Proactive Voices: Benchmarking Full-duplex Human-LLM Speech Interaction

Saihan Chen[1†], **Yuan Ge**[1†], **Jingqi Xiao**[1], **Xiaoqian Liu**[1],
Tong Xiao[1,2], **Zhengtao Yu**[3], **Yan Xiang**[3], **Jingbo Zhu**[1,2‡]

[1]School of Computer Science and Engineering, Northeastern University, Shenyang, China
[2]NiuTrans Research [3]Kunming University of Science and Technology
christinechensh@gmail.com

Abstract

Full-Duplex Speech Large Language Models (LLMs) are central to real-time spoken human-computer interaction. Yet benchmarking and modeling them remain difficult. We introduce FLEXI, a benchmark that shifts the evaluation focus from reactive responses to proactive interactions, explicitly incorporating model-initiated interventions such as emergency interruptions. Across six languages, FLEXI systematically evaluates the latency, quality, and conversational effectiveness of real-time dialogue through six human-LLM interaction scenarios. Evaluation of open source and commercial models reveals limited proactive interaction capabilities, particularly in initiating timely interventions and providing appropriate backchannels. We release code and data to support reproducible evaluation. Finally, we suggest that next token-pair prediction offers a promising path toward achieving truly seamless and human-like full-duplex interaction.

1 Introduction

Large language models (LLMs) have made conversational interfaces more capable, but spoken human-computer interaction is still far from the fluidity of everyday conversation (Clark, 1996; Luger and Sellen, 2016; Ge et al., 2025). A major limitation is that most voice interfaces remain turn-based: the system waits for a user utterance to end before planning and speaking. Full-duplex speech dialogue aims to relax this constraint by allowing the model to listen and speak at the same time (Hurst et al., 2024; Défossez et al., 2024; Wang et al., 2025b). In such settings, interaction is not purely reactive. The model may provide a backchannel while the user is speaking, raise an emergency interruption, or remain silent when speaking would be intrusive.

As shown in Fig. 1, proactive behavior in full-duplex interaction ranges from silence to lightweight feedback and stronger interventions. Such initiative is helpful only when its timing and strength match the interaction state; otherwise, early responses, unnecessary backchannels, or misplaced interruptions can disrupt the user experience. Existing duplex evaluations still focus mainly on user-initiated events, leaving limited analysis of when model-initiated behavior is appropriate, necessary, or harmful.

Motivated by this gap, we introduce FLEXI, a benchmark for **f**ull-dup**lex** human-LLM speech **i**nteraction, designed to evaluate S2S LLMs from a proactive interaction perspective. Our contributions are:

- We introduce FLEXI, a full-duplex benchmark encompassing six distinct human-LLM interaction scenarios from a proactive interaction perspective.
- We construct a multilingual evaluation set of 3,600 samples spanning English, Chinese, German, Spanish, French, and Italian, enabling cross-language analysis of full-duplex interaction.
- We evaluate six open source and commercial systems, identifying gaps in response timing, turn management, and proactive speech.
- We derive design implications for coordinating streaming comprehension, floor management, and lightweight feedback.

2 Related Work

Spoken interaction depends on turn-taking, pauses, and backchannels (Sacks et al., 1974; Skantze, 2021; Ward and Tsukahara, 2000; Wang et al., 2024). Full-duplex speech models address these

[†]Equal contribution.
[‡]Corresponding author.

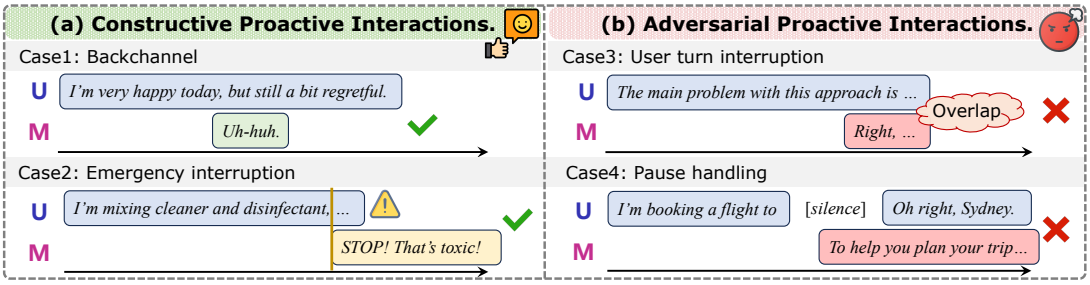


Figure 1: The spectrum of **proactivity** in full-duplex SLMs.

events through external control or scheduling modules (Fu et al., 2025; Chen et al., 2025; Zhang et al., 2025b; Liao et al., 2025; Lu et al., 2025a; Yang et al., 2026), or coupled input–output generation (Défossez et al., 2024; Wang et al., 2025b; Zhang et al., 2025c; Yu et al., 2025; Wang et al., 2025a; Hu et al., 2025). Recent work explores shared speech–text modeling for multilingual translation (Ou et al., 2025) and evaluates multilingual LLM translation on Southeast Asian languages (Lai et al., 2025). These studies concern multilingual content processing; FLEXI focuses on when a speech system should take, retain, or yield the floor.

Duplex benchmarks evaluate turn-taking, user interruption, overlap, and backchannel behavior (Arora et al., 2025; Lin et al., 2025b,a; Peng et al., 2025; Lin et al., 2026). ISBench evaluates speech-language models' instruction following and safety across acoustic conditions (Wang et al., 2026), complementing interaction-level evaluation. LLM judges support semantic scoring (Zheng et al., 2023; Zhou et al., 2024), while proactive agents study context-driven assistance (Lu et al., 2025b; Zhang et al., 2025a; Zhao et al., 2026; Yang et al., 2025). FLEXI connects semantic appropriateness with real-time floor management, distinguishing useful emergency intervention from premature takeover and brief feedback from full responses.

3 Method

3.1 Problem Setup

We formalize interaction events in FLEXI following Nguyen et al. (2023) and Arora et al. (2025). For simplicity, we consider a two-speaker turn-taking event, where $k \in \{1, 2\}$ indexes the two speakers: the **user** and the **model**. The state of speaker$_k$ at time t is defined as:

$$y_t^k = \begin{cases} 1, & \text{speaker}_k \text{ is speaking at time } t \\ 0, & \text{speaker}_k \text{ is } \textit{not} \text{ speaking at time } t \end{cases} \tag{1}$$

Further, we define the following concepts:

- *Inter-pausal units* (*IPUs*) are continuous stretches of speech that are separated by a silence on each side.

$$\text{IPU}^k = \{t \in [a, b] \mid y_t^k = 1\} \tag{2}$$

- *Silences* are the durations that both speakers are not speaking, where a^{sil} and b^{sil} denote the start and end time.

$$\text{SIL} = \{t \in [a^{sil}, b^{sil}] \mid y_t^1 = 0,\ y_t^2 = 0\} \tag{3}$$

- *Pauses* are silences between IPUs of the same speaker. Successive IPUs of the same speaker separated by pauses are grouped to form a *turn*. Formally, given that $\mathcal{T}_j = \{i_1, i_2, \ldots, i_n\}$ is a series of IPU indices forming the j-th turn, and each adjacent pair of IPUs is separated by silence, we define:

$$\begin{aligned} \text{Turn}_j^k = &\left(\bigcup_{r=1}^{n} [a_{i_r}^k, b_{i_r}^k] \right) \\ &\cup \left(\bigcup_{r=1}^{n-1} [b_{i_r}^k, a_{i_{r+1}}^k] \right) \end{aligned} \tag{4}$$

- *Gaps* are silences between IPUs of different speakers, typically occurring around a turn exchange.

- *Overlaps* are times where there are IPUs for both speakers.

$$\text{Overlap} = \{t \in [a^o, b^o] \mid y_t^1 = 1,\ y_t^2 = 1\} \tag{5}$$

- *Backchannels* are overlaps such as 'yeah' and 'exactly', which the listener utters without taking the speaker's turn to acknowledge the current speaker. *Interrupts* are overlaps between two speakers where both speakers are trying to take the turn. Formally, for $\text{IPU}^1 = [a^1, b^1]$ and $\text{IPU}^2 = [a^2, b^2]$, the IPU of speaker 2 interrupts speaker 1 when:

$$\text{IPU}^2_{\text{int}} = \{t \in [a^2, b^2]\} \quad \text{s.t. } a^1 < a^2 < b^1 \tag{6}$$

- An *interrupt* is a successful interruption if $b^1 < b^2$, and an unsuccessful interruption if $b^1 > b^2$. In this paper, interrupt indicates a successful interrupt.

3.2 Evaluation Categories

Based on this formalization, we categorize full-duplex user–LLM interactions into six scenarios from a human–computer interaction perspective.

1) Standard turn-taking. This scenario mirrors traditional turn-based dialogue, where speakers alternate turns separated by short gaps. It evaluates whether the model can detect the end of a user's turn and respond promptly.

Metric: Following Lin et al. (Lin et al., 2025b), Takeover Rate (TOR) is used to measure how often the model controls the conversation and dominates the turn. TOR is averaged across the dataset:

$$\begin{aligned} \text{TOR} &= \frac{1}{N}\sum_{i=1}^{N} \text{TO}_i, \\ \text{TO} &= \begin{cases} 0, & \text{if SIL or BC} \\ 1, & \text{otherwise} \end{cases} \end{aligned} \tag{7}$$

For successful takeovers, we measure latency from the end of user speech to the start of the model response.

2) Pause Handling. Humans naturally pause and hesitate between consecutive turns or within the same sentence. User pauses much longer before or between utterances to think or organize their thoughts. Therefore, it is critical that the model does not take over the turn during user pauses. Instead, it should decide whether to speak or yield based on semantic cues.

Metric: We also use the TOR metric to measure whether the model takes over the turn within a time frame of the silence duration plus the model's average response latency.

3) User Interrupt. This scenario evaluates a crucial aspect of full-duplex interaction: the model's ability to terminate its turn and attend to new user input.

Metric: Turn-Termination Rate (TTR) measures the proportion of cases in which the model terminates its ongoing turn in response to the user's interruption; higher values indicate more reliable yielding. Latency quantifies the average time to termination. Topic Shift Score (TSS), rated by GPT-4o, assesses whether the model's second-round response successfully shifts from the initial query to the user's interrupting query.

4) Model Interrupt. While often avoided in system design (Arora et al., 2025), a model's ability to proactively interrupt a user can be beneficial when immediate intervention is warranted. In `FLEXI`, we instantiate this ability with time-sensitive emergency cases, where effective interruption requires both speech understanding and latency control.

Metric: We use TOR to assess interruptions and an Emergency Detection Score (EDS) rated by GPT-4o to determine if the model correctly identifies the emergency situation.

5) User Backchannel. This scenario tests whether the model can ignore user backchannels (e.g., 'yeah') without misunderstanding them as turn-taking attempts.

Metric: We use TTR to measure unintended turn termination in response to user backchannels; lower values indicate better preservation of the model's ongoing turn. An interruption is acceptable only if the model resumes its original utterance within 1 second, measured by a coherence metric.

6) Model Backchannel. Although less emphasized in HCI, appropriate model-generated backchannels can signal active listening and enhance naturalness. This scenario evaluates the model's ability to produce socially appropriate backchannels.

Metric: We use TOR and Backchannel Rate (BCR) to quantify model behavior. To assess appropriateness, we use GPT-4o to annotate reference backchannel positions and compute the Jensen-Shannon Divergence (JSD) between the model's predicted backchannel distribution (P) and the reference distribution (Q). JSD scores range from 0 (perfect alignment) to 1 (complete divergence).

$$\begin{aligned} \text{JSD}(P\|Q) = &\frac{1}{2}\sum_i P(i)\log\frac{P(i)}{M(i)} \\ &+ \frac{1}{2}\sum_i Q(i)\log\frac{Q(i)}{M(i)} \end{aligned} \tag{8}$$

Table 1: Evaluation of turn-taking and pause handling using takeover rate (TOR), jump-in rate (JIR), and latency (Lat.). Results are reported as English/Chinese; a dash denotes an unsupported language.

Model	Turn-taking			Pause
	TOR(↑)	Lat.	JIR(↓)	TOR(↓)
Moshi	0.995/-	**0.696**/-	0.785/-	0.550/-
Freeze-Omni	**1.000**/0.980	1.147/**0.959**	0.120/0.055	0.937/0.830
VITA-1.5	**1.000**/0.995	4.878/5.679	**0.000/0.000**	0.916/0.740
DuplexCascade	0.970/-	2.091/-	**0.000**/-	**0.073**/-
Gemini	0.995/**1.000**	2.680/4.749	**0.000**/0.005	0.585/0.450
GPT	**1.000/1.000**	2.152/1.786	**0.000**/0.010	0.250/**0.125**

where $M(i) = \frac{1}{2}(P(i) + Q(i))$, and i indexes the time windows.

4 Experiments and Analysis

4.1 Experimental Setups

Data Construction. Owing to the scarcity of real full-duplex conversational data, we follow Lin et al. (Lin et al., 2025b) and evaluate S2S full-duplex LLMs using synthetic speech. The construction process creates controlled test cases in which critical interaction events, such as pauses, backchannels, user interruptions, and emergency cues, are explicitly localized in time, allowing systems to be compared under the same speech input and event position.

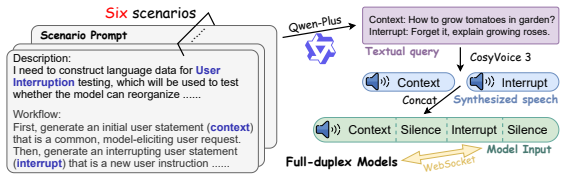


Figure 2: Pipeline of data construction and evaluation.

As shown in Fig. 2, we decompose each scenario into semantic and temporal slots, such as the initial user context, interrupting utterance, backchannel token, or post-pause continuation. Candidate texts are generated with Qwen-plus[1] under scenario-specific instructions, covering daily assistance, knowledge-seeking, planning, extended narratives, false-premise cases, and safety-sensitive situations. We then synthesize each text segment with CosyVoice3 (Du et al., 2025) and concatenate the resulting audio according to the target scenario, preserving event positions as temporal annotations and inserting appropriate silences after user backchannels, interruptions, and query completions.

Table 2: Evaluation of user and model interruption using turn-termination rate (TTR), takeover rate (TOR), latency (Lat.), topic shift score (TSS), and emergency detection score (EDS). TSS and EDS are assessed by GPT-4o on a 0–5 scale. Results are reported as English/Chinese; a dash denotes an unsupported language.

User Interrupt			
Model	TTR(↑)	Lat.	TSS(↑)
Moshi	0.462/-	2.660/-	2.48/-
Freeze-Omni	0.539/0.525	1.751/1.345	2.90/2.78
VITA-1.5	0.226/0.225	4.016/3.483	0.75/0.93
DuplexCascade	0.110/-	4.055/-	2.14/-
Gemini	0.225/0.475	**1.289/0.652**	2.29/3.80
GPT	**0.845/0.985**	2.405/1.603	**4.57/4.06**

Model Interrupt		
Model	TOR(↑)	EDS(↑)
Moshi	0.390/-	0.81/-
Freeze-Omni	**0.535/0.105**	1.84/2.82
VITA-1.5	0.000/0.030	1.72/2.13
DuplexCascade	0.035/-	2.11/-
Gemini	0.040/0.010	4.29/**4.43**
GPT	0.015/0.005	**4.59**/3.88

Finally, we manually validate the samples for semantic coherence, event clarity, cue position, and speech intelligibility, removing or regenerating invalid cases. The released evaluation set contains 3,600 samples across six languages: 1,200 each in Chinese and English (200 per scenario per language), and 300 each in German, Spanish, French, and Italian (50 per scenario per language). Code and dataset are available in the public repository[2].

Models. We evaluate Moshi (Défossez et al., 2024), Freeze-Omni (Wang et al., 2025b), VITA-1.5 (Fu et al., 2025), DuplexCascade (Yang et al., 2026), Gemini-3.1-flash-live-preview (Gemini), and GPT-Realtime-2.1-Mini (GPT). Gemini uses the Live API[3]; GPT-Realtime uses Knox Chat's OpenAI Realtime-compatible streaming interface. Tables 1–3 report English/Chinese results.

Metric aggregation. Cases that do not meet a metric's applicability conditions, owing to the interaction state or the model's behavior, are excluded from its denominator. Effective sample sizes therefore vary by model and metric.

[1] https://www.alibabacloud.com/help/en/model-studio/what-is-qwen-llm

[2] https://github.com/ChristineCHEN274/FLEXI

[3] https://ai.google.dev/gemini-api/docs/live

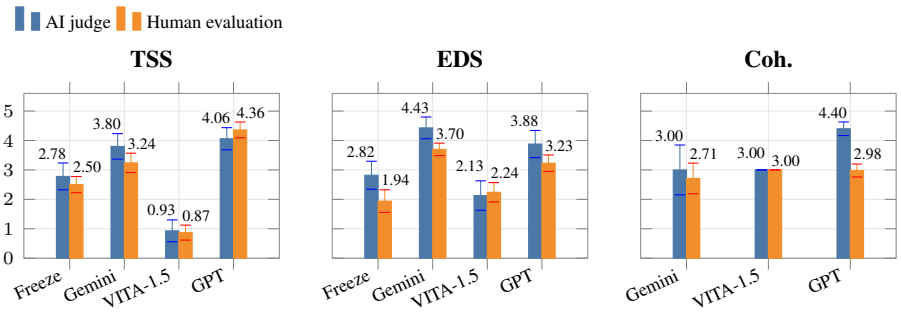


Figure 3: AI-judged and human ratings on paired Chinese samples. Bars show means and error bars show normal-approximation 95% confidence intervals. Valid paired-sample counts for TSS are 50/50/45/50, for EDS 39/50/50/50, and for Coh. 0/14/1/50, ordered as Freeze-Omni, Gemini, VITA-1.5, and GPT-Realtime-2.1-Mini.

4.2 Experimental Results

Turn-Taking and Pause Handling. Table 1 shows a responsiveness–restraint trade-off. Moshi responds fastest but frequently starts before the user finishes. DuplexCascade avoids premature responses and handles pauses conservatively, at higher latency. GPT combines reliable completed-turn takeover with low premature takeover in both languages. Thus, latency must be interpreted alongside JIR and pause TOR to account for premature floor-taking.

User and Model Interrupt. GPT-Realtime yields most reliably to user interruptions, whereas DuplexCascade's conservative takeover policy does not ensure prompt yielding (Table 2). For model interruption (Table 2), high TOR in open source systems often accompanies low EDS, indicating premature speech without emergency recognition. Gemini and GPT-Realtime recognize emergencies more effectively but rarely interrupt. Useful intervention therefore requires both semantic recognition and timely floor-taking.

User and Model Backchannel. Most systems rarely yield to user backchannels, while GPT-Realtime yields frequently but usually resumes coherently (Table 3). Model-generated backchannels remain weak: systems often stay silent or produce full responses at local boundaries. GPT-Realtime takes every such opportunity as a full turn. BCR must therefore be interpreted alongside TOR and reference alignment to assess brief, non-disruptive feedback.

Four-language extension. Table 4 reports a separate CosyVoice3 evaluation set covering German, Spanish, French, and Italian, with 50 samples per scenario and 300 per language. We evaluate `gpt-realtime-2.1-mini` and `Gemini-3.1-flash-live-preview` using the frozen temporal protocol, local Whisper large-v3-turbo transcriptions, and GPT-4o semantic judgments. Compared with Gemini, GPT yields more reliably to user interruptions but more frequently to user backchannels. Gemini achieves higher EDS across these languages, although both systems rarely initiate an interruption. EDS evaluates nonempty responses and therefore does not by itself demonstrate timely intervention.

Human validation. To assess automatic semantic judging, we manually rated fixed subsets of Chinese outputs using the same 0–5 rubrics. Figure 3

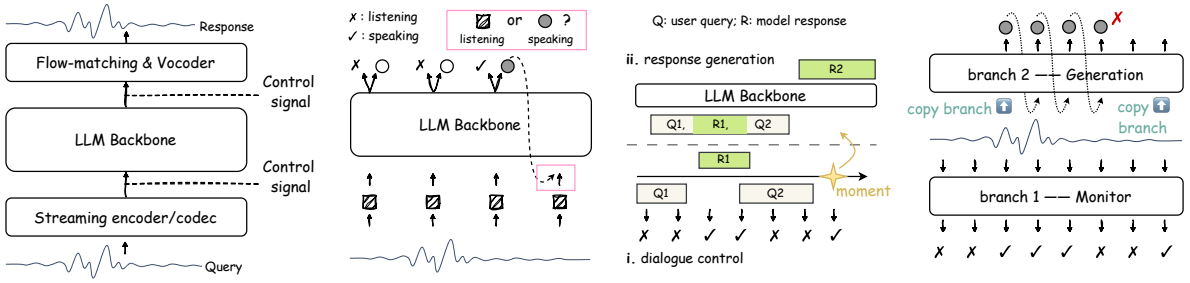
Table 3: Evaluation of user and model backchannels using turn-termination rate (TTR), takeover rate (TOR), coherence (Coh.), backchannel rate (BCR), and Jensen-Shannon divergence (JSD). Coh. reports passed/judged counts under GPT-4o evaluation; 0/0 indicates no eligible cases. JSD is computed against GPT-4o-annotated reference backchannels. Results are reported as English/Chinese; a dash denotes an unsupported language. VITA-1.5's Chinese BCR is approximately 3.15×10^{-4}, rounded to 0.000.

User Backchannel		
Model	TTR(↓)	Coh.
Moshi	0.020/-	(1/4)/-
Freeze-Omni	0.086/**0.000**	(8/17)/(0/0)
VITA-1.5	**0.000**/0.005	(0/0)/(**1/1**)
DuplexCascade	0.010/-	(0/2)/-
Gemini	0.025/0.070	(4/5)/(9/14)
GPT	0.485/0.730	(**94/97**)/(138/146)

Model Backchannel			
Model	TOR(↓)	BCR(↑)	JSD(↓)
Moshi	0.578/-	**0.031**/-	**0.947**/-
Freeze-Omni	0.608/**0.131**	0.003/0.000	0.986/1.000
VITA-1.5	0.578/0.950	0.002/0.000	0.986/0.999
DuplexCascade	**0.570**/-	0.001/-	0.987/-
Gemini	0.765/0.960	0.009/**0.004**	0.989/**0.993**
GPT	1.000/1.000	0.000/0.000	1.000/1.000

Table 4: Results on the four-language set (de/es/fr/it: German/Spanish/French/Italian). Each scenario contains 50 valid samples per model. BC denotes backchannel; Coh. reports passed/judged counts (0/0: no eligible cases). Metric definitions follow Tables 1–3.

Model	Turn-taking			Pause	User Interrupt			Model Interrupt		User BC		Model BC		
	TOR(↑)	Lat.	JIR(↓)	TOR(↓)	TTR(↑)	Lat.	TSS(↑)	TOR(↑)	EDS(↑)	TTR(↓)	Coh.	TOR(↓)	BCR(↑)	JSD(↓)
GPT_{de}	**1.000**	**2.243**	**0.000**	**0.020**	**1.000**	1.967	**3.94**	**0.020**	3.90	0.620	**19/31**	1.000	**0.000**	**1.000**
$\text{Gemini}_{\text{de}}$	**1.000**	2.668	**0.000**	0.100	0.380	**1.738**	3.63	**0.020**	**4.56**	**0.000**	0/0	**0.900**	**0.000**	**1.000**
GPT_{es}	**1.000**	**2.214**	**0.000**	**0.000**	**0.960**	2.044	**4.27**	**0.000**	4.17	0.480	**14/24**	1.000	**0.000**	**1.000**
$\text{Gemini}_{\text{es}}$	**1.000**	2.477	**0.000**	0.280	0.340	**1.793**	3.41	**0.000**	**4.49**	**0.060**	1/3	**0.740**	**0.000**	**1.000**
GPT_{fr}	0.980	**2.386**	**0.000**	**0.000**	**0.940**	2.183	**4.21**	**0.000**	4.08	0.480	12/24	1.000	0.000	1.000
$\text{Gemini}_{\text{fr}}$	**1.000**	2.395	**0.000**	0.300	0.400	**1.461**	3.90	**0.000**	**4.32**	**0.120**	**4/6**	**0.820**	**0.004**	**0.995**
GPT_{it}	**1.000**	**2.199**	**0.000**	**0.000**	**0.960**	2.180	4.27	**0.000**	3.93	0.600	**18/30**	1.000	0.000	1.000
$\text{Gemini}_{\text{it}}$	**1.000**	3.022	**0.000**	0.580	0.360	**1.316**	**4.33**	**0.000**	**4.49**	**0.020**	0/1	**0.800**	**0.008**	**0.996**



a. Positions of control signal b. Failure of next token prediction c. Dialogue manager framework d. Dual-branch framework

Figure 4: Full-duplex modeling challenges and control designs: (a) control-signal positions, (b) limitations of standard next-token prediction, (c) a dialogue-manager framework, and (d) a dual-branch framework.

compares mean AI-judged and human scores on the same valid samples. Human ratings are lower for most model–metric pairs, with the largest observed gap for GPT-Realtime on Coh.; GPT-Realtime TSS and VITA-1.5 EDS are the main exceptions. Under human evaluation, GPT-Realtime remains strongest on TSS, whereas coherence comparisons for Gemini and VITA-1.5 are less stable because few outputs are eligible.

Design implications. Figure 4 summarizes control designs for coordinating streaming comprehension, floor management, and low-latency generation. Models must distinguish a completed turn from a hesitation, an interruption from a backchannel, and an emergency from an ordinary continuation. Concurrent input–output modeling, including next token-pair prediction (Défossez et al., 2024; Wang et al., 2025a), is one promising approach to maintaining awareness while speaking. However, differences in training, control policies, and deployment conditions prevent attributing performance differences solely to architecture. Future models should learn when and how strongly to intervene, rather than optimize response speed or speaking frequency alone.

5 Conclusion

We introduced FLEXI, a six-scenario benchmark for full-duplex spoken interaction. Evaluation across open source and commercial systems reveals persistent gaps in timely yielding, emergency intervention, and lightweight backchanneling. Semantic competence does not guarantee appropriate floor management. These findings motivate joint evaluation and modeling of response content, timing, and intervention strength.

Limitations

FLEXI relies on synthetically constructed speech. Although controlled construction ensures comparable event timing across systems, it cannot capture the full variability of natural conversations. The evaluated model set is also limited by the availability of open implementations and commercial real-time APIs. Finally, semantic scoring relies on GPT-4o and may therefore reflect judge-specific bias. Despite human validation on fixed Chinese subsets, further evaluation at a larger scale, with natural dialogue and more diverse full-duplex systems, remains necessary.

Acknowledgments

This work was supported in part by the National Science Foundation of China (Nos. 62276056 and U24A20334), the Yunnan Fundamental Research Projects (No.202401BC070021), the Yunnan Science and Technology Major Project (No. 202502AD080014), the Liaoning Provincial Science and Technology Plan Project (No. 2026JH40/10100033), the Fundamental Research Funds for the Central Universities (Nos. N26XQD028, N25BSS054 and N25BSS094), and the Program of Introducing Talents of Discipline to Universities, Plan 111 (No.B16009).

References

Siddhant Arora, Zhiyun Lu, Chung-Cheng Chiu, Ruoming Pang, and Shinji Watanabe. 2025. Talking Turns: Benchmarking audio foundation models on turn-taking dynamics. In *The Thirteenth International Conference on Learning Representations*.

Qian Chen, Yafeng Chen, Yanni Chen, Mengzhe Chen, Yingda Chen, Chong Deng, Zhihao Du, Ruize Gao, Changfeng Gao, Zhifu Gao, et al. 2025. MinMo: A multimodal large language model for seamless voice interaction. *arXiv preprint arXiv:2501.06282*.

Herbert H Clark. 1996. *Using language*. Cambridge university press.

Alexandre Défossez, Laurent Mazaré, Manu Orsini, Amélie Royer, Patrick Pérez, Hervé Jégou, Edouard Grave, and Neil Zeghidour. 2024. Moshi: a speech-text foundation model for real-time dialogue. Technical report.

Zhihao Du, Changfeng Gao, Yuxuan Wang, Fan Yu, Tianyu Zhao, Hao Wang, Xiang Lv, Hui Wang, Chongjia Ni, Xian Shi, Keyu An, Guanrou Yang, Yabin Li, Yanni Chen, Zhifu Gao, Qian Chen, Yue Gu, Mengzhe Chen, Yafeng Chen, and 3 others. 2025. CosyVoice 3: Towards in-the-wild speech generation via scaling-up and post-training. *arXiv preprint arXiv:2505.17589*.

Chaoyou Fu, Haojia Lin, Xiong Wang, Yi-Fan Zhang, Yunhang Shen, Xiaoyu Liu, Yangze Li, Zuwei Long, Heting Gao, Ke Li, et al. 2025. VITA-1.5: Towards GPT-4o level real-time vision and speech interaction. *arXiv preprint arXiv:2501.01957*.

Yuan Ge, Junxiang Zhang, Xiaoqian Liu, Bei Li, Xiangnan Ma, Chenglong Wang, Kaiyang Ye, Yangfan Du, Linfeng Zhang, Yuxin Huang, et al. 2025. SageLM: A multi-aspect and explainable large language model for speech judgement. *arXiv preprint arXiv:2508.20916*.

Ke Hu, Ehsan Hosseini-Asl, Chen Chen, Edresson Casanova, Subhankar Ghosh, Piotr Żelasko, Zhehuai Chen, Jason Li, Jagadeesh Balam, and Boris Ginsburg. 2025. Efficient and direct duplex modeling for speech-to-speech language model. *arXiv preprint arXiv:2505.15670*.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. 2024. GPT-4o system card. *arXiv preprint arXiv:2410.21276*.

Hua Lai, Silei Han, Ying Li, Zhengtao Yu, and Zirui Guo. 2025. Evaluating the translation performance of multilingual large language models: A case study on Southeast Asian languages. In *Machine Translation: 20th China Conference, CCMT 2024*, volume 2365 of *Communications in Computer and Information Science*, pages 27–47, Singapore. Springer.

Borui Liao, Yulong Xu, Jiao Ou, Kaiyuan Yang, Weihua Jian, Pengfei Wan, and Di Zhang. 2025. FlexDuo: A pluggable system for enabling full-duplex capabilities in speech dialogue systems. *arXiv preprint arXiv:2502.13472*.

Guan-Ting Lin, Shih-Yun Shan Kuan, Jiatong Shi, Kai-Wei Chang, Siddhant Arora, Shinji Watanabe, and Hung-yi Lee. 2026. Full-Duplex-Bench-v2: A multi-turn evaluation framework for duplex dialogue systems with an automated examiner. *arXiv preprint arXiv:2510.07838*.

Guan-Ting Lin, Shih-Yun Shan Kuan, Qirui Wang, Jiachen Lian, Tingle Li, and Hung-yi Lee. 2025a. Full-Duplex-Bench v1.5: Evaluating overlap handling for full-duplex speech models. *arXiv preprint arXiv:2507.23159*.

Guan-Ting Lin, Jiachen Lian, Tingle Li, Qirui Wang, Gopala Anumanchipalli, Alexander H Liu, and Hung-yi Lee. 2025b. Full-Duplex-Bench: A benchmark to evaluate full-duplex spoken dialogue models on turn-taking capabilities. *arXiv preprint arXiv:2503.04721*.

Xiangyu Lu, Wang Xu, Haoyu Wang, Hongyun Zhou, Haiyan Zhao, Conghui Zhu, Tiejun Zhao, and Muyun Yang. 2025a. DuplexMamba: Enhancing real-time speech conversations with duplex and streaming capabilities. *arXiv preprint arXiv:2502.11123*.

Yaxi Lu, Shenzhi Yang, Cheng Qian, Guirong Chen, Qinyu Luo, Yesai Wu, Huadong Wang, Xin Cong, Zhong Zhang, Yankai Lin, et al. 2025b. Proactive Agent: Shifting LLM agents from reactive responses to active assistance. In *Proceedings of the 42nd International Conference on Machine Learning*.

Ewa Luger and Abigail Sellen. 2016. 'like having a really bad PA' the gulf between user expectation and experience of conversational agents. In *Proceedings of the 2016 CHI conference on human factors in computing systems*, pages 5286–5297.

Tu Anh Nguyen, Eugene Kharitonov, Jade Copet, Yossi Adi, Wei-Ning Hsu, Ali Elkahky, Paden Tomasello,

Robin Algayres, Benoît Sagot, Abdelrahman Mohamed, and Emmanuel Dupoux. 2023. Generative spoken dialogue language modeling. *Transactions of the Association for Computational Linguistics*, 11:250–266.

Jiale Ou, Hongfei Xu, and Hongying Zan. 2025. Joint multi-modal modeling for speech-to-text translation as multilingual neural machine translation. In *Machine Translation: 20th China Conference, CCMT 2024*, volume 2365 of *Communications in Computer and Information Science*, pages 99–115, Singapore. Springer.

Yizhou Peng, Yi-Wen Chao, Dianwen Ng, Yukun Ma, Chongjia Ni, Bin Ma, and Eng Siong Chng. 2025. FD-Bench: A full-duplex benchmarking pipeline designed for full duplex spoken dialogue systems. *arXiv preprint arXiv:2507.19040*.

Harvey Sacks, Emanuel A. Schegloff, and Gail Jefferson. 1974. A simplest systematics for the organization of turn-taking for conversation. *Language*, 50(4):696–735.

Gabriel Skantze. 2021. Turn-taking in conversational systems and human-robot interaction: A review. *Computer Speech & Language*, 67:101178.

Chen Wang, Junhong Wu, Shuo Ren, Jun Lin, Guangfu Wang, and Jiajun Zhang. 2026. ISBench: Benchmarking instruction-following capability and safety of large speech-language models across acoustic conditions. In *Machine Translation: 21st China Conference, CCMT 2025*, volume 2906 of *Communications in Computer and Information Science*, pages 49–57, Singapore. Springer.

Jinhan Wang, Long Chen, Aparna Khare, Anirudh Raju, Pranav Dheram, Di He, Minhua Wu, Andreas Stolcke, and Venkatesh Ravichandran. 2024. Turn-taking and backchannel prediction with acoustic and large language model fusion. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 12121–12125. IEEE.

Qichao Wang, Ziqiao Meng, Wenqian Cui, Yifei Zhang, Pengcheng Wu, Bingzhe Wu, Irwin King, Liang Chen, and Peilin Zhao. 2025a. NTPP: Generative speech language modeling for dual-channel spoken dialogue via next-token-pair prediction. In *Forty-second International Conference on Machine Learning*.

Xiong Wang, Yangze Li, Chaoyou Fu, Yike Zhang, Yunhang Shen, Lei Xie, Ke Li, Xing Sun, and Long Ma. 2025b. Freeze-Omni: A smart and low latency speech-to-speech dialogue model with frozen LLM. *ICML*.

Nigel Ward and Wataru Tsukahara. 2000. Prosodic features which cue back-channel responses in English and Japanese. *Journal of Pragmatics*, 32(8):1177–1207.

Bufang Yang, Lilin Xu, Liekang Zeng, Kaiwei Liu, Siyang Jiang, Wenrui Lu, Hongkai Chen, Xiaofan Jiang, Guoliang Xing, and Zhenyu Yan. 2025. ContextAgent: Context-aware proactive LLM agents with open-world sensory perceptions. In *Proceedings of the 39th Annual Conference on Neural Information Processing Systems*.

Jianing Yang, Yusuke Fujita, and Yui Sudo. 2026. DuplexCascade: Full-duplex speech-to-speech dialogue with VAD-free cascaded ASR-LLM-TTS pipeline and micro-turn optimization. *arXiv preprint arXiv:2603.09180*.

Wenyi Yu, Siyin Wang, Xiaoyu Yang, Xianzhao Chen, Xiaohai Tian, Jun Zhang, Guangzhi Sun, Lu Lu, Yuxuan Wang, and Chao Zhang. 2025. SALMONN-Omni: A standalone speech LLM without codec injection for full-duplex conversation. *arXiv preprint arXiv:2505.17060*.

Daoan Zhang, Pai Liu, Xiaofei Zhou, Yuan Ge, Guangchen Lan, Jing Bi, Christopher Brinton, Ehsan Hoque, and Jiebo Luo. 2025a. VisualActBench: Can VLMs see and act like a human? *arXiv preprint arXiv:2512.09907*.

Hao Zhang, Weiwei Li, Rilin Chen, Vinay Kothapally, Meng Yu, and Dong Yu. 2025b. LLM-enhanced dialogue management for full-duplex spoken dialogue systems. *arXiv preprint arXiv:2502.14145*.

Qinglin Zhang, Luyao Cheng, Chong Deng, Qian Chen, Wen Wang, Siqi Zheng, Jiaqing Liu, Hai Yu, Chao-Hong Tan, Zhihao Du, and ShiLiang Zhang. 2025c. OmniFlatten: An end-to-end GPT model for seamless voice conversation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14570–14580, Vienna, Austria. Association for Computational Linguistics.

Yuheng Zhao, Xueli Shu, Liwen Fan, Lin Gao, Yu Zhang, and Siming Chen. 2026. ProactiveVA: Proactive visual analytics with LLM-based UI agent. *IEEE Transactions on Visualization and Computer Graphics*, 32(1):451–461.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, et al. 2023. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. In *Proceedings of the 37th Conference on Neural Information Processing Systems, Datasets and Benchmarks Track*.

Hongli Zhou, Hui Huang, Yu Long, et al. 2024. Mitigating the bias of large language model evaluation. In *Proceedings of the 23rd Chinese National Conference on Computational Linguistics (Volume 1: Main Conference)*, pages 1310–1319.