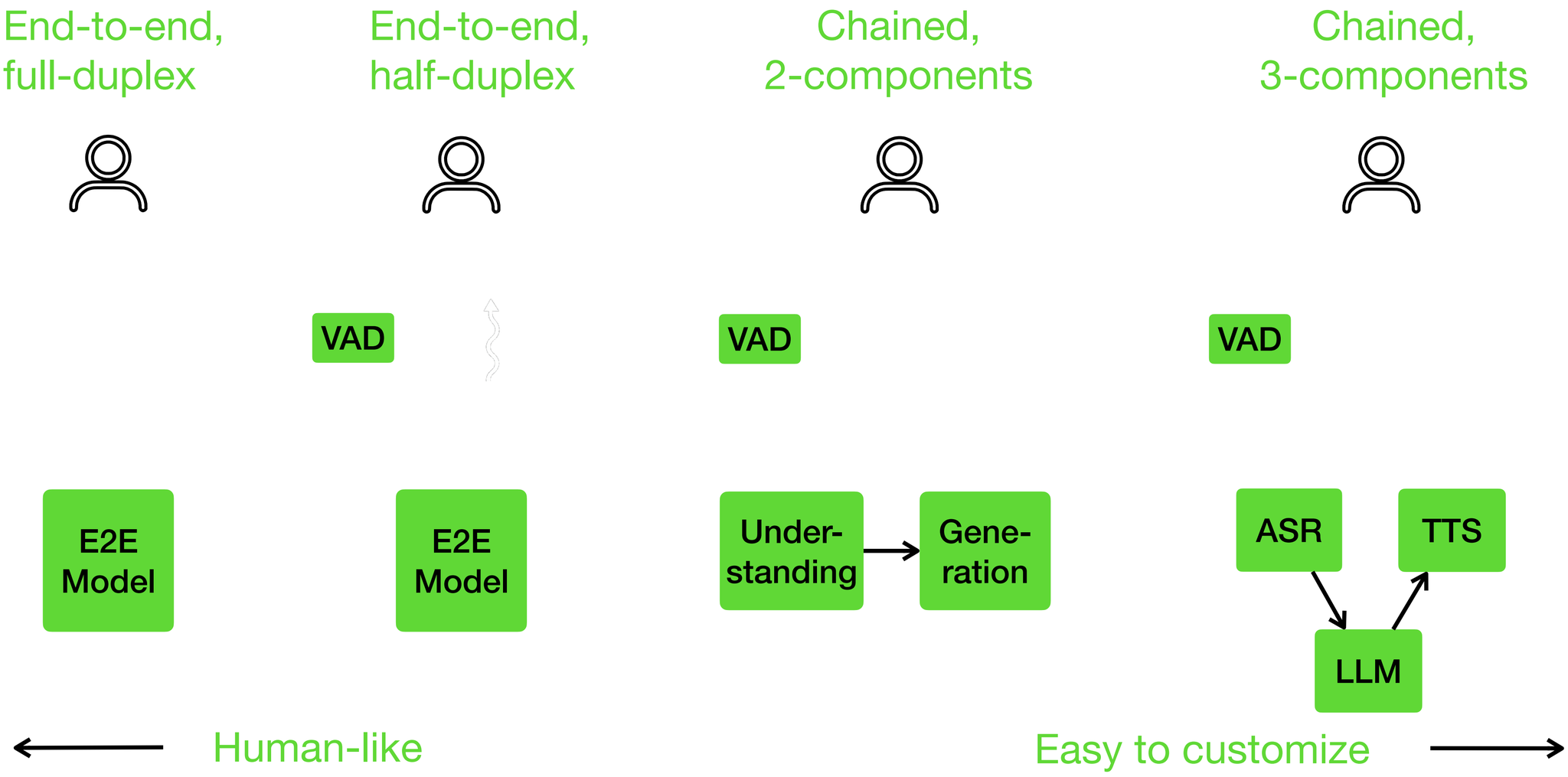


Xcodec 2, Ye et al, 2025

Have tokens, will train …

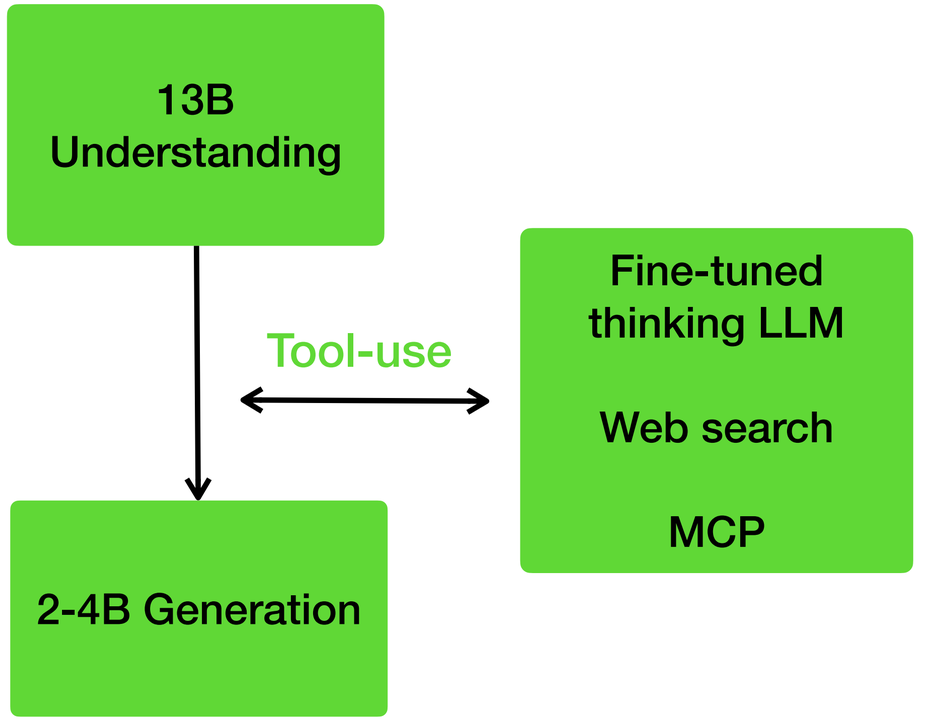
- **Standard model**
  - Tokens in, tokens out, maximize sequence likelihood
  - Well established optimizer libraries
- **Making the model work (fine tuning, alignment, …)**
  - **Rehearsal** data (don't lose the abilities of the LLM)
  - Design **dialogue** structure & prompts, e.g.
    - Instruction -> Audio -> Transcription (e.g. for ASR)
    - Instruction -> Audio -> Instruction -> Audio (e.g. for voice cloning)
    - Instruction -> Audio -> Text -> Audio -> Text -> Audio (two speaker cloning)
  - **Reward** signals, e.g.
    ASR accuracy, Audio model as a judge (calibrate against humans), Style, …
  - **GRPO** optimization

Voice Agent Architectures



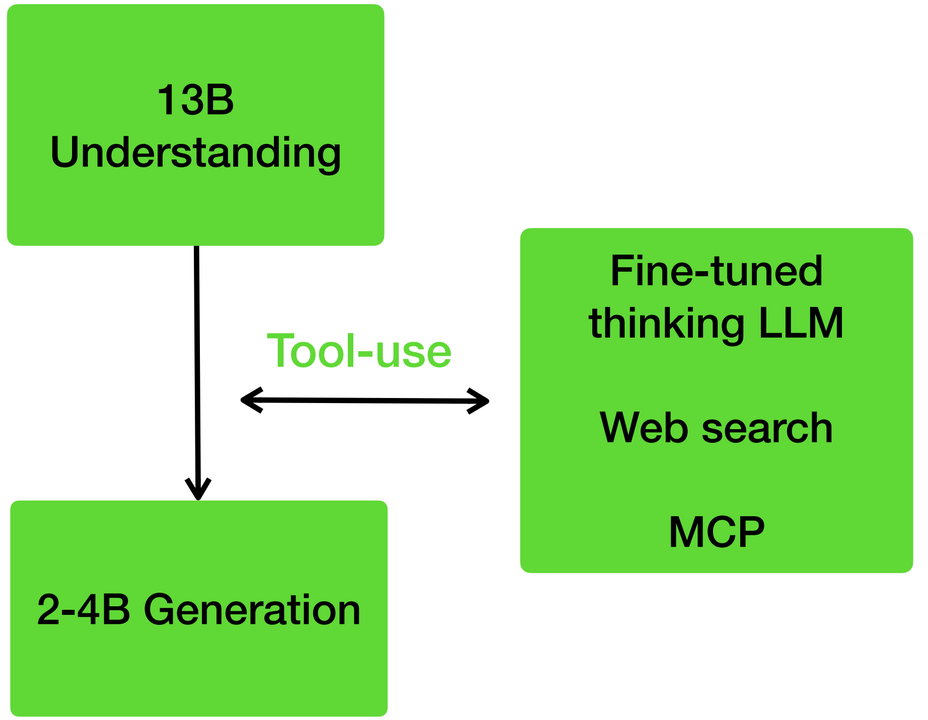
Chaining 2 Audio Models

- **Base LLM**
  - Continue pre-training / post-training on different data mixture
- **Understanding**
  - 10M+ hours of audio
  - Trillions of text tokens
  - Latency hiding
- **Generation**
  - 10M+ hours of audio
  - LLM finetuned with domain specific data



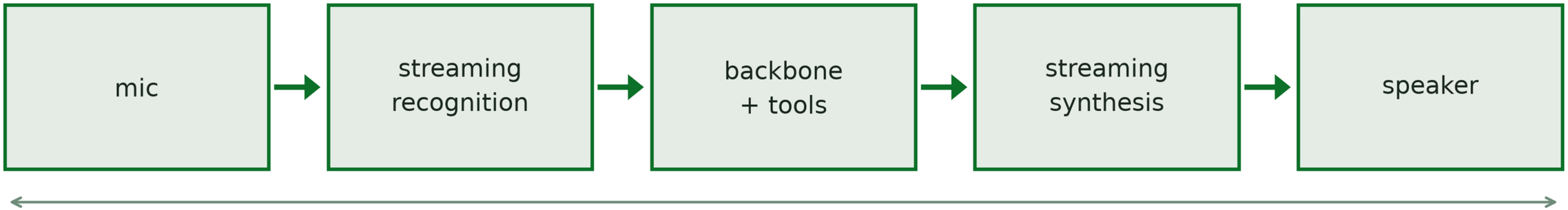
Benefits

- **Listen, talk, and think simultaneously**
- **Context Engineering**
  - surface the most relevant product knowledge & playbook guidance
- **Orchestrator**
  - handle strategy planning, intent analysis, and live task tracking



Latency Budget

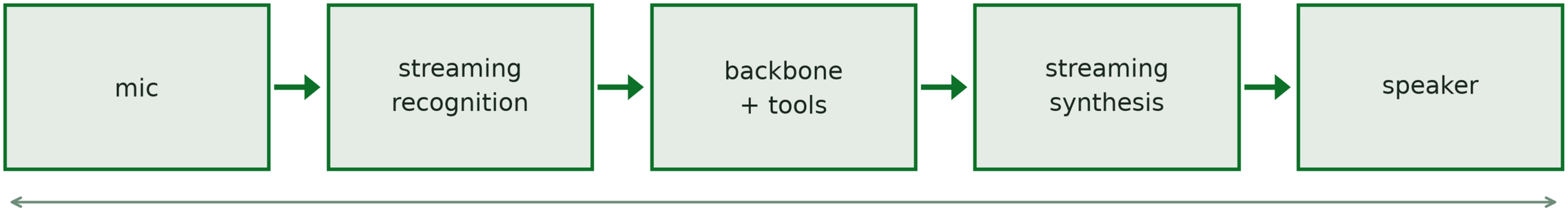
- Every stage streams; nothing waits for a whole turn
- The tool call runs while audio is still playing



VOICE ACTIVITY DETECTION	Turn end detected in under 300 ms; server, semantic, and manual modes
UNDERSTANDING	Streaming STT with chunk-prefill — transcription begins before the user finishes
GENERATION	Incremental; first tokens leave before the response is complete
SYNTHESIS	Streaming TTS; 150 ms to first token in the small model
INTERRUPTION	Output stops in 0.125 s; the response is truncated at the interruption point

Latency Budget

- Every stage streams; nothing waits for a whole turn
- The tool call runs while audio is still playing



first audio: 750 ms median, 1.16 s at p90

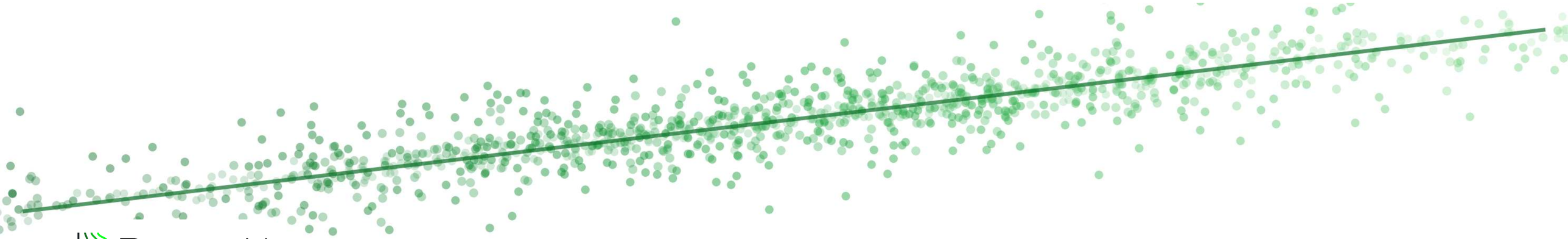
Pitfalls

- Barge-in that yields but never resumes
- Prosody that contradicts the words
- Tool calls that stall the audio stream
- Language switches that reset the voice

Show and tell
Building it
Evaluation
Picking good benchmarks

3

Human performance without the cost



BosonAI

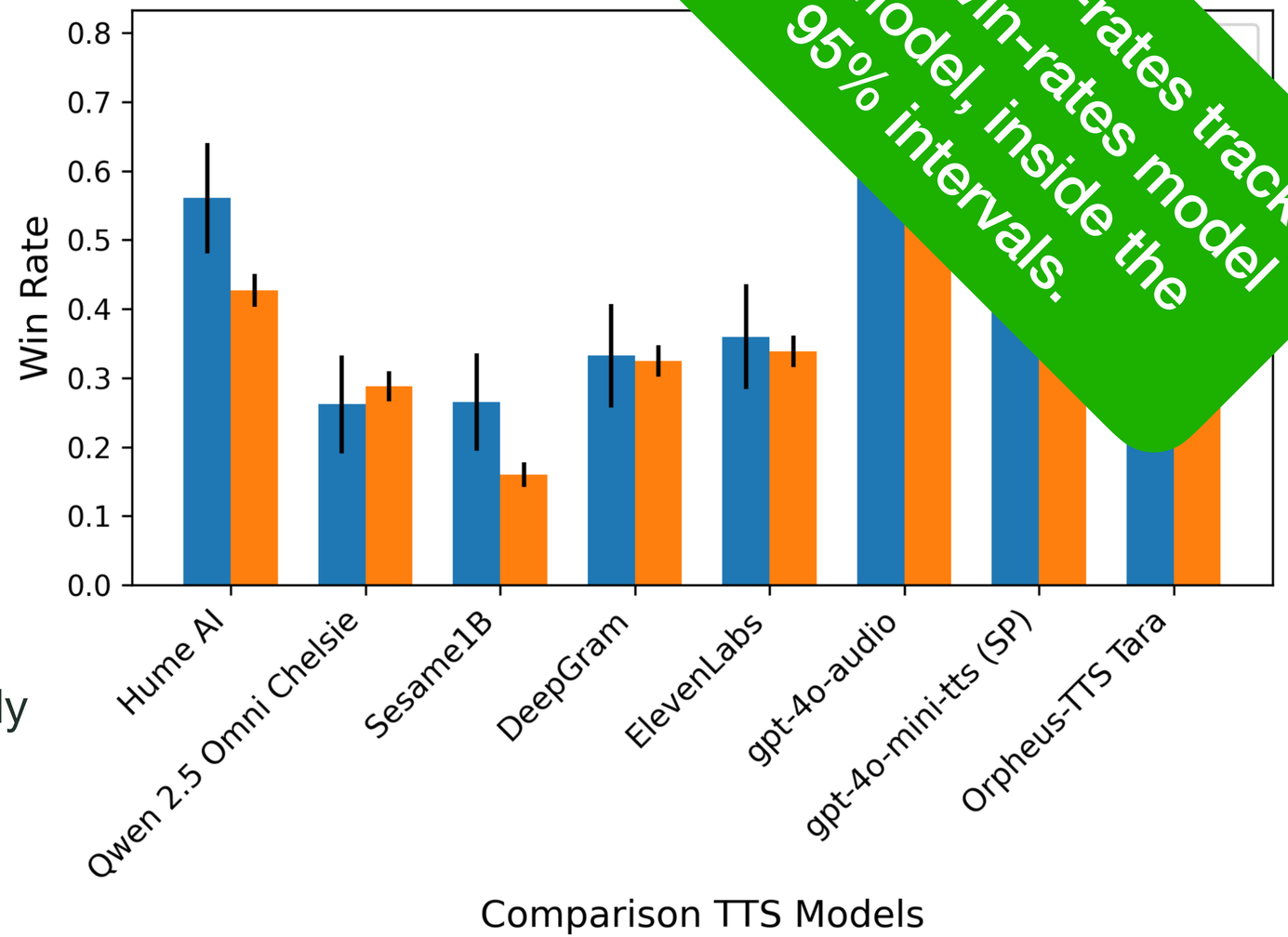
Humans are expensive.
Humans are random.

Can we trust LLMs
to help us improve LLMs?

Speech quality

EmergentTTS-Eval

- **Six families of hard speech**
  - emotions
  - paralinguistics
  - foreign words
  - complex syntax
  - URLs and formulas
  - questions
- **Human generated seeds**
  - Grown by LLMs
  - Targets structure, phonetics, prosody
  - 1,645 test cases
- **Automatic evaluation**



Judge win-rates track human win-rates model for model, inside the 95% intervals.

BosonAI

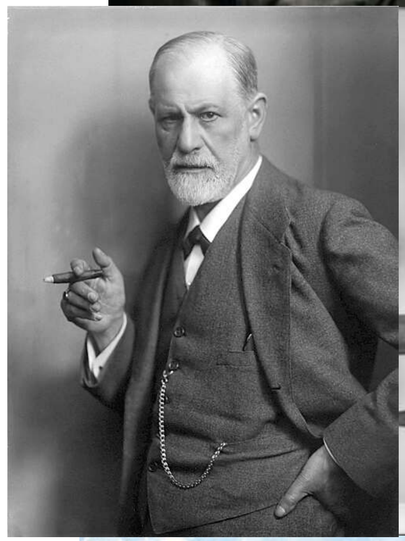
EmergentTTS-Eval, arXiv:2505.23009 (NeurIPS 2025). Tested on 11Labs, Deepgram, 4o-mini-TTS.

What about Building a Conversational Agent

- **Intuitive dialog design**
  - Needs to work for ‘regular users’
  - Scene, tools and context
    e.g. Genie Worksheets (Joshi et al, 2024)
- **RL and Planning**
  - Turn by turn is entirely reactive
  - Plan larger strategies (action space is infinite)
    e.g. Prompt based MCTS (Yu, Chen and Yu, 2023)
- **Use evidence rather than psychology for planning**

Learning human behavior / emotions

- **Interaction in literature / movies are unrealistic**
  - Magical negro trope
  - Manic Pixie Dream Girl
  - RomCom hero = stalker
- **Psychology & Social Science**
  - Might be correct for extreme cases
  - Small sample size
  - Doesn’t cover short-term interactions
- **With large audio models we can finally learn the real emotional dynamics of humans.**



BosonAI

Humans are expensive.
Humans are random.

Can we trust LLMs
to help us improve LLMs?

Proactivity

Proactivity

- **User**
  Plain-text packet is final. I'll send tonight and load the hatchback after work.
- **Model A**
  Sounds like a clean plan. Drive safe, and go get it.
- **Model B**
  Before you go, a fast final check: projector + power cable; media player + charger; file tested and looping; tripod + mounting plate; extension cord; adapter; gaffer tape; measuring tape; small level; wall label on phone; backup copy on USB. If you can, pack the car tonight so tomorrow is just: grab computer, leave, install. Load in reverse order of install — the projector stand and extension cord go in last, so they're the first things out.

Harfi et al., ProactBench, arXiv:2605.09228.

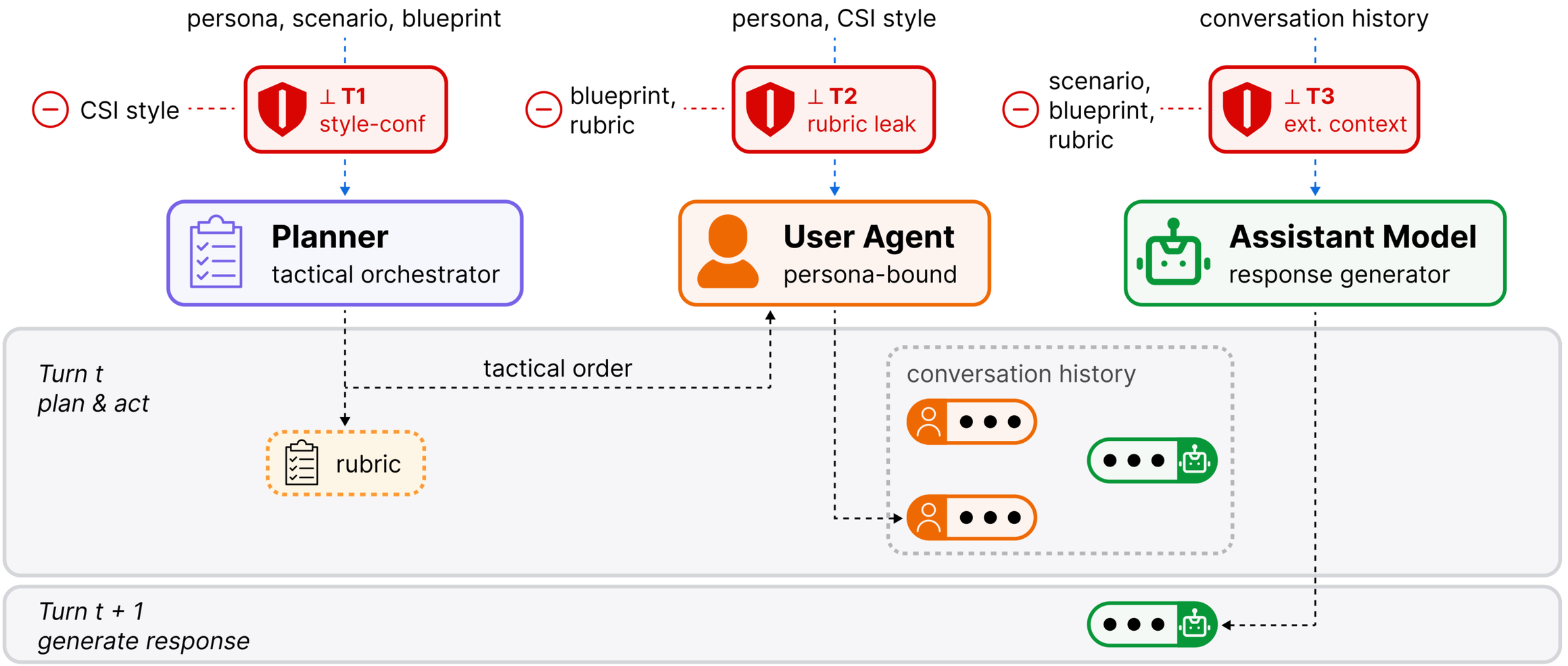
Our goal - infer what the user wants and help *proactively*

- **Design dialogue**
  Plant anchors for where the user might require support
- **See whether model picks up cues regarding help**
  Excludes instruction-following, clarification-seeking, generic helpfulness, sycophantic closings, etc.

EMERGENT	Infer an unstated need from one anchor	Early in the dialogue
CRITICAL	Synthesise several anchors into one need	Mid-dialogue
RECOVERY	Offer grounded forward-looking value	After the user signals done

Harfi et al., ProactBench, arXiv:2605.09228.

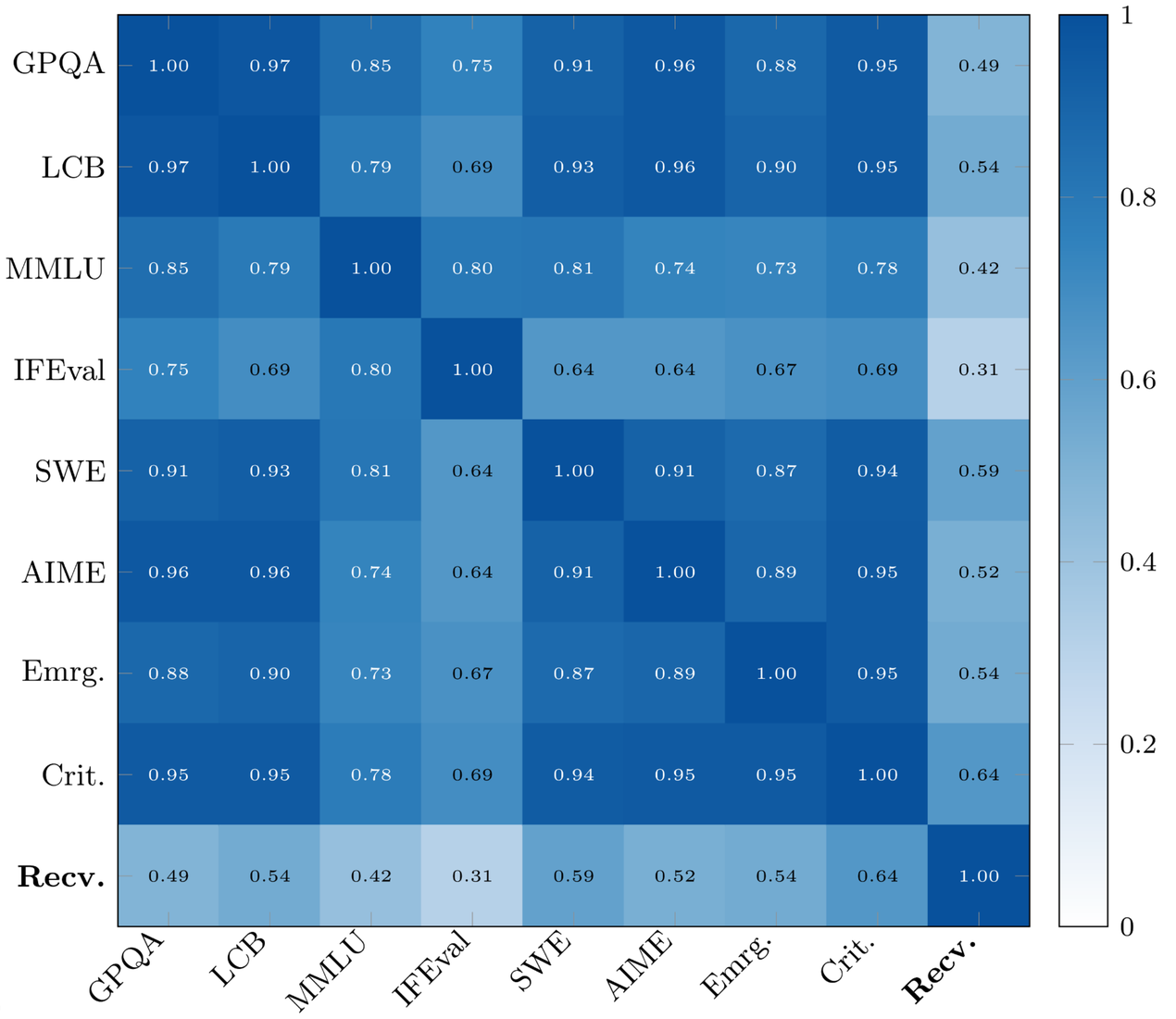
Dialog Design



Harfi et al., ProactBench, arXiv:2605.09228.

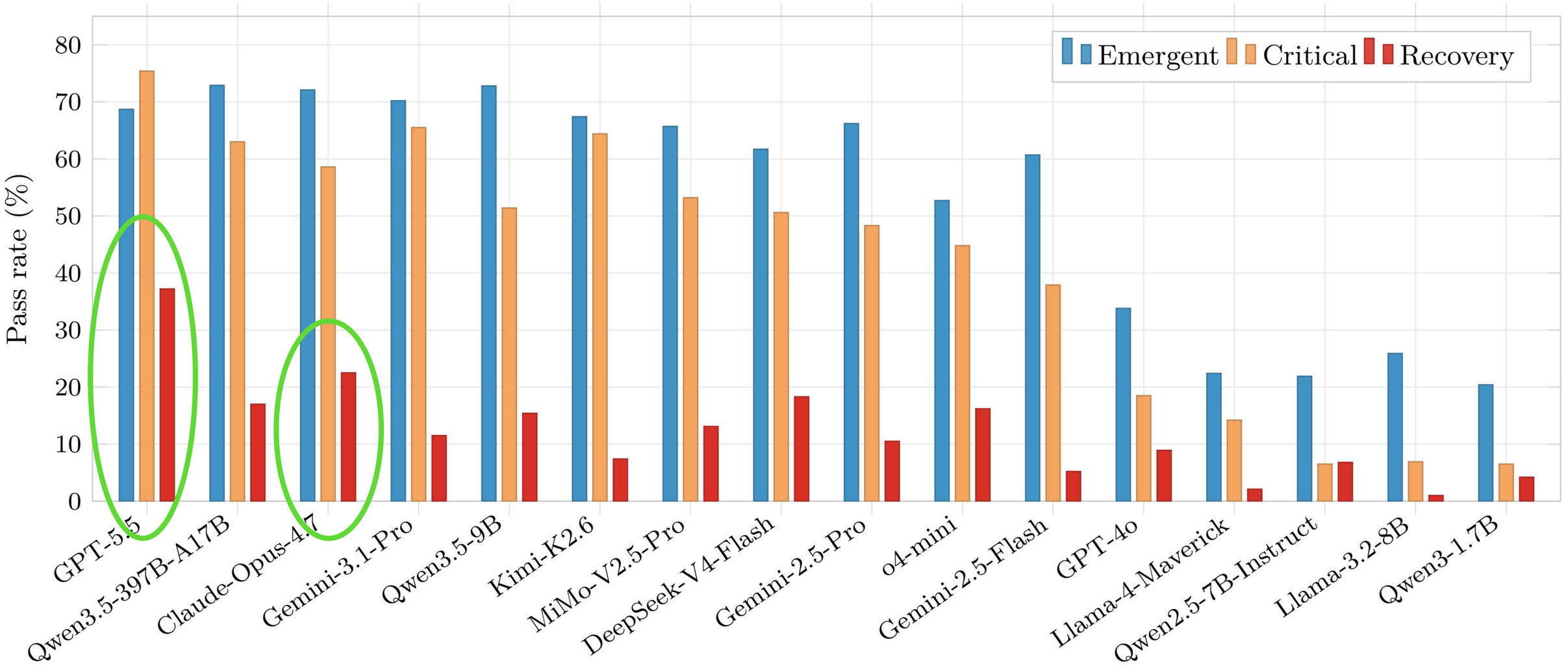
Recovery Is Hard and Weakly Predicted

- 198 dialogues, 624 trigger points
- 24 communication styles (CSI traits)
- 16 frontier and open-weight models
- Six standard benchmarks: r = 0.64–0.97
- Recovery: mean r = 0.51 [0.29, 0.71]
- **Raters prefer Recovery-rubric replies in 80% of 144 non-tie pairs**



Harfi et al., ProactBench, arXiv:2605.09228.

Recovery is hard



BosonAI

Harfi et al., ProactBench, arXiv:2605.09228.

Humans are expensive.
Humans are random.

Can we trust LLMs
to help us improve LLMs?

Interruption handling

Interruption handling

- Prior work scores when the agent yields but doesn't measure what happens later.
- Workflows modeled as state machines (10 domains, fictional data, different stages, never clean)

Type	What recovery requires
NORMAL	Address the cut-in, then resume
IMPATIENT	Skip ahead without re-explaining
CORRECTION	Accept without pushback, integrate, continue
TOPIC SWITCH	Answer the new topic, steer back without blending
FILLER	Continue exactly; no restart, no acknowledgement
PUSHBACK	De-escalate, offer an alternative

Example A: Filler interruption (backchannel)

Assistant: "...and there is elevator access from the garage. Back—"

User (interrupts)**:** "mm-hm"

✗ **Bad response:** "Before we continue, I just want to ..."

Starts over after a simple backchannel, and jumps to a different workflow stage.

✓ **Good response:** "to your registration, we still need to schedule your first appointment. Is now a good time?"

Continues exactly from where it was cut off.

Example B: Correction interruption

Assistant: "Got it, I'll use john.smith@example-company.com. Reading that back—"

User (interrupts)**:** "Oh wait, sorry, no hyphen. It should be examplecompany as one word."

✗ **Bad response:**
"Sure, it's john.smith@examplecompany.com. I'll use that for the email."

Accepts correction but drops the workflow (was about to send summary).

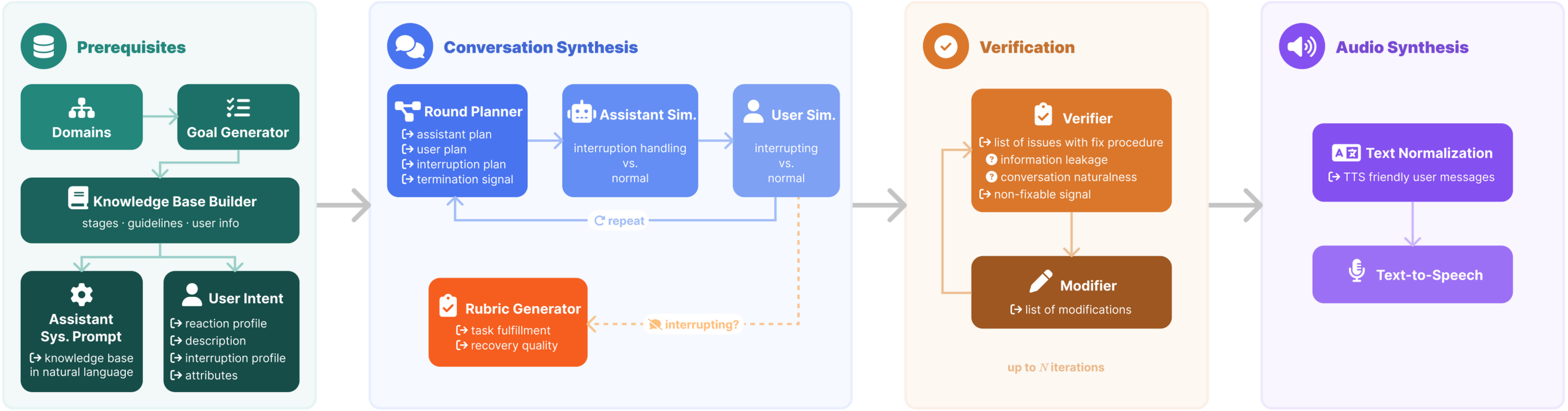
✓ **Good response:**
"Got it, I'll use john.smith@examplecompany.com. I'm sending the brief summary now with Jane Doe's direct contact and scheduling link."

Integrates correction and resumes the workflow.

Salimi et al, IHBench, arXiv:2606.19595

User Simulator

- **Prerequisites → rounds → rubrics → verify**
  - Round planner, assistant sim, user sim, so undelivered content cannot leak
  - Verifier discards what it can't repair
- **Two metrics**
  - TF = task-fulfillment win rate
  - RQ = recovery-quality pass rate



Task Fulfillment and Recovery Quality Separate Models Along Different Axes

Model	TF win	RQ pass
GPT Realtime 2	**0.728**	0.624
GPT Realtime 1.5	**0.654**	**0.655**
Gemini 3 Flash	**0.632**	0.605
Gemini 2.5 Flash	0.586	**0.704**
Gemma 4 12B	0.505	0.540
Qwen3-Omni-30B	0.304	**0.676**
Qwen2-Audio-7B	0.044	0.395

DEPTH DECAY

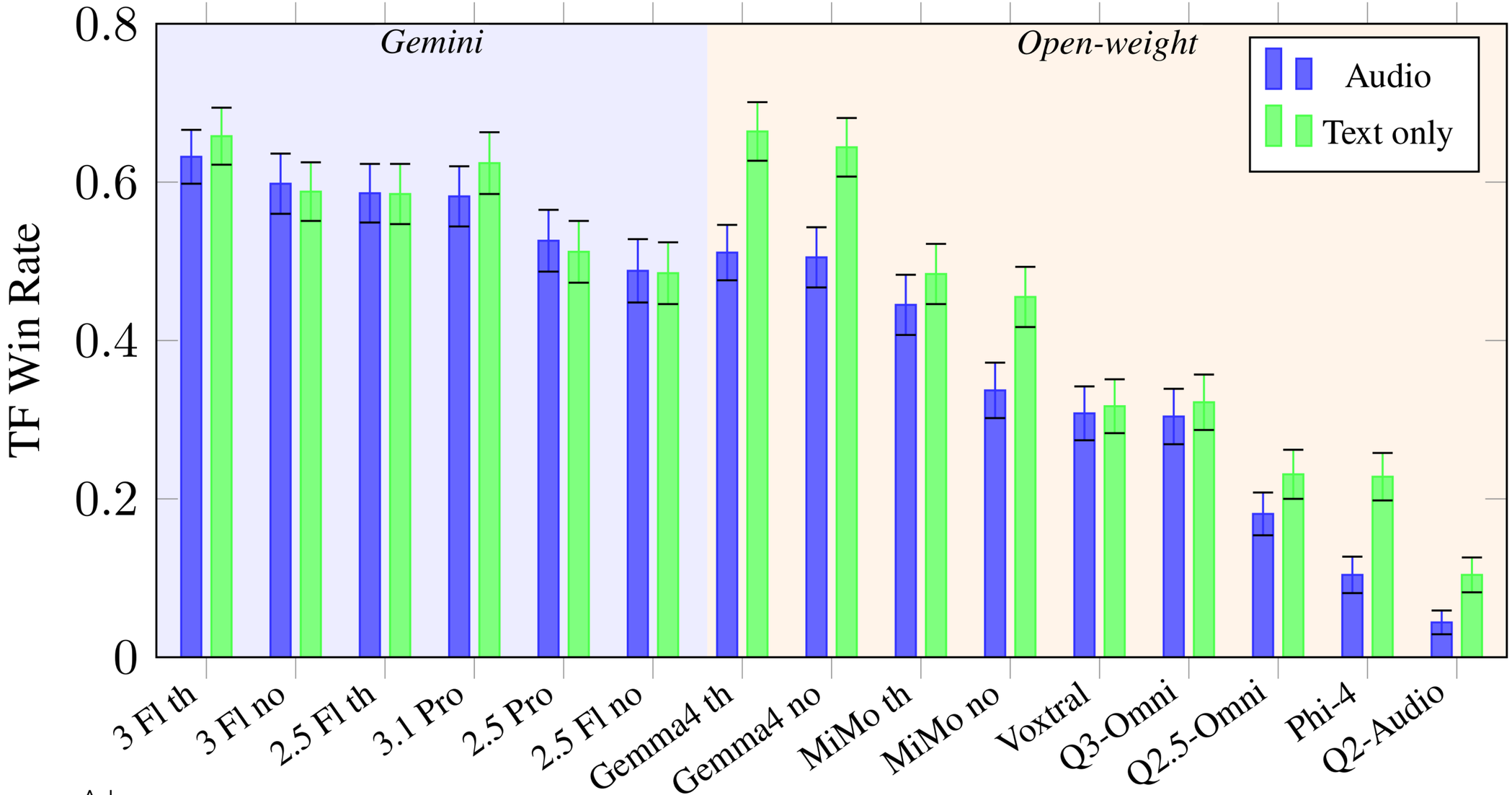
25 of 27 configs decline with conversation depth. Open-weight slopes average –0.053 against –0.015 frontier.

FILLER CONTINUATION

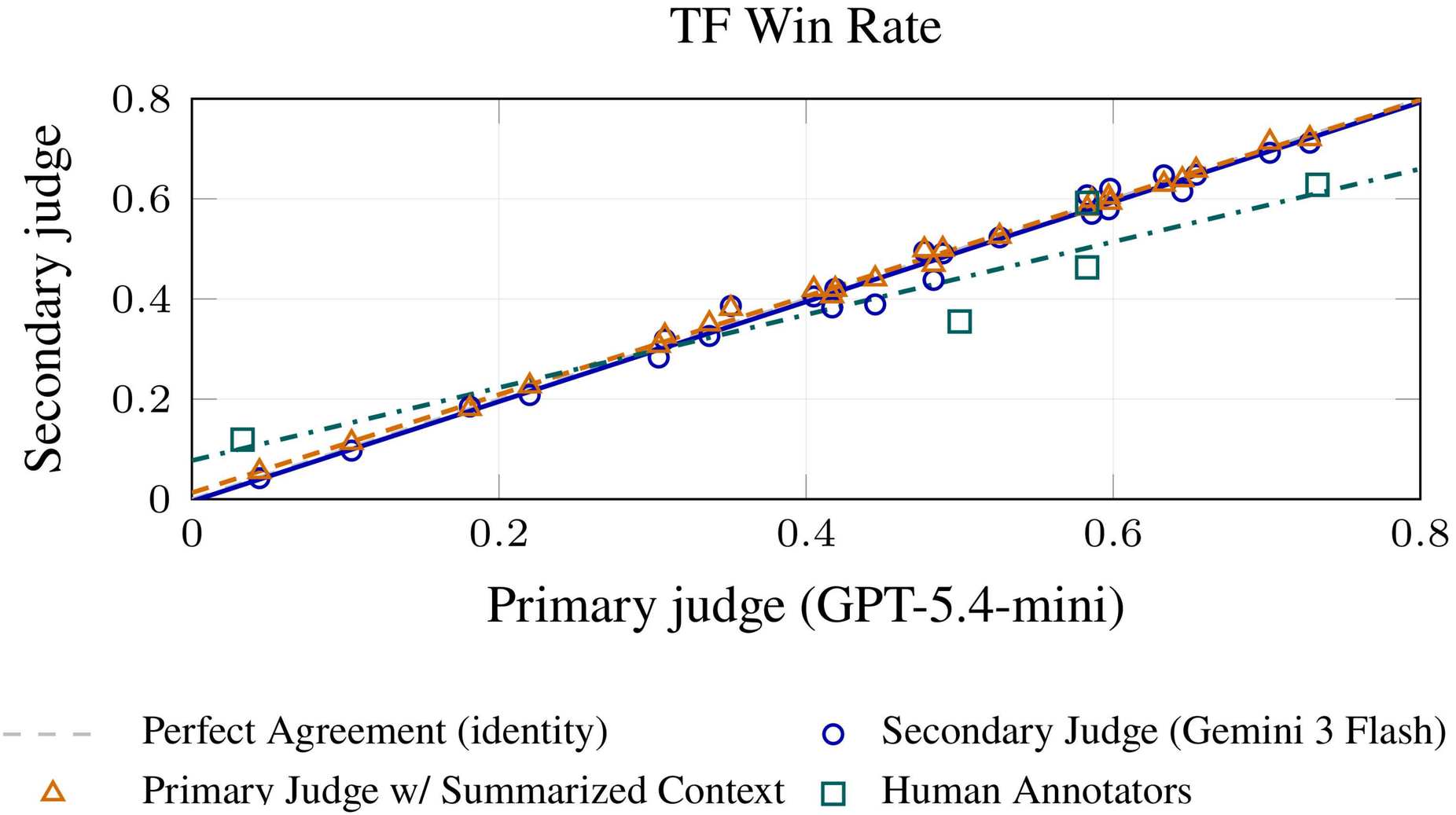
After **‘mm-hm’**: GPT continues correctly 7–31% of the time, Gemini 2.5 62–68%.

Salimi et al, IHBench, arXiv:2606.19595

Audio Matches Text for Gemini, Not for Open Weights



Sanity Check - Second LLM Judge Reproduces the Ranking



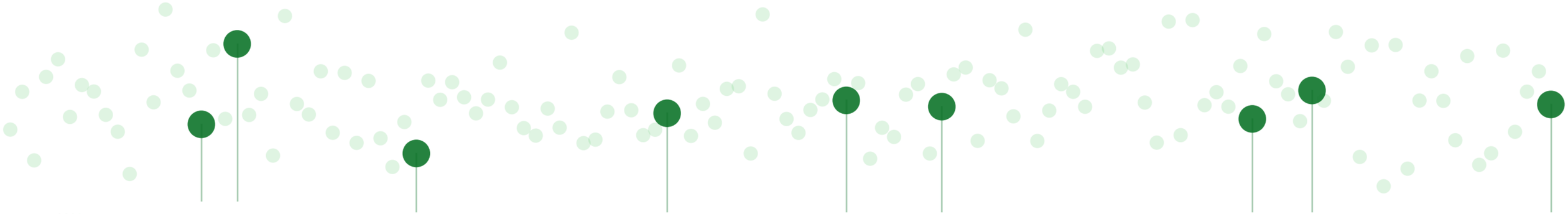
Salimi et al, IHBench, arXiv:2606.19595

Show and tell
Building it
Evaluation

4

Picking good benchmarks

Do we really need them all?



BosonAI

An ever increasing number of tasks (benchmarks)

Collection	Models	Tasks
MMLU	5,452	57
Open LLM v2	4,507	6
MTEB	263	56
AlpacaEval 2	223	3
LiveBench	195	3
BigCodeBench	155	14
WildBench	63	11
Arena-Hard	60	1
HELM Lite	30	11
MT-Bench	5	8
Total tasks		170+

TWO COMPETING PRESSURES

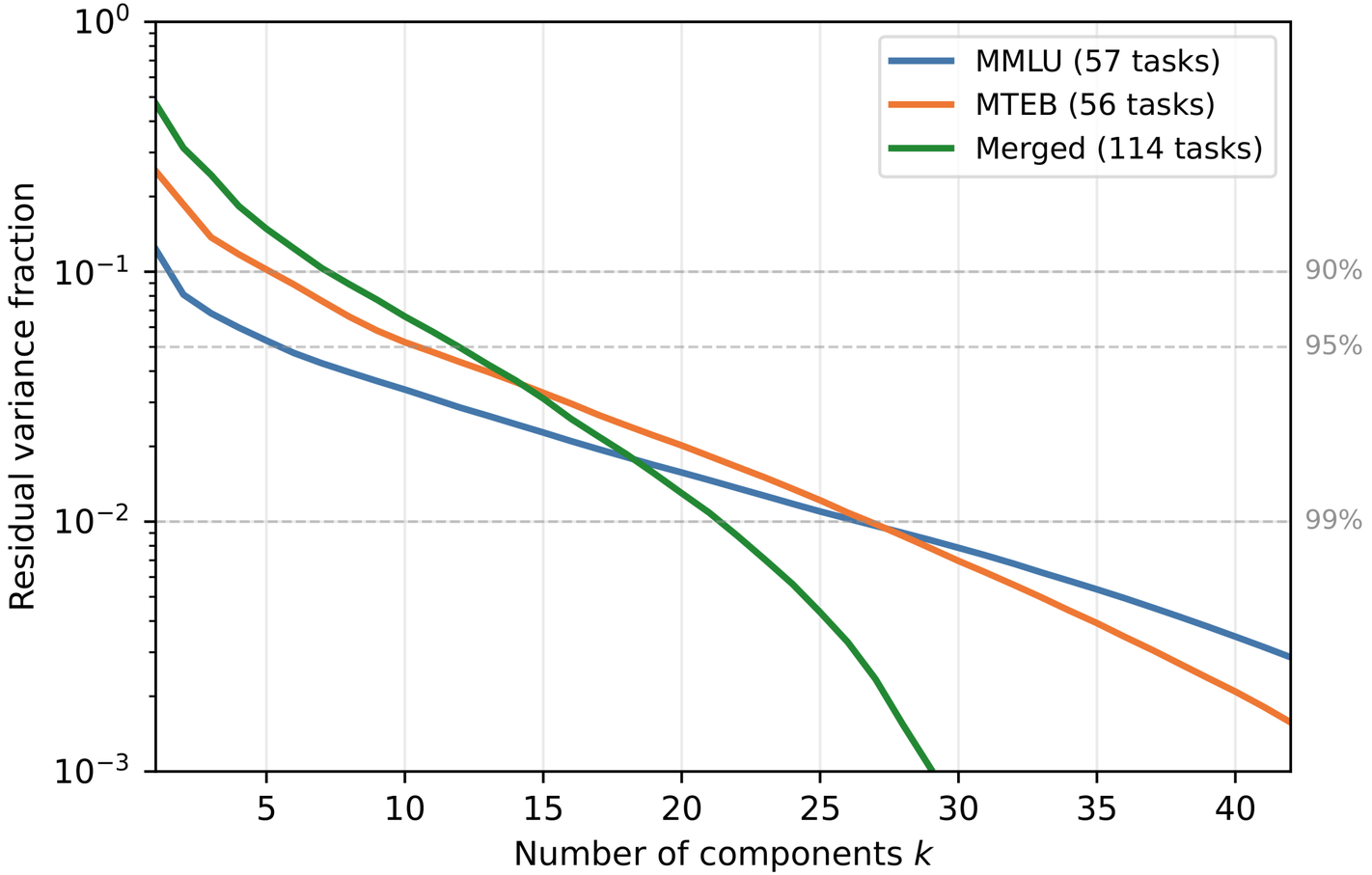
Completeness wants math, code, knowledge, reasoning, safety.
Cost wants none of it (GPU / humans)

KEY IDEA

Many benchmarks measure the same capability.
A small subset should suffice but which ones?

Ten public leaderboards — Submodular Benchmark Selection, arXiv:2605.02209.

Benchmark Scores Live in a Low-Dimensional Subspace



90% OF THE VARIANCE

MMLU: 2 of 57 components.
MTEB: 6 of 56.
Merged: 8 of 114.

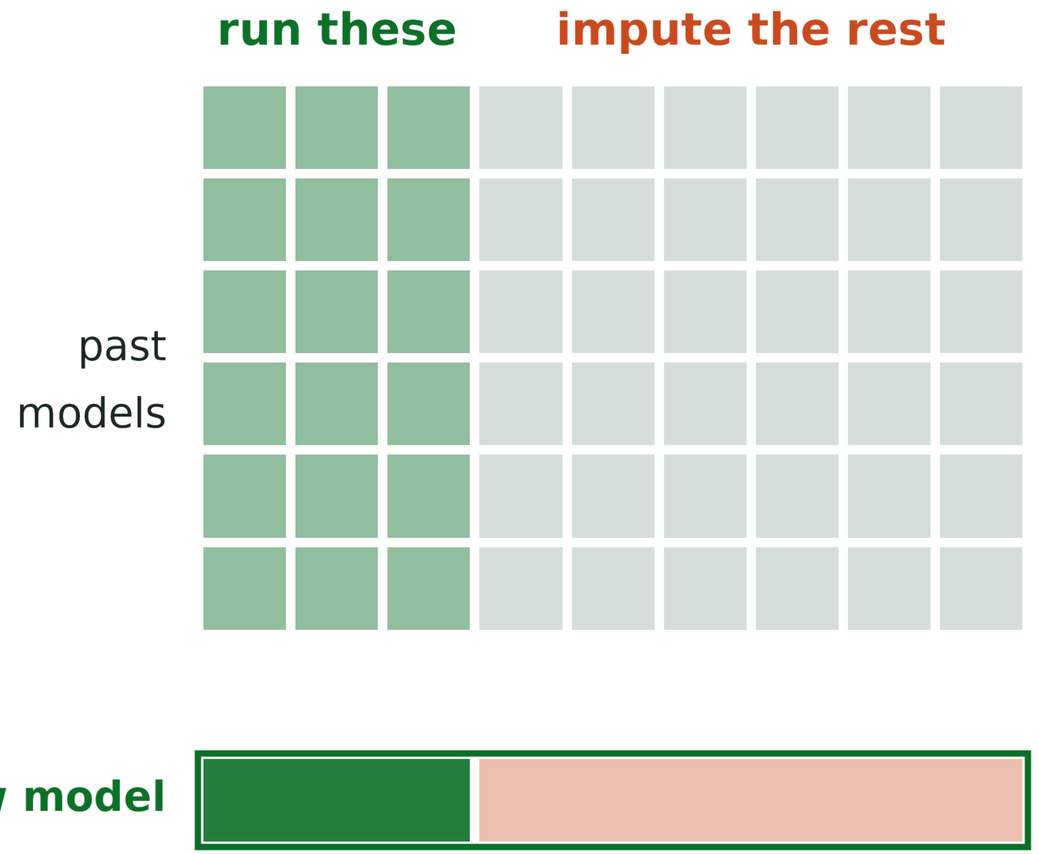
WHAT THAT PROMISES

A handful of well-chosen benchmarks carries nearly all the signal. Five will reach $R^2 \approx 0.91$ — with a guarantee attached.

Residual variance 1 – ρ(k) of the score correlation matrix, log scale — arXiv:2605.02209.

Choose a subset and extrapolate

- Scores matrix
  - M models × N benchmarks
  - Leaderboards give μ and Σ for free
  - Choose A before the model exists
- New model
  - run A, regress the rest
  - Fit on past models' scores



DESIGN QUESTIONS

Which subset A, at each budget k
How good can the imputation get?

Smola, Submodular Benchmark Selection, arXiv:2605.02209.

Let’s pretend everything is Gaussian

- Rows of B drawn i.i.d. from N(μ, Σ) (obviously wrong, but useful)
- Conditional mean fills the gaps
- Covariance: a Schur complement

$$\widehat{B}_{i,\bar{\mathcal{A}}} = \mu_{\bar{\mathcal{A}}} + \Sigma_{\bar{\mathcal{A}}\mathcal{A}} \Sigma_{\mathcal{A}\mathcal{A}}^{-1} \left(B_{i,\mathcal{A}} - \mu_{\mathcal{A}} \right)$$

$$\Sigma_{\bar{\mathcal{A}}|\mathcal{A}} = \Sigma_{\bar{\mathcal{A}}\bar{\mathcal{A}}} - \Sigma_{\bar{\mathcal{A}}\mathcal{A}} \Sigma_{\mathcal{A}\mathcal{A}}^{-1} \Sigma_{\mathcal{A}\bar{\mathcal{A}}}$$

- **Collateral benefit**
  Fill in missing values via EM

CONDITIONAL COVARIANCE

Only depends on which benchmarks are in A, never on the observed scores. A can be chosen before the model exists (convenient but weird).

LINEAR MODEL

The conditional mean is the best linear predictor whatever the true distribution.

IN PRACTICE

Normality test fails hard (but it works)

Two Objectives: Entropy and Mutual Information

HOW MUCH SIGNAL IS IN THE DATA?

Entropy — the volume of the selected set. Rewards benchmarks diverse from one another.

$$f_1(\mathcal{A}) = H(X_{\mathcal{A}}) = \tfrac{1}{2}\log\det\left(2\pi e\,\Sigma_{\mathcal{A}\mathcal{A}}\right)$$

WHAT DOES A SAY ABOUT THE REST?

Mutual information — coupling between A and its complement. Rewards predictive hubs.

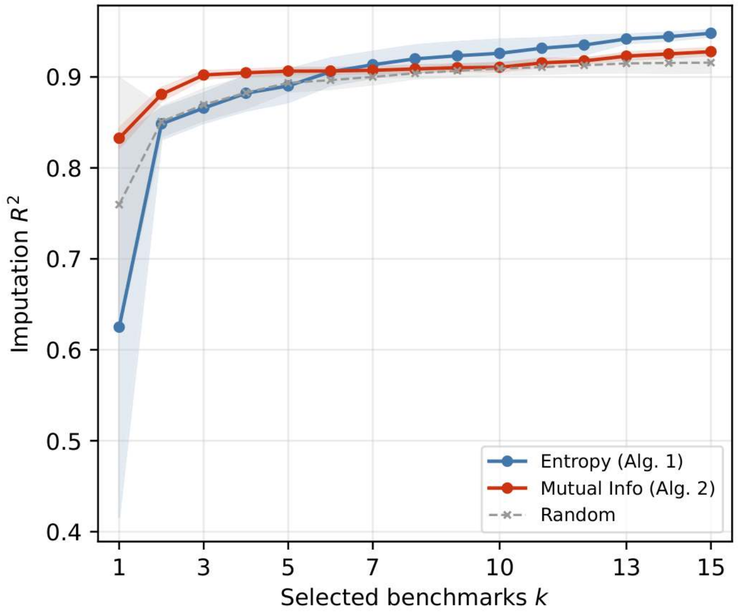
$$f_2(\mathcal{A}) = I(X_{\mathcal{A}}; X_{\bar{\mathcal{A}}}) = \tfrac{1}{2}\left[\log\det\Sigma_{\mathcal{A}\mathcal{A}} + \log\det\Sigma_{\bar{\mathcal{A}}\bar{\mathcal{A}}} - \log\det\Sigma\right]$$

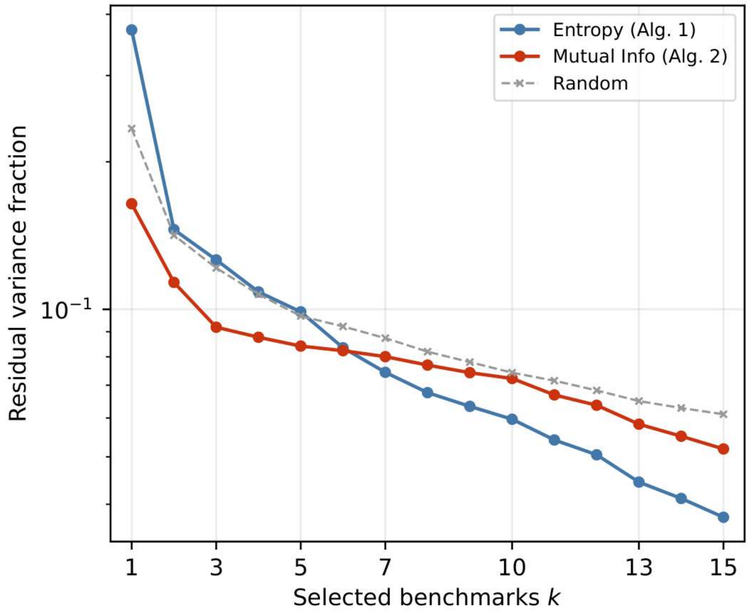
$$\Delta(v \mid \mathcal{A}) = \tfrac{1}{2}\underbrace{\log\sigma^2_{v|\mathcal{A}}}_{\text{grows with diversity}} - \tfrac{1}{2}\underbrace{\log\sigma^2_{v|\bar{\mathcal{A}}\setminus v}}_{\text{shrinks with coupling}}$$

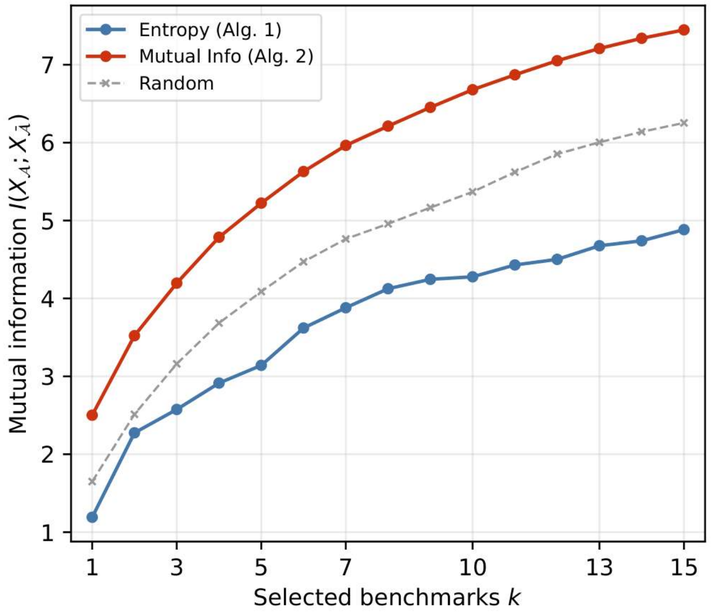
SUBMODULAR GREEDY SELECTION (Guestrin & Krause, 2005)

Pick one benchmark at a time. Guaranteed within (1 - 1/e) of optimality (Nemhauser et al. 1978).

MI Wins Exactly Where the Budget Is Tight







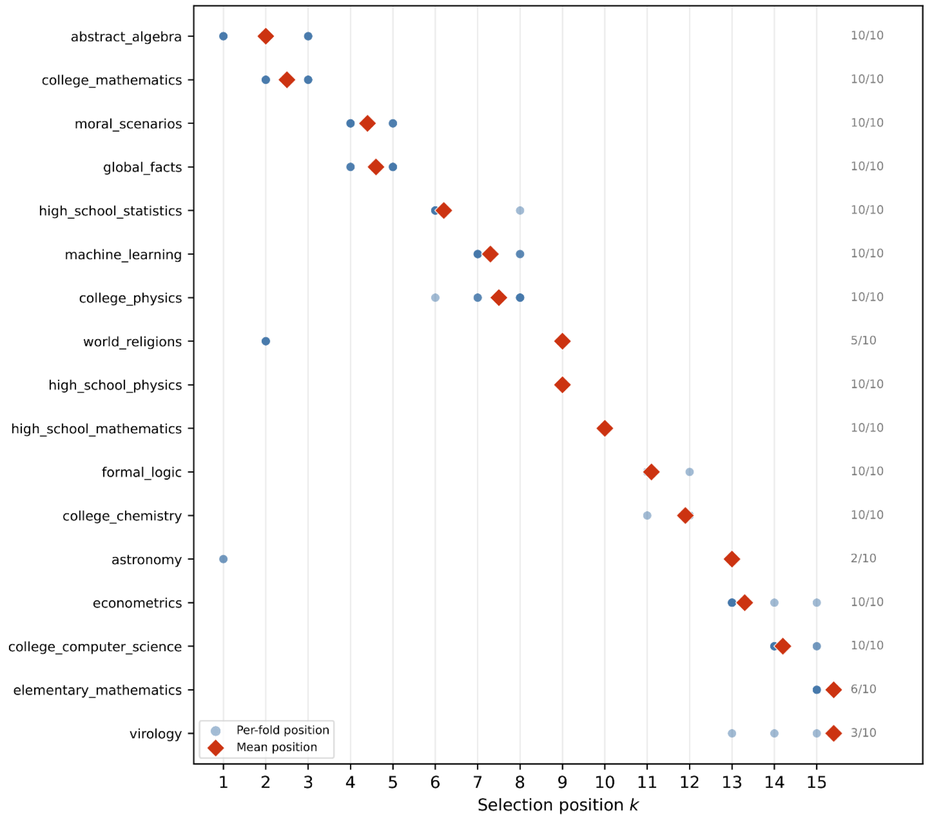
SURROGATE GAP

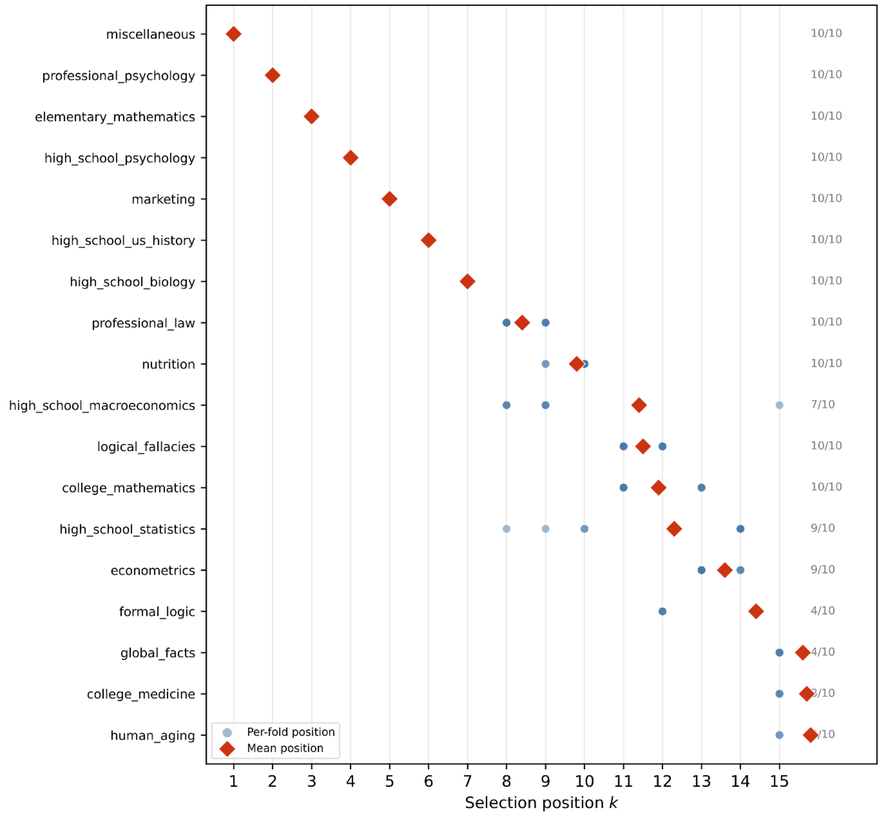
k = 1: R^2 0.83 (MI) vs 0.65 (entropy); crossover near k ≈ 7. Entropy leaves lower residual variance yet imputes worse — diversity inside A is not coupling with the rest.

BosonAI

Entropy Collects Outliers, MI Collects Hubs

- Entropy first picks: abstract algebra, moral scenarios
- Mutual Information first picks: miscellaneous, professional psychology

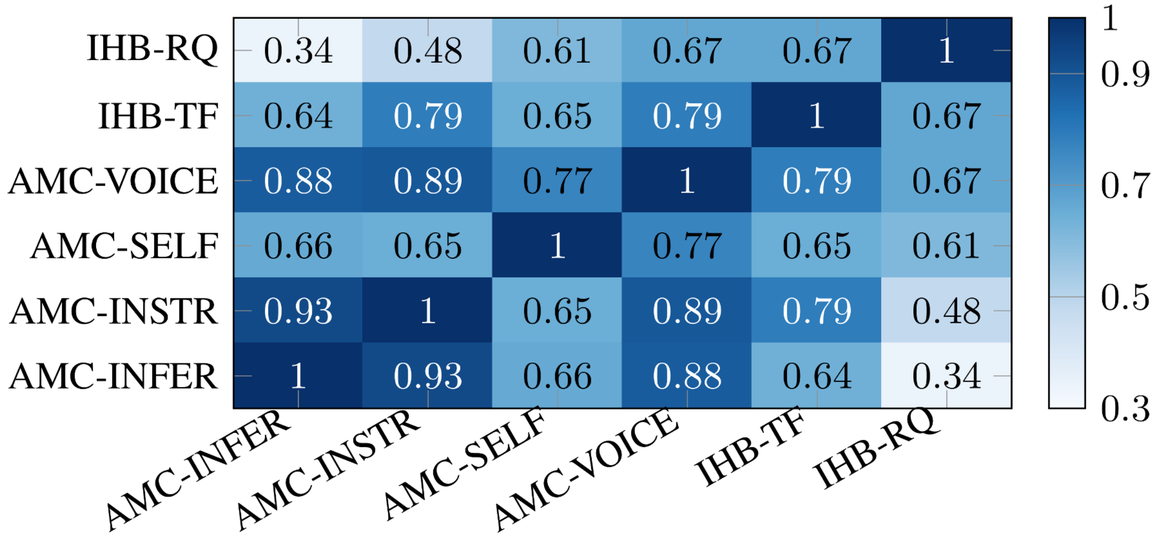




BosonAI

Using this to design our own benchmarks

- ProactBench
  - Recovery plus 9 other benchmarks
  - Greedy entropy ranks it #2 of 9
- IHBench
  - Recovery Quality plus 6 other benchmarks
  - Greedy entropy ranks it #2 of 6



SANITY CHECK

Selection knows nothing about ‘human-facing’ aspect. Covariance helps determine new axes

- **Interruption Benchmark**
  arXiv 2606.19595
- **Proactivity Benchmark**
  arXiv 2605.09228
- **Benchmark Selection**
  arXiv 2605.02209
- **TTS Benchmark**
  arXiv 2505.23009
- **Roleplay Benchmark**
  arXiv 2502.00595

Summary

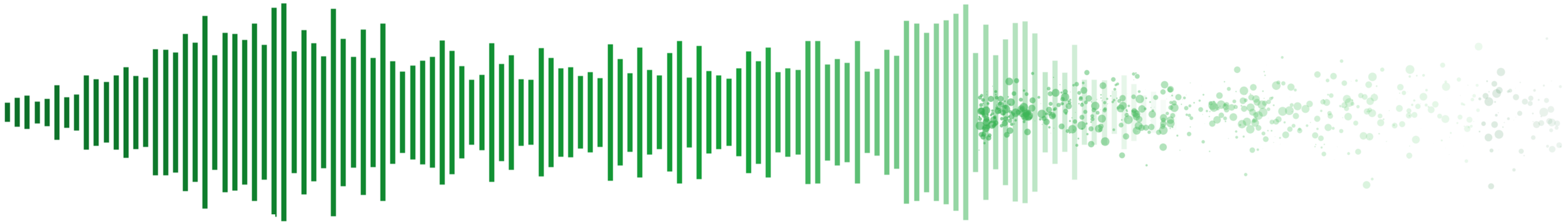
- Realtime Live Audio and Video
- Building pipeline
- Benchmarks and how to choose them

We are hiring

- Santa Clara (USA) and Toronto (Canada)

Two more things …

- d2l.smola.org (D2L second edition WiP)
- github.com/FeynRL-project (lightweight RL)



BosonAI