

VoxTurn: Answering Several Turn-Taking Questions in One Pass with One Model

An open 0.8B speech LLM for end-of-turn detection and related voice-agent decisions

Yang Xiao

The University of Melbourne



Model & plugins



GitHub



Interactive demo

Abstract

A cascaded voice agent must decide, throughout each silence of its user, whether the user has finished their turn: answering too early cuts the user off, answering too late makes the agent slow. It often needs related decisions too: will the user soon go on, is speech during its own turn an interruption or a backchannel, is a sound speech at all. With open detectors, each extra decision costs another model or another call, and on the 14-language eot-bench they cut users off in more than a quarter of mid-turn pauses. Because they are trained on segmented utterances, this mismatch is the largest single source of error. We present VoxTurn, a 0.8B speech LLM that casts these decisions as multiple-choice questions, scores the hidden state of each option, and answers any set of them in one forward pass with a block-isolating attention mask, exactly as if asked one at a time. Trained on real conversational pauses with labels derived from their timing, and re-queried while a silence lasts, VoxTurn reaches 11.5% false cut-offs in English at a 300 ms latency budget, comparable to two commercial detectors, and a 14-language mean of 19.9%, the lowest among open models, at 8.3 ms per decision. In Pipecat and LiveKit it cuts off 17% and 9% of turns, and it ranks fifth of 24 in TurnBench end-of-turn detection without any TurnBench training data. The model, code and framework plugins are released.

11.5%

English false cut-offs at 300 ms
on eot-bench

19.9%

14-language mean, lowest
among open models

8.3 ms

per decision; 4 questions in
9.0 ms

#5 / 24

TurnBench end-of-turn,
zero-shot

1 Introduction

Turn-taking is the timing backbone of spoken dialogue [3]. A cascaded voice agent listens while the user speaks and, whenever the user falls silent, must decide whether to answer: answering too early cuts the user off mid-thought, answering too late makes the conversation sluggish. A voice activity detector (VAD) [4] with a fixed silence timeout can only trade one error for the other, so recent systems add an end-of-turn (EOT) model that judges what was said and how. These range from text-only language models over the transcript [5, 6] and small audio classifiers [7] to larger audio models in commercial products [6, 8, 9] and in the open [10]. EOT-Bench [1] now compares them on real user turns in 14 languages, and TurnBench [2] on two-party conversations.

Although many works address turn-taking, open detectors leave three problems.

1. **They lag behind commercial detectors, most clearly beyond English.** On eot-bench, their 14-

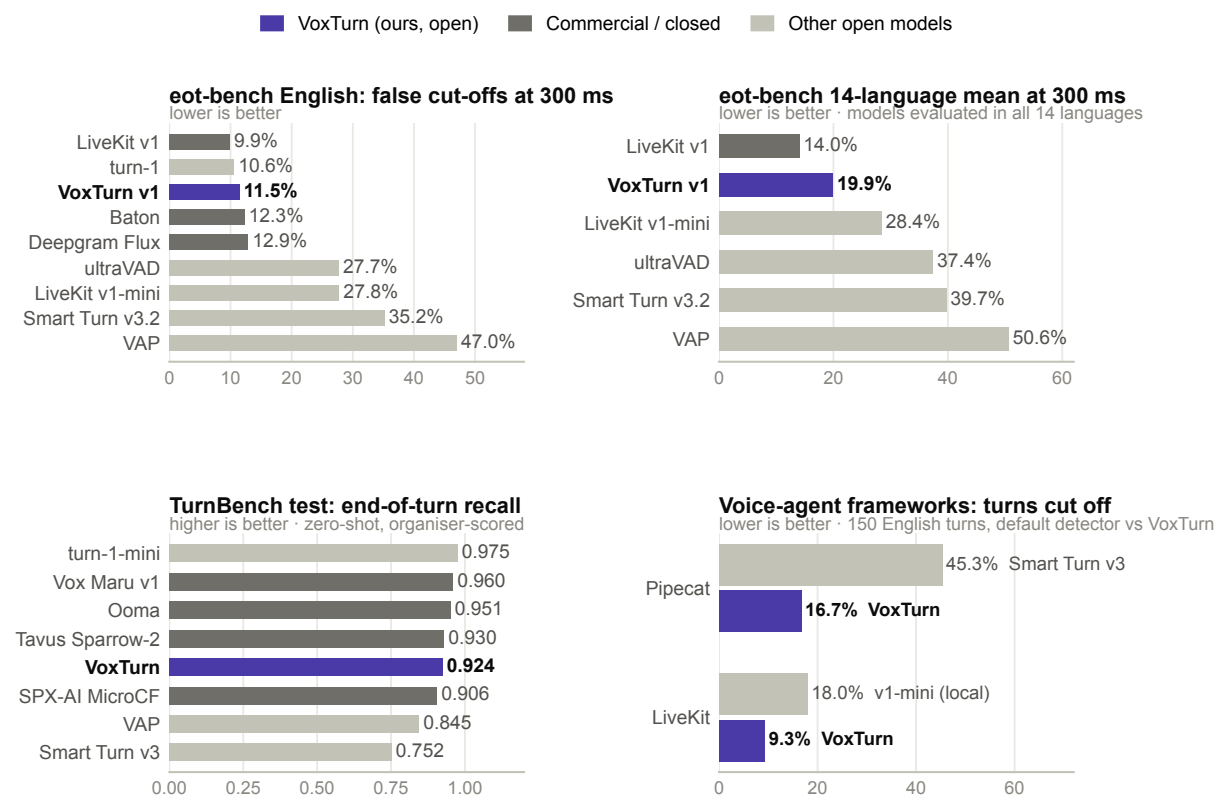


Figure 1. VoxTurn at a glance. Top: eot-bench [1], the share of mid-turn (hold) silences in which a detector takes the turn when it must answer within a 300 ms mean latency budget, in English (left) and averaged over 14 languages (right; only systems evaluated in all 14 languages). Bottom left: TurnBench [2] test-set end-of-turn recall, scored by the organisers; VoxTurn is applied zero-shot. Bottom right: share of 150 English turns that the agent would cut off, replayed through each framework’s VAD and endpointing with its default turn detector or with VoxTurn. Baseline numbers are official benchmark results or model cards.

Contents

1	Introduction	1
2	Related work	3
3	VoxTurn	4
3.1	Turn-taking decisions as multiple-choice questions	4
3.2	Model and readout	4
3.3	One-pass multi-question inference	5
3.4	Training objective	5
3.5	Deployment: re-querying during a silence	5
4	Training data and recipe	6
5	Evaluation suite	6
6	Results	7
6.1	eot-bench English: the latency–accuracy trade-off	7
6.2	Multilingual results	8
6.3	What drives the gains	9
6.4	Other decisions from the same pass	9
6.5	Two-party conversations: TurnBench	11
6.6	Efficiency and use in voice-agent frameworks	12
7	Conclusion	13
A	Per-language eot-bench results	16
B	Training hyperparameters	16
C	TurnBench decision policy	16

language mean is 28% or more false cut-offs at a 300 ms budget. We trace much of this gap to training data. Widely used training sets consist of segmented utterances, many of them synthetic [7, 11], each labelled once at its end, while a deployed detector is asked at varying times inside real pauses, where the answer depends on how long the silence has lasted.

2. **A deployed agent needs more than one decision:** whether a pause will end soon, whether speech during its own turn is an interruption or a backchannel, and whether a sound is speech at all. Each open detector answers one of these per call, so every extra answer costs another model or another call.
3. **Frameworks query the detector once,** when the VAD reports silence, and fall back to a timeout if the answer is not yet confident, although a good model grows more certain as the silence goes on.

VoxTurn addresses each problem. For the first, we train on real conversational pauses from phone calls and face-to-face dialogue, draw each decision time from anywhere in the pause, and derive labels from the pause timing. For the second, we write each decision as a multiple-choice question for a speech LLM built on Qwen3-ASR-0.6B [12], score the hidden state of each option instead of generating text, and answer any set of questions in one forward pass with an attention mask that isolates question blocks after a shared audio prefix, with exactly the result of asking them one by one. The same pass lets us supervise every training sample with all labels its timing determines. For the third, we re-query VoxTurn every 0.25 s while the silence lasts.

Key results

- ▶ **Real pauses matter most.** Real conversational pauses, with decision times drawn throughout each pause and labels derived from its timing, bring most of the gain on eot-bench: English false cut-offs fall from 26.2 to 16.5% and the 14-language mean from 33.0 to 20.0%, more than a speech-aligned backbone, multi-question training or four times more training (§6.3).
- ▶ **One pass, five decisions.** A one-pass multi-question readout, exactly equivalent to asking the questions one by one, lets one 0.8B model answer five turn-taking decisions in 8.3 ms per decision, and four questions in 9.0 ms instead of 33.2 ms (§3.3, §6.6).
- ▶ **Best open model on eot-bench.** 11.5% English false cut-offs at 300 ms, comparable to two commercial detectors, and a 19.9% 14-language mean, the lowest among open models evaluated in all 14 languages (§6.1, §6.2).
- ▶ **Works in real frameworks.** Re-queried during each silence, VoxTurn cuts off 17% and 9% of turns in Pipecat and LiveKit instead of 45% and 18%, and answers sooner (§6.6).
- ▶ **Transfers to two-party conversation.** Without TurnBench training, it ranks fifth of 24 in end-of-turn and sixth of 23 in interruption detection on the TurnBench test leaderboard (§6.5).

2 Related work

Turn-taking models. Voice Activity Projection [13] predicts the future voice activity of both speakers from dialogue audio, and full-duplex speech models such as Moshi [14] make turn-taking implicit. Cascaded agents instead use a dedicated detector on the user channel: TurnGPT [5] and the LiveKit text detector [6] score the transcript, Smart Turn [7] classifies the last seconds of audio, and turn-1 [10] combines speech-model features with the agent’s previous utterance. Easy Turn [11] and TEN [15] extend the binary decision to dialogue states such as backchannel and wait, and the SemanticVAD data [16] label dialogue text as complete, incomplete, interruption or backchannel. VoxTurn covers these decisions with one model and one pass.

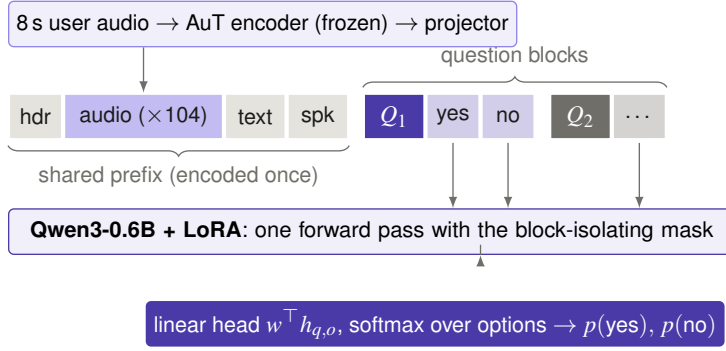
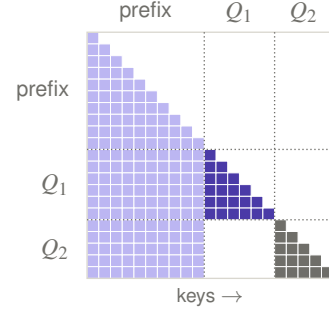
(a) Sequence and readout**(b) Attention mask and position ids****position ids**prefix $0, \dots, P-1$ Q_1 $P, \dots, P+|Q_1|-1$ Q_2 $P, \dots, P+|Q_2|-1$

Figure 2. Architecture. (a) VoxTurn reads the last 8 s of user audio (and optional text) once and answers several questions in one forward pass; a linear head scores the hidden state at each option’s last token. (b) Attention mask (filled: the query row attends to the key column) and position ids: each question block sees only the prefix and itself, at the positions it would have if asked alone.

LLMs as classifiers. Scoring answer options instead of generating free text is a standard way to use language models for multiple-choice questions [17]. VoxTurn applies it to speech, reading a learnt linear head on the hidden state of each option, and adds a mask that lets several questions share one encoding of the audio.

Benchmarks. eot-bench [1] scores detectors on real single-speaker user turns in 14 languages, asking for a decision repeatedly as each silence unfolds. TurnBench [2] scores end-of-turn and interruption events in two-party conversations with one channel per speaker. We report both, together with the test sets of Easy Turn [11] and TEN [15] and a nonverbal-vocalisation set [18] (Section 5).

3 VoxTurn

3.1 Turn-taking decisions as multiple-choice questions

At a decision time t , VoxTurn sees the 8 s of user audio ending at t (left-padded with silence), optionally a transcript or the partner’s last words, and a marker of who is speaking (user or agent). It answers five questions, each with a fixed option set (Table 1). Every question has 12 phrasings in English and in Chinese; training samples one, inference uses a fixed one.

3.2 Model and readout

The backbone is Qwen3-ASR-0.6B [12]: its AuT audio encoder (frozen) turns 128-bin log-mel features into 13 tokens per second (104 for 8 s), a projector maps them into Qwen3-0.6B [19], and we add LoRA [20] (rank 16, $\alpha = 32$) to every linear layer of the language model. The input sequence is [chat header][audio][text][speaker marker] followed by the questions, each written as “Question: ... Options: ...”. Nothing is generated: a linear head w scores the final hidden state $h_{q,o}$ of each option o of question q , and

$$p(o | q) = \text{softmax}_o(w^\top h_{q,o}) \quad (1)$$

Table 1. The five questions VoxTurn answers, with their option sets and the training sources that supervise them (Table 2).

Question	What it asks	Options	Supervised by
eot	Has the user finished their turn?	yes / no	Smart Turn, pause sources, SemanticVAD
resume	When will the user go on?	within 0.5 s / later / not in this turn	pause timing (auxiliary)
state4	Easy Turn dialogue state	complete / incomplete / backchannel / wait	Easy Turn
intent	Is speech during the agent’s turn an interruption?	interruption / backchannel	SemanticVAD
invalid	Is the sound not valid input (cough, laughter, noise, music)?	yes / no	MUSAN, VocalSound; real speech as negatives

over that question’s options. Only the projector, the LoRA weights and the head (11.8M parameters) are trained.

3.3 One-pass multi-question inference

Let the prefix occupy positions $1..P$ and question block k occupy a contiguous span B_k after it (Figure 2b). We use the attention mask

$$M_{ij} = \mathbf{1}[j \leq i] \cdot \mathbf{1}[j \leq P \vee \text{blk}(i) = \text{blk}(j)], \quad (2)$$

and give the m -th token of every block the position id $P + m - 1$, the id it would have if its block were the only question. Prefix tokens never attend to a block, so their states do not depend on the questions; by induction over layers, every token of B_k then attends to exactly the keys, at exactly the positions, that it would see in the sequence $[\text{prefix}; B_k]$. With a full-attention backbone, one pass therefore returns the same probabilities as K separate passes (we measure a maximum difference of 1.5×10^{-5} in fp32), with no dependence on question order. The prefix (≈ 120 tokens) is encoded once and each extra question adds about 15 tokens. The argument does not hold for recurrent or linear-attention layers, whose state cannot be masked per block, which is one reason we use a full-attention backbone.

3.4 Training objective

Each training sample has a main question from its source (Table 2). With the one-pass readout we add, at almost no cost, every auxiliary question whose label is exact: all real speech is labelled *not invalid*, and every pause sample gets a *resume* label from its timing. For a decision time o into a pause of length d , a hold pause with remaining silence $d - o \leq 0.5$ s is *within 0.5 s*, a longer one is *later*, and an end-of-turn pause is *not in this turn*. We also ask the main question a second time with another phrasing in the same pass. The loss is

$$\mathcal{L} = \text{CE}_{\text{main}} + \lambda_{\text{aux}} \overline{\text{CE}}_{\text{aux}} + \lambda_{\text{cons}} \frac{1}{2} [\text{KL}(p \parallel p') + \text{KL}(p' \parallel p)], \quad (3)$$

where p, p' are the main question’s answers under the two phrasings and $\lambda_{\text{aux}} = \lambda_{\text{cons}} = 0.5$.

3.5 Deployment: re-querying during a silence

For serving we merge the LoRA weights, run the encoder on fixed 8 s windows without host synchronisation, and replay CUDA graphs captured per input shape. In a voice agent, VoxTurn is asked when the VAD reports a pause and then again every 0.25 s while the silence lasts; the turn ends as soon as

Table 2. Training mixture: source, language, main question, size (clips, or pauses for the conversational sources) and sampling share. Each pause sample uses a fresh random decision time. Shaded: real conversational pauses.

Source	Type	Lang.	Question	Size	Share
Smart Turn v3.2 [7]	segmented utterances	23 langs	eot	266k	18%
Easy Turn [11]	segmented utterances	zh	state4	1.19M	25%
MagicData-RAMC [21]	conversational pauses	zh	eot	181k	11%
SpokenWOZ [22]	conversational pauses (phone)	en	eot	95k	14%
Seamless Interaction [23]	conversational pauses (face-to-face)	en	eot	179k	14%
SemanticVAD [16]	dialogue text	en, zh	eot, intent	15k	4%
MUSAN [24], VocalSound [25]	non-speech sounds	–	invalid	17k	14%

$p(\text{eot} = \text{yes}) \geq \theta$ (default $\theta = 0.8$), with the framework’s timeout as a fallback. This uses the model’s growing confidence during a pause, which a single check discards.

4 Training data and recipe

Data. Table 2 lists the mixture. Smart Turn and Easy Turn provide segmented utterances, much of them synthetic. The pause sources are real conversations: for SpokenWOZ (8 kHz task-oriented phone calls, user channel) and Seamless Interaction (face-to-face, each speaker’s own track) we run Silero VAD [4] on both channels and label every silence of at least 100 ms between two segments of a speaker as *end of turn* if the other speaker takes the floor (≥ 1 s of speech) and as *hold* if nobody speaks before the speaker resumes within 5 s; overlaps and backchannels are skipped. RAMC pauses come from its time-aligned transcripts. For every pause the decision time is drawn uniformly from the first $\min(d, 1.5 \text{ s})$ of the silence, so the model learns how the answer changes as a pause unfolds. Seamless audio is band-limited to 8 kHz half of the time so that bandwidth does not reveal the source. Easy Turn samples are audio only, audio with transcript or transcript only (6:3:1); invalid negatives are real speech from the other sources; labels within a source are rebalanced by n_{label}^α .

Training. We train on 300k samples with batch 128 (4 micro-batches of 32), AdamW ($\text{lr } 2 \times 10^{-4}$, $\beta = (0.9, 0.98)$, weight decay 0.01), 100 warm-up steps and cosine decay to 10%, gradient clipping at 1.0 and bf16 autocast (Appendix B). One run takes 78 minutes on one NVIDIA L40S. Re-running the r3 recipe on different hardware changed English false cut-offs by 0.1–0.5 points, which we take as the run-to-run noise.

5 Evaluation suite

Table 3 summarises what each benchmark measures. The headline benchmark is eot-bench, whose protocol is worth spelling out because it differs from utterance-level classification.

The eot-bench protocol. eot-bench [1] contains user turns in 14 languages with their silences labelled *hold* (the user goes on) or *end of turn*. A model is scored every 100 ms into each silence (minimum silence 100 ms), and the eot-harness sweeps decision policies (threshold, action delay, timeout). A *false cut-off* is a hold silence in which the policy takes the turn; the *latency* is the mean time from the start of an end-of-turn silence to the response. For each detector we report the lowest false cut-off rate at a mean response latency of 300 ms and 600 ms, the lowest mean latency at 5% and 10% cut-offs, and the AUC between hold and end-of-turn silences 200 ms into the pause. Every number is therefore a point on a trade-off curve (Figure 3a), and lower is better for both cut-offs and latency.

Table 3. Evaluation suite. “Turns” are single-speaker user turns; “FPR” is the false-positive rate. No evaluation audio is used for training.

Benchmark	What is tested	Languages	Metric
eot-bench [1]	hold vs. end-of-turn silences in real user turns, scored every 100 ms into each silence	14	false cut-offs at a 300 / 600 ms latency budget; latency at 5 / 10% cut-offs; AUC at 200 ms
TurnBench [2]	end-of-turn (EOT) and interruption (INT) events in 30 h of two-party conversation	en	recall at an FPR, median latency
Easy Turn test [11]	four dialogue states from audio	zh	accuracy
TEN test [15]	finished vs. unfinished turns from text	en, zh	per-class recall
OpenSLR 99 [18]	nonverbal vocalisations (16 categories) vs. 1,786 valid speech clips	–	flagged / falsely rejected
Framework replay	150 eot-bench English turns through Pipecat [26] and LiveKit Agents [27]	en	turns cut off, mean response delay

Table 4. eot-bench. English false cut-offs (%) at a mean-latency budget and mean latency (ms) at a cut-off budget; Chinese and the 14-language mean at 300 ms. Lower is better everywhere. –: not supported or not reported. Sorted by English @300.

System	Open	Params	English				Chinese	14 lang.
			cut-offs (%) at		latency (ms) at		@300	mean
			300 ms	600 ms	5%	10%		
LiveKit v1 [6]	–	–	9.9	4.5	543	295	20.6	14.0
turn-1 [10]	✓	2.8B	10.6	4.5	553	322	–	–
VoxTurn v1	✓	0.8B	11.5	6.0	669	335	22.9	19.9
Baton [8]	–	–	12.3	4.8	577	350	–	–
Deepgram Flux [9]	–	–	12.9	9.9	1151	548	–	–
ultraVAD [28]	✓	–	27.7	11.9	899	663	38.8	37.4
LiveKit v1-mini [6]	✓	–	27.8	12.1	1071	698	27.5	28.4
Smart Turn v3.2 [7]	✓	–	35.2	14.8	1051	739	42.9	39.7
VAP [13]	✓	–	47.0	14.6	1131	749	51.2	50.6

Baselines. Baseline numbers are the official eot-bench results; turn-1’s are from its model card [10]. TurnBench numbers are from the organisers’ test leaderboard (October 2026). eot-bench audio is never used for training.

6 Results

6.1 eot-bench English: the latency–accuracy trade-off

Table 4 compares VoxTurn with the systems on eot-bench. In English, VoxTurn cuts the user off in **11.5%** of hold silences at a 300 ms latency budget, fewer than the commercial Baton (12.3%) and Deepgram Flux (12.9%) and behind only LiveKit’s cloud detector and the $3.5\times$ larger turn-1.

Figure 3a shows the full trade-off. Three groups emerge. LiveKit v1, Baton and VoxTurn form the lowest band at every budget; Deepgram Flux is close to them at tight budgets but flattens beyond about 500 ms; and the other open models (LiveKit v1-mini, Smart Turn v3.2) need $1.6\text{--}2.2\times$ the latency of VoxTurn for the same 5% or 10% cut-off rate. At tight budgets VoxTurn tracks the commercial detectors, and it stays

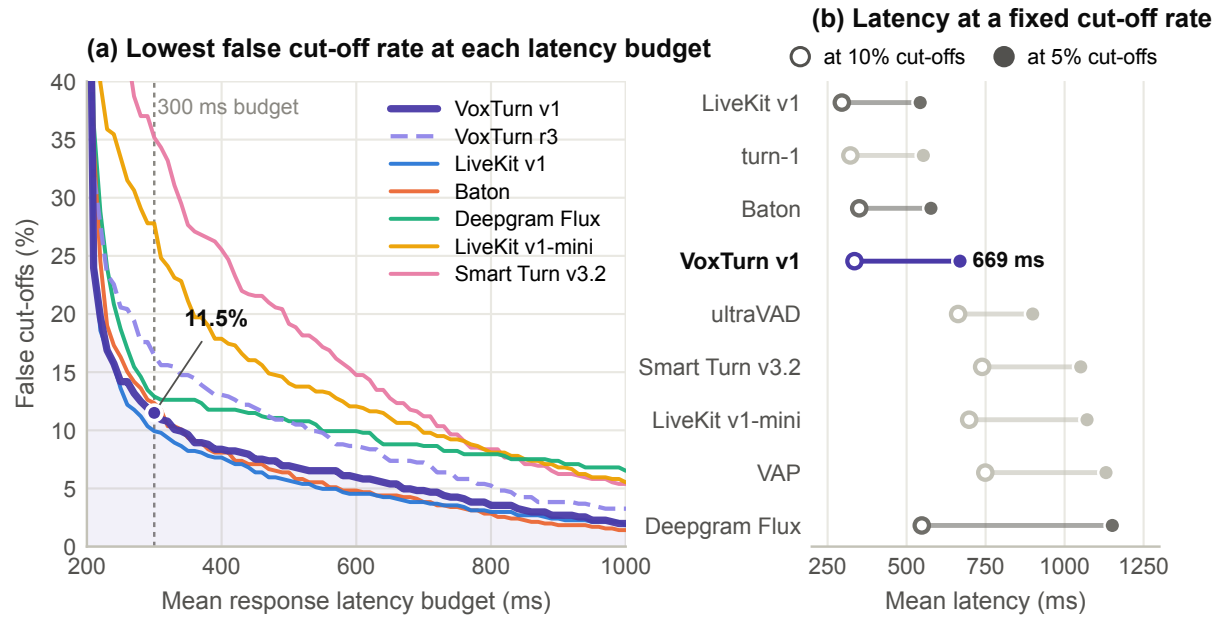


Figure 3. eot-bench English. (a) Lowest false cut-off rate at each mean-latency budget, from the eot-harness policy sweep; the shaded area is under the VoxTurn v1 curve. VoxTurn r3 is our earlier Whisper-based model (Table 5). (b) Mean latency each system needs to keep false cut-offs at 10% (open circle) and 5% (filled circle); shorter is better.

	ar	de	en	es	fr	hi	id	it	ja	ko	nl	pt	tr	zh	mean
LiveKit v1 (commercial)	16.2	11.8	9.9	15.5	12.7	10.5	22.3	16.9	5.6	5.5	12.5	20.1	15.7	20.6	14.0
VoxTurn v1	21.1	15.4	11.5	21.4	13.4	37.5	30.8	21.0	11.1	12.8	15.9	24.1	20.1	22.9	19.9
LiveKit v1-mini	24.4	24.8	27.8	26.0	22.4	27.4	43.5	33.7	23.5	12.3	32.9	36.9	35.3	27.5	28.4
ultraVAD	45.5	35.5	27.7	41.6	34.4	39.1	45.0	40.4	12.0	33.8	42.3	42.8	44.9	38.8	37.4
Smart Turn v3.2	39.2	35.5	35.2	38.7	38.9	42.6	46.2	44.1	34.9	30.1	42.3	42.7	43.3	42.9	39.7
VAP	54.6	53.6	47.0	52.6	54.5	49.5	55.3	50.3	48.8	43.5	42.3	55.3	50.4	51.2	50.6

Color scale: false cut-offs at 300 ms (%) from 0 to 50.

Figure 4. eot-bench, 14 languages: false cut-offs (%) at a 300 ms latency budget for every system evaluated in all 14 languages, rows sorted by the mean. Outlined: the best open model in each column. Only English and Chinese (purple) have conversational-pause training data.

far below the other open models at every budget. Figure 3b reads the same curves the other way: at 5% cut-offs VoxTurn answers after 669 ms, about 400 ms sooner than Smart Turn and LiveKit v1-mini, and at 10% cut-offs after 335 ms, within 40 ms of LiveKit v1.

6.2 Multilingual results

Across the 14 languages VoxTurn has a mean of **19.9%**, against 28.4% for the next open model (Table 4, Figure 4). It is the best open model in 12 of the 14 languages; only in Hindi and Korean is LiveKit v1-mini lower (27.4 vs. 37.5% and 12.3 vs. 12.8%). Its strongest languages are Japanese (11.1%), English (11.5%), Korean (12.8%) and French (13.4%), and ten of the 14 languages are below 22%. Read at a fixed 5% cut-off rate instead, VoxTurn answers sooner than every other open model in all 14 languages (Appendix A, Table 8).

The multilingual gain came with the real pause data: adding English and Chinese conversational pauses (r0 → r3 in Table 5) lowered the 14-language mean from 33.0% to 20.0%, although no pause data was

Table 5. The path to v1 (eot-bench, 300k training samples each). Rows r0–r3 use Whisper-small [29] with Qwen3.5-0.8B; r3[†] is r3 re-run on the L40S, the control for the rows below it.

Configuration	English			Chinese	14 lang.
	@300 (%) ↓	@5% (ms) ↓	AUC ↑	@300 (%) ↓	mean (%) ↓
r0: segmented utterances, RAMC at 0.2 s	26.2	921	.896	42.1	33.0
r1: + RAMC decision time within pause	22.6	936	.901	32.7	–
r2: + SpokenWOZ pauses	16.3	850	.944	29.1	–
r3: + Seamless pauses	16.5	808	.940	26.5	20.0
4× more training (1.2M)	15.9	833	.948	24.8	21.0
r3 [†]	16.6	855	.939	26.3	20.5
MLP head + pitch branch	18.4	966	.927	29.6	21.6
Qwen3-ASR-0.6B backbone	12.8	709	.961	21.9	19.7
+ multi-question (v1)	11.5	669	.963	22.9	19.9

added in the other twelve languages. The pause cues learnt from these conversations therefore transfer to other languages.

6.3 What drives the gains

Table 5 and Figure 5 trace the development.

Data first. With segmented utterances (r0), AUC stayed below 0.90 whatever we tuned. Drawing RAMC decision times from anywhere in a pause (r1) helped Chinese most (−9.4 points), and adding real English pauses from task-oriented phone calls (r2) gave the largest single English gain: 26.2 → 16.3% false cut-offs and AUC 0.896 → 0.944. Face-to-face pauses (r3) added little in English but improved Chinese and latency. The pause data, not the amount of training, is what matters: four times more updates gain only 0.7 points, and neither a larger readout head nor an explicit pitch branch helps.

Then the backbone. The second large step is the backbone. Replacing Whisper-small and a lightly trained adapter with Qwen3-ASR-0.6B, whose encoder and language model are already aligned on large-scale speech, lowers English cut-offs from 16.6 to 12.8% and the Chinese calibration error of $p(\text{eot})$ on eot-bench from 0.30 to 0.11.

Then multi-question training. Multi-question training improves English at every latency budget between 250 and 500 ms by 0.4–1.8 points (11.5%, AUC 0.963) while adding the resume, invalid and consistency supervision; the English $p(\text{eot})$ is well calibrated (ECE 0.007). In Chinese it costs 1.0 point at 300 ms and the 14-language mean is nearly unchanged (19.7 vs. 19.9%).

6.4 Other decisions from the same pass

The same model and pass also answer the other questions (Figure 6). On the Easy Turn test set, VoxTurn reaches 88.3% accuracy over the four dialogue states (complete 91.7, incomplete 87.0, backchannel 84.0, wait 86.0) from audio alone. It flags 99.6% of the OpenSLR 99 nonverbal vocalisations (coughs, laughter, sighs and 13 other categories) as invalid input while rejecting only 0.06% of 1,786 valid speech clips. On TEN’s text test set it recognises 83.5% (English) and 89.7% (Chinese) of finished turns and 99.0% and 97.9% of unfinished ones.

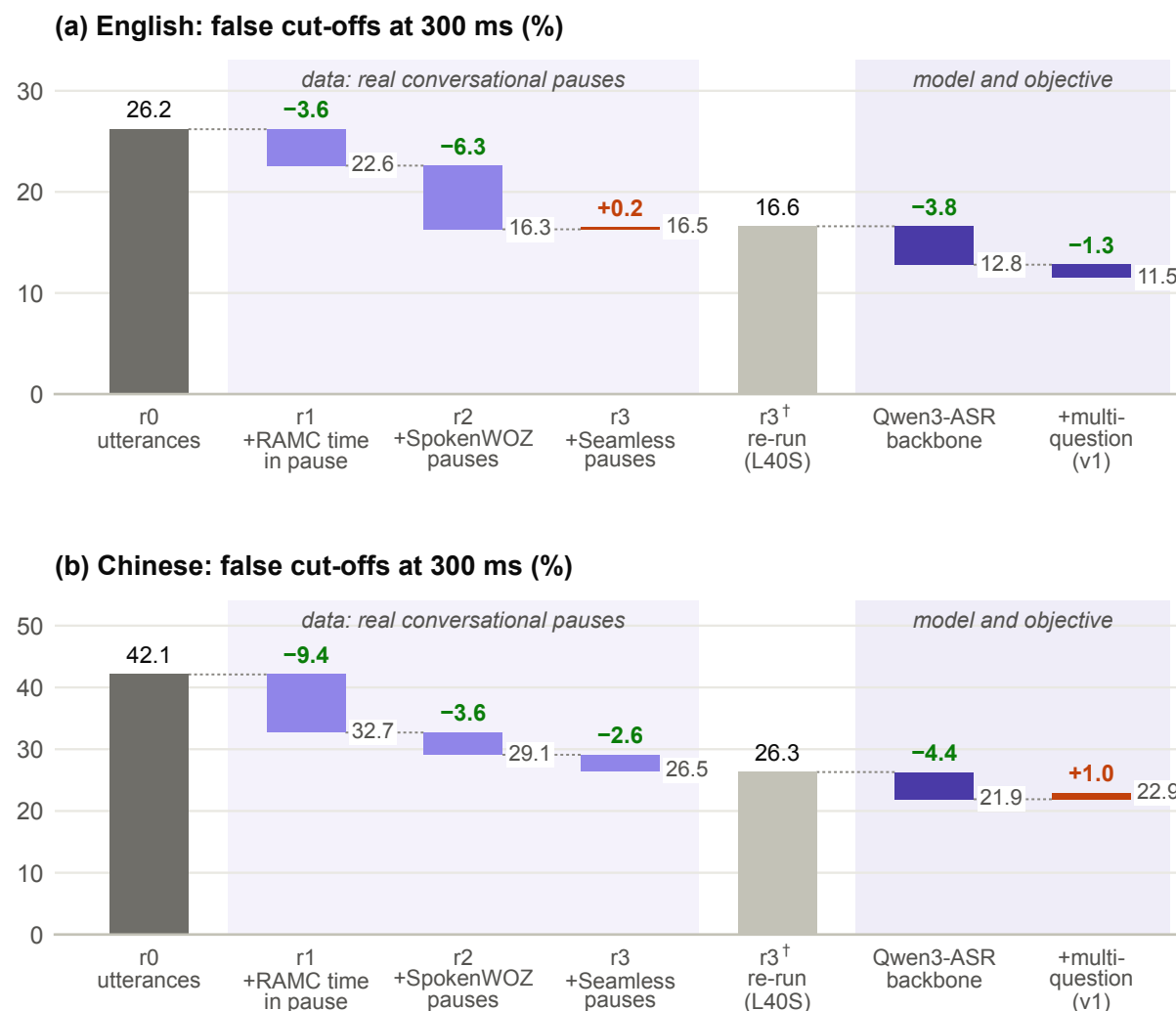


Figure 5. The path to v1 as a waterfall. Floating bars show the change in false cut-offs at 300 ms from the step before; the number above each gives the change (green: fewer cut-offs, orange: more) and the number beside it the new value. Grey bars are absolute values. $r3^\dagger$ re-runs r3 on different hardware and restarts the chain. Data steps (shaded left) bring the largest gains in both languages.

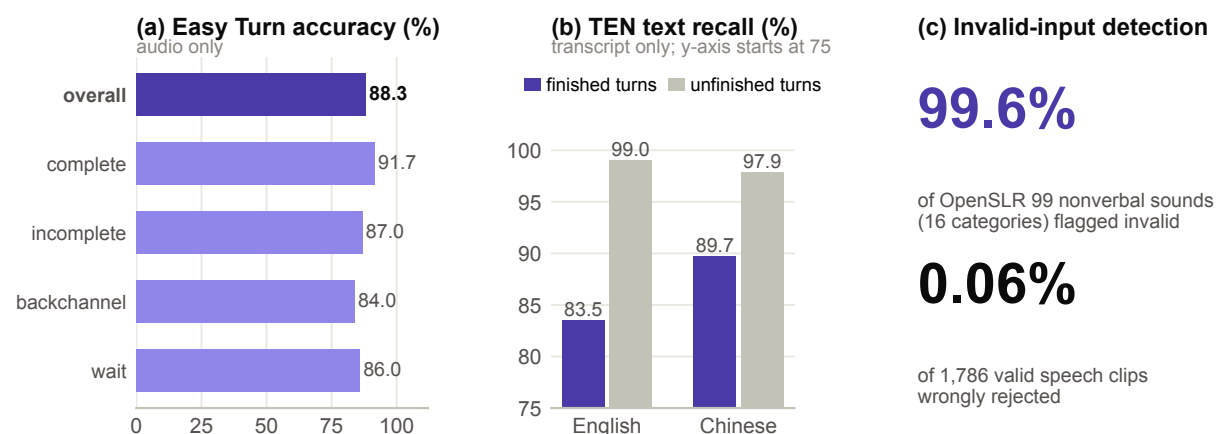


Figure 6. Other questions answered by the same model and pass. (a) Easy Turn test set, four dialogue states from audio alone. (b) TEN text test set, recall of finished and unfinished turns. (c) Invalid-input detection on OpenSLR 99 nonverbal vocalisations and on valid speech.

Table 6. TurnBench. Recall (R) at a false-positive rate (FPR) and median latency (ms). Top: test set, scored by the organisers (October 2026; †: external submission); ranks are among all submissions. Bottom: dev (38 conversations, $FPR \leq 0.1$) and EOT policy ablations re-searched on dev.

System	EOT			INT		
	R ↑	FPR	ms ↓	R ↑	FPR	ms ↓
turn-1-mini† [30]	.975	.075	595	.973	.048	551
Vox Maru v1†	.960	.072	548	.964	.085	368
Ooma†	.951	.082	388	.968	.076	567
Tavus Sparrow-2†	.930	.084	615	.958	.126	959
SPX-AI MicroCF†	.906	.079	678	.984	.073	747
VAP [13]	.845	.055	368	.945	.107	994
Smart Turn v3 [7]	.752	.047	1017	.107	.093	159
VoxTurn (EOT #5 of 24, INT #6 of 23)	.924	.076	557	.951	.112	332
<i>Dev set, EOT</i>						
VoxTurn, threshold, own channel	.700	.100	597	.960	.100	337
two-channel score	.810	.100	351			
+ decision policy	.930	.094	565			
Policy, re-searched	.925	.087	667			
– confirmation event	.895	.087	762			
– partner trigger	.885	.098	751			
VAD in place of VoxTurn	.924	.098	821			

■ VoxTurn (ours, open) ■ Commercial / closed ■ Other open models

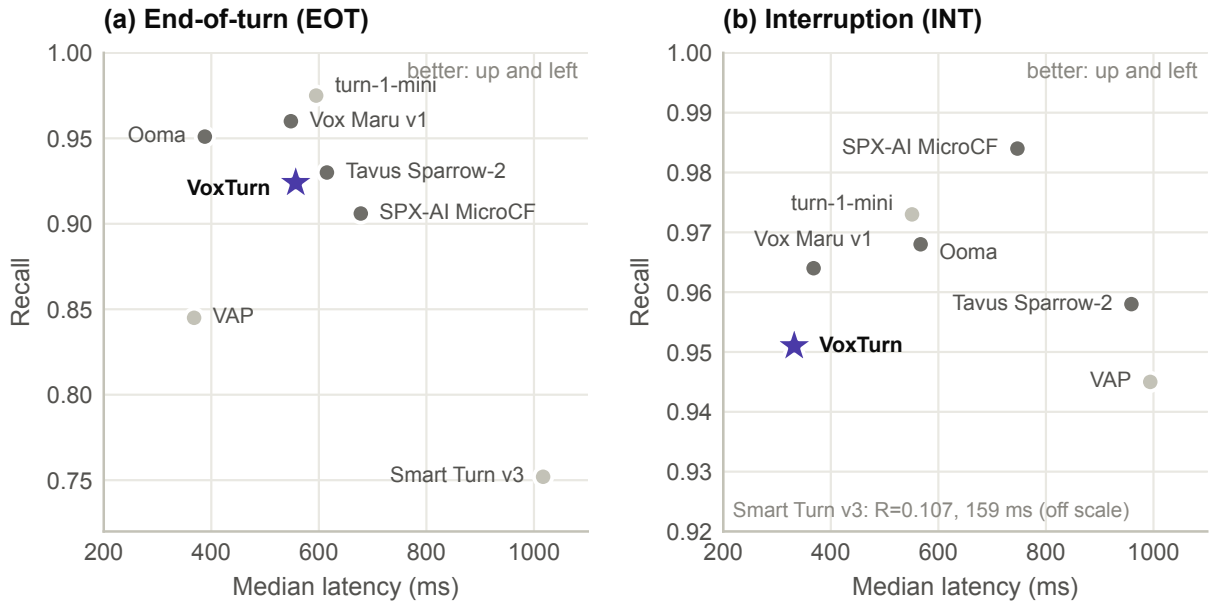


Figure 7. TurnBench test set: recall against median latency (upper left is better). Systems operate at different false-positive rates (Table 6). Smart Turn v3’s INT recall (0.107) is off the scale of panel (b).

6.5 Two-party conversations: TurnBench

TurnBench [2] has 30 h of triple-annotated two-party English conversations, one channel per speaker. Systems emit causal event times for **EOT** (a turn has ended) and **INT** (speech over the partner is a floor-taking interruption, not a backchannel); an event counts within $[-0.25, +3]$ s of a gold event, and

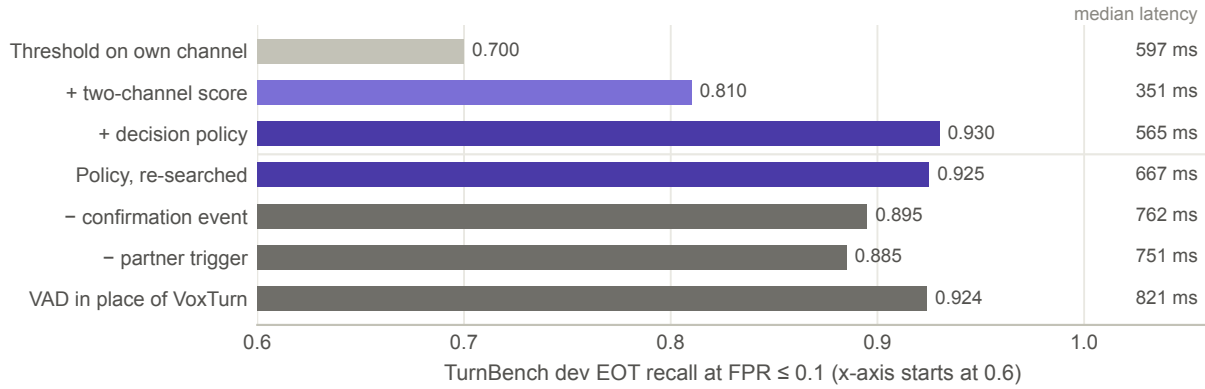


Figure 8. TurnBench dev, EOT decision policy. Recall at $FPR \leq 0.1$ as the policy is built up (top three) and when its parts are removed from the re-searched policy (bottom three); the right column gives the median latency.

Table 7. Efficiency and framework replay. Left: latency per call (L40S, bf16, 8 s audio, end to end, p50). Right: 150 eot-bench English turns through each framework’s Silero VAD and endpointing (LiveKit: VAD minimum silence 0.3 s); VoxTurn re-checks every 0.25 s. A turn is cut off if any of its pauses ends it; delay is the mean from the final silence.

Call	ms	Framework	Turn detector	Cut off	Delay
1 question, unmerged eager model	≈38	Pipecat	Smart Turn v3 (default)	45.3%	1.32 s
1 question, serving mode	8.3		VoxTurn	16.7%	0.36 s
4 questions in 4 passes	33.2	LiveKit	v1-mini (local)	18.0%	0.73 s
4 questions in one pass	9.0		VoxTurn	9.3%	0.47 s
4 clips in one batch, 1 question each	16.0				

firing in a mid-turn pause or backchannel is one false positive per span. VoxTurn runs zero-shot: at 10 Hz per channel it answers eot and resume in silences, and state4 and invalid in the first 2.5 s of speech begun near the partner’s; INT thresholds $p(\text{not backchannel})$ $p(\text{not invalid})$. For EOT we add a two-channel decision policy, described in Appendix C, whose ten parameters are random-searched for the best dev recall at $FPR \leq 0.1$.

Test leaderboard. On the held-out test set (Table 6, Figure 7), VoxTurn ranks fifth of 24 in EOT (0.924) and sixth of 23 in INT (0.951), and is the fastest of the INT top six (332 ms), well ahead of the strongest baseline in [2], VAP. turn-1-mini, first in EOT, reports 22.6% English false cut-offs on eot-bench [30] against 11.5% for VoxTurn.

What the policy adds. On dev (Figure 8), the partner’s speech lifts the own-channel EOT recall from 0.700 to 0.810, and the full policy to 0.930; the partner trigger and the confirmation event add 4.0 and 3.0 points. With VAD in place of VoxTurn the recall holds (0.924) but answers come later (821 vs. 667 ms).

6.6 Efficiency and use in voice-agent frameworks

Latency. One decision takes **8.3 ms** in serving mode (Table 7, 1.9 GB of GPU memory), and four questions in one pass take 9.0 ms instead of 33.2 ms, a $3.7\times$ saving whose answers differ from separate passes only by bf16 rounding. Serving mode keeps the eot-bench result (11.6% vs. 11.5% at 300 ms).

Framework replay. We integrated VoxTurn into Pipecat [26] as a turn analyzer and into LiveKit Agents [27] as a streaming audio turn detector, both with the 0.25 s re-check policy, and replayed 150 English eot-bench turns through each framework’s VAD and endpointing rules. A turn counts as cut off if the agent would have answered in any of its mid-turn pauses; delay is measured from the start of the final

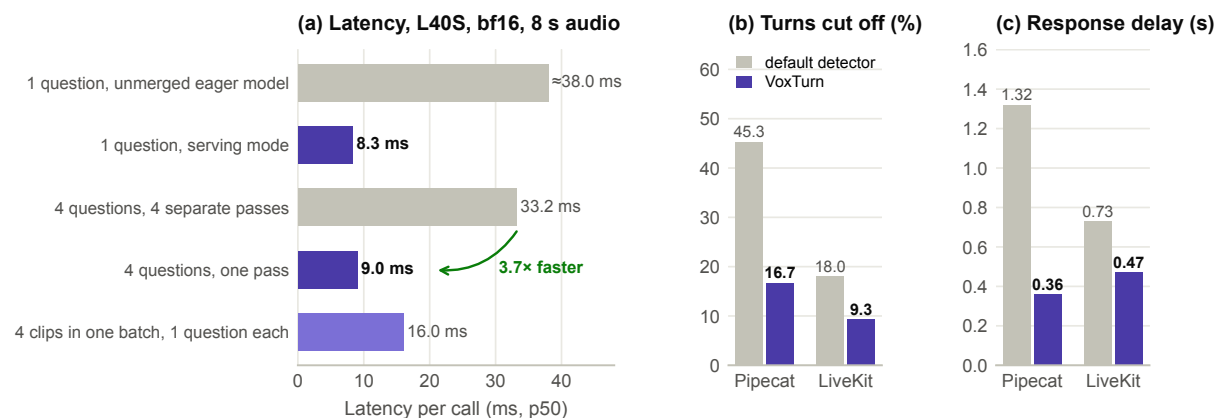


Figure 9. Efficiency and deployment. (a) Latency per call; answering four questions in one pass costs about as much as one question. (b, c) Framework replay of 150 English eot-bench turns: share of turns the agent would cut off and mean delay from the start of the final silence, with each framework’s default local detector and with VoxTurn re-checked every 0.25 s.

silence. In Pipecat, VoxTurn cuts off **16.7%** of turns against 45.3% with the default Smart Turn v3 and answers after 0.36 s instead of 1.32 s on average. In LiveKit, with the VAD’s minimum silence lowered to 0.3 s, it cuts off half as many turns as LiveKit’s local audio detector (**9.3%** vs. 18.0%) and answers sooner (0.47 s vs. 0.73 s; 90th percentile 0.78 s vs. 2.50 s).

The re-check matters. Asked only once (threshold 0.5), VoxTurn in Pipecat waits 1.26 s on average, because a pause that is not yet convincing when the VAD reports it falls back to the framework’s 3 s timeout. Replays of such turns in six languages, with both models’ probabilities over time, can be explored in the interactive demo.¹

7 Conclusion

VoxTurn answers several turn-taking questions with a speech LLM in one forward pass. Real conversational pauses, a speech-aligned backbone and supervision of every label a pause’s timing determines bring an open 0.8B model to 11.5% English false cut-offs at 300 ms and the best 14-language mean among open models, at 8.3 ms per decision, and to fifth place in TurnBench end-of-turn detection. Weights, code and plugins are at <https://huggingface.co/AustinXiao/voxtturn>, and the demo at <https://swagshaw.github.io/voxtturn/>.

Acknowledgements

Experiments used computational resources of the National Computational Infrastructure (NCI Australia) and of the University of Melbourne. Generative AI tools were used to assist with writing code, running and summarising experiments, and drafting and editing this report; the authors reviewed and take responsibility for all content.

References

- [1] LiveKit. eot-bench: LiveKit end-of-turn benchmark. <https://github.com/livekit/eot-bench>, 2026.

¹<https://swagshaw.github.io/voxtturn/>

- [2] Freeman Jiang, Ramon Sanabria, Soham Deshmukh, Bandhav Veluri, Simon Michael Vuch Williams, Elliott K. Suen, Garreth Lee, Kevin Yoonho Choi, Takuya Umeki, Riku Kubo, Sathvik Udupa, Chien-yu Huang, Shih-Yun Shan Kuan, Zhuoyan Tao, Satyapriya Krishna, Sefik Emre Eskimez, Yu Tsao, Hung-yi Lee, and Shinji Watanabe. TurnBench: A multi-domain benchmark for turn-taking dynamics in spoken dialogue. In *Proc. IEEE Spoken Language Technology Workshop (SLT)*, 2026. arXiv:2608.25218.
- [3] Gabriel Skantze. Turn-taking in conversational systems and human-robot interaction: A review. *Computer Speech & Language*, 67:101178, 2021. doi: 10.1016/j.csl.2020.101178.
- [4] Silero Team. Silero VAD: pre-trained enterprise-grade voice activity detector. <https://github.com/snakers4/silero-vad>, 2024.
- [5] Erik Ekstedt and Gabriel Skantze. TurnGPT: a transformer-based language model for predicting turn-taking in spoken dialog. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, 2020. arXiv:2010.10874.
- [6] LiveKit. LiveKit turn detector. <https://docs.livekit.io/agents/build/turns/turn-detector/>, 2026.
- [7] Pipecat. Smart turn: an open-source audio turn detection model. <https://github.com/pipecat-ai/smart-turn>, 2025.
- [8] JoinIn AI. Baton turn-taking engine. <https://joinin.ai>, 2026.
- [9] Deepgram. Introducing Flux: Conversational speech recognition. <https://deepgram.com/learn/introducing-flux-conversational-speech-recognition>, 2025.
- [10] p99lab. turn-1. <https://huggingface.co/p99lab/turn-1>, 2026.
- [11] Guojian Li, Chengyou Wang, Hongfei Xue, Shuiyuan Wang, Dehui Gao, Zihan Zhang, et al. Easy turn: Integrating acoustic and linguistic modalities for robust turn-taking in full-duplex spoken dialogue systems. *arXiv preprint arXiv:2509.23938*, 2025.
- [12] Xian Shi, Xiong Wang, Zhifang Guo, Yongqi Wang, Pei Zhang, Xinyu Zhang, et al. Qwen3-ASR technical report. *arXiv preprint arXiv:2601.21337*, 2026.
- [13] Erik Ekstedt and Gabriel Skantze. Voice activity projection: Self-supervised learning of turn-taking events. In *Proc. Interspeech*, 2022. arXiv:2205.09812.
- [14] Alexandre Défossez, Laurent Mazaré, Manu Orsini, Amélie Royer, Patrick Pérez, Hervé Jégou, Edouard Grave, and Neil Zeghidour. Moshi: a speech-text foundation model for real-time dialogue. *arXiv preprint arXiv:2410.00037*, 2024.
- [15] TEN Framework. TEN turn detection. <https://github.com/TEN-framework/ten-turn-detection>, 2025.
- [16] KE-Team. SemanticVAD dataset. <https://huggingface.co/datasets/KE-Team/SemanticVAD-Dataset>, 2025.
- [17] Joshua Robinson, Christopher Michael Rytting, and David Wingate. Leveraging large language models for multiple choice question answering. In *Proc. ICLR*, 2023. arXiv:2210.12353.
- [18] Deeply Inc. Deeply nonverbal vocalization dataset (OpenSLR SLR99). <https://www.openslr.org/99/>.
- [19] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.
- [20] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *Proc. ICLR*, 2022. arXiv:2106.09685.

-
- [21] Zehui Yang, Yifan Chen, Lei Luo, Runyan Yang, Lingxuan Ye, Gaofeng Cheng, et al. Open source MagicData-RAMC: A rich annotated Mandarin conversational (RAMC) speech dataset. In *Proc. Interspeech*, 2022. arXiv:2203.16844.
 - [22] Shuzheng Si, Wentao Ma, Haoyu Gao, Yuchuan Wu, Ting-En Lin, Yinpei Dai, et al. SpokenWOZ: A large-scale speech-text benchmark for spoken task-oriented dialogue agents. In *Proc. NeurIPS Datasets and Benchmarks Track*, 2023. arXiv:2305.13040.
 - [23] Vasu Agrawal, Akinniyi Akinyemi, Kathryn Alvero, Morteza Behrooz, Julia Buffalini, Fabio Maria Carlucci, et al. Seamless interaction: Dyadic audiovisual motion modeling and large-scale dataset. *arXiv preprint arXiv:2506.22554*, 2025.
 - [24] David Snyder, Guoguo Chen, and Daniel Povey. MUSAN: A music, speech, and noise corpus. *arXiv preprint arXiv:1510.08484*, 2015.
 - [25] Yuan Gong, Jin Yu, and James Glass. VocalSound: A dataset for improving human vocal sounds recognition. In *Proc. ICASSP*, 2022. arXiv:2205.03433.
 - [26] Pipecat. Pipecat: open source framework for voice and multimodal conversational AI. <https://github.com/pipecat-ai/pipecat>, 2026.
 - [27] LiveKit. LiveKit agents. <https://github.com/livekit/agents>, 2026.
 - [28] Fixie AI. ultraVAD. <https://huggingface.co/fixie-ai/ultraVAD>, 2025.
 - [29] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *Proc. ICML*, 2023. arXiv:2212.04356.
 - [30] p99lab. turn-1-mini. <https://huggingface.co/p99lab/turn-1-mini>, 2026.

A Per-language eot-bench results

Table 8. eot-bench per language: false cut-offs (%) at a 300 ms mean-latency budget (top) and mean latency (ms) at 5% false cut-offs (bottom); lower is better. Bold: best open model per column.

	ar	de	en	es	fr	hi	id	it	ja	ko	nl	pt	tr	zh	mean
<i>False cut-offs at 300 ms (%)</i>															
LiveKit v1 (closed)	16.2	11.8	9.9	15.5	12.7	10.5	22.3	16.9	5.6	5.5	12.5	20.1	15.7	20.6	14.0
VoxTurn v1	21.1	15.4	11.5	21.4	13.4	37.5	30.8	21.0	11.1	12.8	15.9	24.1	20.1	22.9	19.9
LiveKit v1-mini	24.4	24.8	27.8	26.0	22.4	27.4	43.5	33.7	23.5	12.3	32.9	36.9	35.3	27.5	28.4
ultraVAD	45.5	35.5	27.7	41.6	34.4	39.1	45.0	40.4	12.0	33.8	42.3	42.8	44.9	38.8	37.4
Smart Turn v3.2	39.2	35.5	35.2	38.7	38.9	42.6	46.2	44.1	34.9	30.1	42.3	42.7	43.3	42.9	39.7
VAP	54.6	53.6	47.0	52.6	54.5	49.5	55.3	50.3	48.8	43.5	42.3	55.3	50.4	51.2	50.6
<i>Latency at 5% cut-offs (ms)</i>															
LiveKit v1 (closed)	706	672	543	643	635	454	832	785	321	324	610	703	718	799	625
VoxTurn v1	751	801	669	782	662	606	913	912	433	542	728	796	898	814	736
LiveKit v1-mini	811	947	1071	926	798	624	938	966	766	712	884	897	1080	919	881
ultraVAD	807	864	899	986	846	616	939	1005	462	731	910	849	1075	926	851
Smart Turn v3.2	814	963	1051	980	873	639	945	980	837	734	884	836	1013	959	893
VAP	839	1068	1131	1067	917	655	1009	1012	796	800	884	899	1120	942	938

B Training hyperparameters

Table 9. VoxTurn v1 training configuration.

Setting	Value
Backbone	Qwen3-ASR-0.6B (AuT encoder frozen, Qwen3-0.6B language model)
Trained parameters	projector, LoRA, linear head (11.8M)
LoRA	rank 16, $\alpha = 32$, every linear layer of the language model
Audio context	8 s, left-padded with silence; 13 tokens/s (104 tokens)
Training samples	300k
Batch	128 (4 micro-batches of 32)
Optimiser	AdamW, $\text{lr } 2 \times 10^{-4}$, $\beta = (0.9, 0.98)$, weight decay 0.01
Schedule	100 warm-up steps, cosine decay to 10%
Gradient clipping	1.0
Precision	bf16 autocast
Loss weights	$\lambda_{\text{aux}} = \lambda_{\text{cons}} = 0.5$
Decision time in a pause	uniform over the first $\min(d, 1.5 \text{ s})$
Compute	78 minutes on one NVIDIA L40S

C TurnBench decision policy

For EOT we add a two-channel *decision policy*: each silence opens a candidate, committed at the first frame where

- (i) the silence has lasted $d_{\min}$ and $p(\text{eot}) (1 - p(\text{resume} \leq 0.5 \text{ s})) \geq \theta_{\text{own}}$, or
- (ii) the partner, starting at most 1 s before or during the silence, has spoken for m s and VoxTurn reads it as

floor-taking ($p(\text{not backchannel}) p(\text{not invalid}) \geq \theta_{\text{floor}}$), or

(iii) a deadline passes.

Events wait 0.5 s after the speech offset. As a pause costs at most one false positive, a confirming event follows at silence t_{conf} , and the last silence ends with an event at the end of the audio. The ten parameters are random-searched for the best dev recall at FPR ≤ 0.1 .