

Still Between Us? Evaluating and Improving Voice Assistant Robustness to Third-Party Interruptions

Dongwook Lee¹, Eunwoo Song², Che Hyun Lee³, Heeseung Kim^{4†}, Sungroh Yoon^{1,3,5†}

¹Interdisciplinary Program in Artificial Intelligence, Seoul National University (SNU)

²Department of EE, Yonsei University ³Department of ECE, SNU

⁴Department of AI, University of Seoul ⁵AIIS, ASRI, INMC, and ISRC, SNU

Abstract

While recent Spoken Language Models (SLMs) have been actively deployed in real-world scenarios, they lack the capability to discern Third-Party Interruptions (TPI) from the primary user’s ongoing flow, leaving them vulnerable to contextual failures. To bridge this gap, we introduce **TPI-Train**, a dataset of 88K instances designed with *speaker-aware hard negatives* to enforce acoustic cue prioritization for interruption handling, and **TPI-Bench**, a comprehensive evaluation framework designed to rigorously measure the interruption-handling strategy and precise speaker discrimination in deceptive contexts. Experiments demonstrate that our dataset design mitigates *semantic shortcut learning*—a critical pitfall where models exploit semantic context while neglecting acoustic signals essential for discerning speaker changes. We believe our work establishes a foundational resource for overcoming text-dominated unimodal reliance in SLMs, paving the way for more robust multi-party spoken interaction. The code for the framework is publicly available at <https://tpi-va.github.io/>.

1 Introduction

Recent Spoken Language Models (SLMs) (Wu et al., 2025; Stacey et al., 2024a; Kim et al., 2024; Xu et al., 2025a) have significantly advanced the capabilities of voice assistants (VAs), enabling increasingly natural and human-like conversations. These models excel in dyadic interactions, adeptly handling complex queries and maintaining conversational flow with a single speaker. However, this proficiency is largely confined to isolated dyadic settings, as current models struggle to remain robust against speakers other than the primary user (Wang et al., 2025a; Kwon et al., 2025). A practically significant and plausible failure case

arises in scenarios involving third-party interruptions (TPI), where VAs often misinterpret a multi-speaker dialogue as a single, continuous utterance from the primary user. For instance, if a speaker asks, “Should we order the new pasta?” and a third-party interjects with, “No, let’s just get the usual,” ordinary VAs might process the entire sequence as a *self-repair* (Levelt, 1983) utterance from the initial speaker—a common phenomenon in voice assistant interactions (Goel et al., 2023; Stacey et al., 2024b; Liu et al., 2024). This erroneous concatenation leads to nonsensical or inappropriate responses as described in Figure 1 and Appendix A, degrading the user experience, eroding trust, and ultimately discouraging further engagement with the voice assistant (Baughan et al., 2023).

We hypothesize that this shortcoming does not merely indicate deficiencies in dialogue-level reasoning but stem from a more fundamental limitation: a lack of sensitivity to acoustic cues for distinguishing speakers. To bridge this gap, we argue that effective handling of TPI requires two essential abilities: (1) *Discerning Speaker Interruption*—the ability to robustly detect interruptions and (2) *Situation-Discriminative Response*—the ability to flexibly adapt response strategies (e.g., answering, disregarding, etc.) to align with user-defined preferences.

However, when and how to react to interruptions is non-trivial, as there is no universally optimal answer strategy for handling interruptions (Xie et al., 2022; Cao et al., 2025). Instead, the appropriate strategy is inherently subjective, contingent upon user preferences (Tabassum et al., 2020), which means it cannot be solved by a rigid solution but instead requires a flexible customized framework. To this end, we propose a comprehensive framework that unifies the entire workflow, spanning from response strategy specification to dataset construction and evaluation. To demonstrate its practical utility, we also instantiate a reference response strategy

[†]Corresponding authors.

(Email: sryoon@snu.ac.kr; gmltmd789@uos.ac.kr)

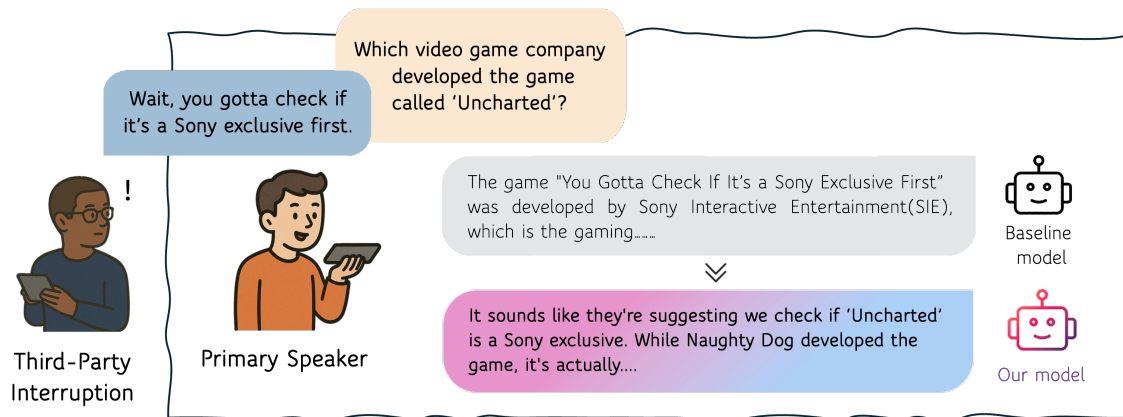


Figure 1: **Example of a TPI query sampled from our TPI-Train.** Recent spoken language models mistake the third-party interruption for a continuous utterance from the primary speaker, while our model correctly identifies the interruption and responds in a TPI-aware manner. Detailed failure cases are described in Appendix A.

within our framework and demonstrate that models trained under this strategy produce effective and natural responses under TPI situations through human evaluations.

Within our framework, we introduce **TPI-Train**, a dataset constructed by reinterpreting seven foundational interruption categories from canonical dyadic taxonomies (Murata, 1994; Goldberg, 1990) in triadic contexts to achieve theoretical validity and realism, expanding it into 26 diverse TPI scenarios. In addition, we propose **TPI-Bench** containing two main benchmarks: TPI-Test, which evaluates models’ ability to produce “situation-discriminative responses” under genuine interruptions, and Janus-Test, which tests whether a model can discern an interrupter from acoustically single-speaker audio yet semantically resemble interruptions as illustrated in Figure 3. Finally, to validate real-world transferability, we curate a compact yet high-quality real-world test set of interruption samples. Along with these benchmarks, we also propose two LLM-based metrics—Response Strategy Following (RSF) and Overall Helpfulness (OH)—to enable interpretable evaluation.

Our experiments reveal that current SLMs struggle to handle third-party interruptions, often misattributing two-speaker utterances as a single voice. While fine-tuning on interruption data effectively imparts interruption-handling capabilities, the model remains prone to exploiting semantic shortcuts—relying on lexical interruption cues rather than acoustic shifts—a similar phenomenon observed in vision-language multimodal learning, where multimodal reasoning is often bypassed in favor of unimodal cues (Koishigarina et al., 2025; Yuksekgonul et al., 2023). To enforce

reliance on acoustic evidence over textual cues, we construct TPI-Train by integrating speaker-aware hard negatives into the training dataset. Along with TPI-Train for interruption-specific supervision and dyadic interaction for core conversational competence, we can produce a balanced and robust TPI-aware model without sacrificing general abilities. Its ability to discriminate changes in speaker is supported by well-separated embedding representations and our reference answer strategies are validated through human evaluations, which confirm that our strategies and the trained model generate responses aligned with user preferences. Together, these results underscore the practical value of our framework for building TPI-aware voice assistants.

Our contributions are summarized as follows:

- We define TPI-awareness and propose the first comprehensive framework, offering an end-to-end pipeline that spans from defining subjective response strategies to dataset construction and evaluation, thereby extending SLMs capabilities beyond dyadic interactions.
- We construct TPI-Train across 26 diverse scenarios, integrating *speaker-aware hard negatives*, mitigating semantic shortcut learning and promoting acoustically grounded speaker discrimination.
- We introduce TPI-Bench, which consists of TPI-Test, Janus-Test and real-world test set, for evaluating situation-discriminative responses and acoustic sensitivity. In addition, through human evaluation, we demonstrate that our example strategy within the framework is well aligned with human preferences.

2 Task Definition

2.1 Problem Setting

We investigate a scenario where the main interaction between a *primary speaker* and a model is interrupted by a third party. This scenario requires the model to not only understand the primary query, but also to recognize and handle interruptions in an interruption-aware manner if required. We formalize this setting as follows. Let a *primary speaker* utterance be denoted by U_p and a *third-party speaker* utterance by U_{tp} . An interruption event is represented as the ordered pair $U_{p \rightarrow tp} = (U_p, U_{tp})$, where the third-party utterance (U_{tp}) intrudes upon the primary one. Given an interruption event $U_{p \rightarrow tp}$, the model is required to generate a response sequence $\mathcal{Y} = \{y_1, y_2, \dots, y_K\}$, where each $y_k \in \mathcal{V}$ and \mathcal{V} denotes the set of textual tokens in the model’s vocabulary. The response generation process is modeled as a conditional distribution: $P_\theta(\mathcal{Y} | U_p, U_{tp}) = \prod_{k=1}^K P_\theta(y_k | y_{<k}, U_p, U_{tp})$, which reduces to $P_\theta(\mathcal{Y} | U_p) = \prod_{k=1}^K P_\theta(y_k | y_{<k}, U_p)$ in the absence of interruption. While we focus on a speech-input, text-output setting in this work, the formulation is modality-agnostic and can be readily extended to a speech-to-speech setting.

2.2 A Framework for Response Strategy

The formulation of a universally appropriate response to interruptions is inherently not trivial, as the ideal behavior often varies based on the user’s preferences and conversational situations (Tabasum et al., 2020; Cao et al., 2025). Therefore, rather than prescribing a single response strategy, we propose a flexible framework that enables users to customize the ideal responses of a voice assistant based on their principles. To this end, we present one possible response strategy as a reference, which we reflect in the answer portion of our dataset, and later demonstrate its effectiveness through comprehensive LLM and human evaluations in Section 5.1. We release the pipeline of our framework, enabling users to build custom responses based on their own response strategies.

Our framework follows a two-stage process. We begin by recognizing that not every interruption warrants a response from the voice assistant—for example, simple acknowledgments or unrelated remarks may not contribute meaningfully to the ongoing interaction. Accordingly, we first classify each interruption event $U_{p \rightarrow tp}$ into one of two high-level categories: **Actionable** (C_A), when the model

should respond, or **Ignorable** (C_I), when it should not. Second, under our reference strategy, we define the criteria for C_A as interruptions conveying information relevant to the primary speaker’s intent, with all remaining cases categorized as C_I . Third, based on this classification, we determine how the model should respond: π_A and π_I . In this paper, we provide the class criteria and the policies as references, which can be freely customized by the user. The specific criteria for C_A and C_I , along with the details of our π_A and π_I , are further elaborated in Section 3.1 and Appendix D.

2.3 Third-Party Interruption Awareness

We define *third-party interruption awareness* through two key capabilities:

1) **Discerning Speaker Interruptions.** The model should distinguish whether an interruption has occurred, as this discrimination enables the appropriate application of the strategies and, in turn, facilitates the generation of desirable responses to a primary speaker. This capacity requires the model to detect speaker changes using acoustic cues, thereby handling speaker interruption robustly while maintaining previous performance in single-speaker conditions.

2) **Situation-Discriminative Response.** As the desirable responses vary according to user preferences, the model should generate a response that aligns with the predefined answer strategy.

3 Dataset

3.1 TPI-Train

Corpus Construction. We first construct a comprehensive corpus, covering 26 scenarios grounded in seven taxonomies of conversational interruptions from prior work (Murata, 1994; Goldberg, 1990). We extend them to the triadic setting involving a primary speaker, a third-party interrupter, and a conversational model. This adaptation yields 26 distinct scenarios, which include cases such as critical corrections, helpful clarifications, and conflicts (Appendix B). We extract primary speaker utterances U_p from VoiceAssistant-400k (Xie and Wu, 2024), a large-scale speech dataset that primarily consists of various single-turn interactions. For each utterance, we randomly select one of the 26 scenarios and generate a corresponding third-party interruption. We provide the LLM with the scenario description and the transcript of the primary utterance, prompting it to produce a context-

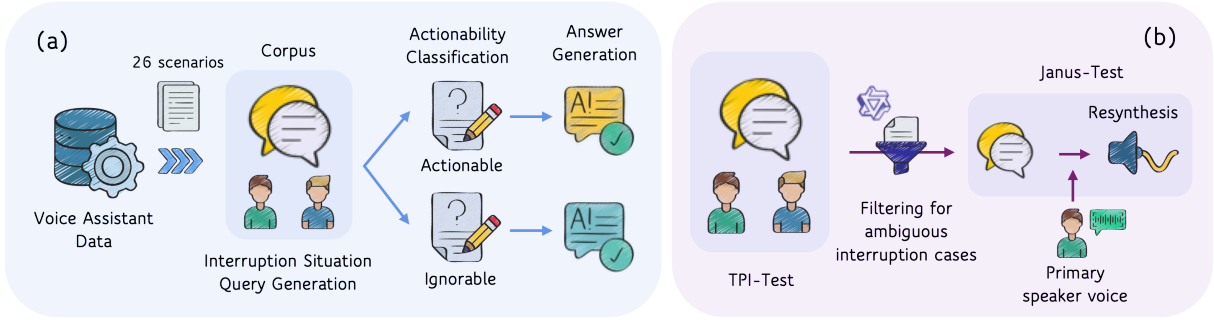


Figure 2: **Overview of our corpus and TPI-Bench construction pipeline.** (a) Our corpus is generated from voice assistant data to adapt various interruption scenarios. For the train dataset, queries are classified as *Actionable* or *Ignorable* with answers generated according to the predefined response strategy. (b) From the sampled TPI-Bench data, we identify instances that are ambiguous as to whether they involve one or two speakers when relying solely on the text. These samples are then re-synthesized using only the primary speaker’s voice to construct the Janus-Test.

appropriate interruption. Here, we consider two major types of interruption events based on the timing of the intrusion: (1) **within-sentence interruption**, where U_{tp} intrudes while U_p is still ongoing, cutting into the primary speaker’s utterance before it reaches completion; and (2) **after-completion interruption**, where U_{tp} occurs after U_p has formed a complete linguistic unit, commonly a full sentence. Those are synthesized using a text-to-speech (TTS) model (Resemble AI, 2025), which synthesizes text in a reference speaker’s voice. We also filter out unrealistic samples using a reasoning model, Qwen3-235B-A22B-Thinking-2507 (Yang et al., 2025). Further details of the implementation and filtering process are described in Appendix F. This process yields around 80K realistic two-speaker samples that capture diverse TPI scenarios.

Response Strategy. For training, we include not only the spoken inputs but also the corresponding model responses. The core principle of our response strategies lies in the initial classification of each interruption as described in Section 2.2. Each interruption case is categorized as either *actionable* (C_A), where the model considers responding, or *ignorable* (C_I), where it does not.

Inspired by prior literature that distinguishes interruptions as cooperative or disruptive (Yang et al., 2022; Murata, 1994; Goldberg, 1990), we define a condition for actionable cases as third-party utterances that provide information directly contributing to the primary user’s objective—enhancing dialogue efficiency, improving task alignment, and preventing errors through supplementary helpful input. Based on this principle, we distill our focus into four representative actionable categories: (i) *Correction & Disambiguation*, (ii) *Feasibility Constraint*, (iii) *Goal-oriented Suggestion*, and (iv) *Co-*

operative Addition & Refinement. Each category’s definition, example, and π_A are elaborated in Appendix D and Figure 7.

Although some other residual cases could be interpreted as actionable, at this stage of the study, we group all remaining cases under the label of *ignorable* interruptions for practical purposes, in which the third-party utterance does not contribute to the user’s task—for instance, off-topic remarks or disruptive interjections. We use an LLM to automatically assign labels and generate textual responses consistent with the appropriate strategy.

Speaker-Aware Hard Negatives. Prior works in the vision-language domain (Koishigarina et al., 2025; Yuksekgonul et al., 2023) have shown that multimodal learning can exhibit unimodal shortcut learning—over-relying on one modality and failing to exploit complementary signals. We posit similar tendencies in the speech domain where models rely disproportionately on semantic cues for interruption detection, overlooking acoustic signals such as speaker changes.

To make the model prioritize acoustic cues, we introduce speaker-aware hard negatives designed to explicitly discourage such semantic shortcut learning. Specifically, we first employ an LLM to select, from the corpus, transcripts that are semantically ambiguous—those that, when stripped of the speaker labels, could plausibly be interpreted as single-speaker monologues (e.g., self-repair or stream-of-consciousness utterance, Appendix I). For these selected instances, we generate counterfactual audio samples using only the voice of a primary speaker. By keeping the transcripts identical, the model is forced to rely on acoustic signals as the sole informative cue for detecting interruptions, effectively reducing reliance on semantic shortcuts.

4 Benchmark

4.1 TPI-Bench

TPI-Bench evaluates two abilities: 1) situation-discriminative response (TPI-Test) and 2) speaker interruption detection (Janus-Test), with an additional real-world test set used for validating transferability beyond synthetic data.

TPI-Test. This set consists of 2K samples, sampled from the corpus, ensuring no overlap with the training data, while maintaining a balanced distribution across all TPI scenarios. It serves as the standard evaluation set for measuring the model’s ability to generate situation-aligned responses according to predefined answer strategies.

Janus-Test. To isolate the impact of acoustic signals, we construct Janus-Test consisting of 2K samples selected from our corpus by filtering ambiguous samples where the concatenation of U_p and U_{tp} is semantically coherent enough to be interpreted as a single-speaker utterance (Figure 3). This adversarial setup compels the model to ignore misleading semantic cues and leverage acoustic cues to correctly infer the number of speakers in the utterance.

Real-world test set. To validate transferability from synthetic data, we collected real-world voice data under TPI scenarios, consisting of 100 high-quality audio samples. We sourced data from three distinct multi-party conversations for diverse scenarios: (1) **AMI Meeting Corpus** (Carletta et al., 2005), representing real-world meeting scenarios; (2) **Friends-MMC** (Wang et al., 2024b), a multi-party sitcom dataset; and (3) **Human Recordings**, collected in varied acoustic environments (e.g., reverb-heavy rooms, outdoors) to mimic daily usage. We utilized a rigorous two-stage filtering pipeline involving 1) LLM-based filtering and 2) human verification. Detailed curation criteria and the filtering process are provided in Appendix G.

All samples in TPI-Bench are filtered to have a word error rate (WER) of zero using whisper-large-v3 (Radford et al., 2022).

4.2 Evaluation Strategies

Response Strategy Following (RSF). Response Strategy Following (RSF) is a binary metric that evaluates whether a model follows the predefined strategy (either π_A or π_I) for actionable (C_A) or ignorable (C_I) interruptions. In TPI-Test, a score of 1 indicates correct strategy adherence. In contrast, in Janus-Test, a score of 1 indicates an error,

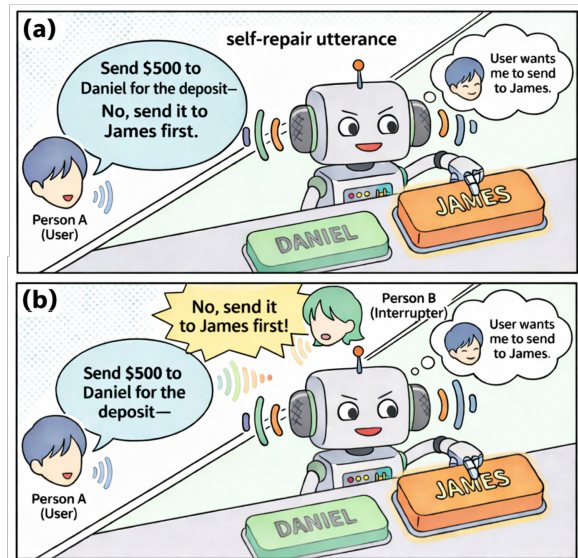


Figure 3: **Illustration of the intuition behind TPI-Bench.** In order to verify whether the model leverages acoustic information to distinguish the presence of an interruption, we introduce the (a) Janus-Test and (b) TPI-Test. We extract (b) instances of interruptions where the user could have plausibly spoken the entire segment alone, and (a) re-synthesize them as a single continuous utterance by one speaker. Consequently, acoustic cues become the only discernible signal, allowing comparison to determine whether the model truly distinguishes interruptions when responding.

where the model incorrectly infers single-speaker utterances as interruptions. Thus, a desirable model should achieve high scores on TPI-Test and low scores on Janus-Test.

Overall Helpfulness (OH). Overall Helpfulness (OH) evaluates the naturalness and usefulness of the model’s response under interruption situations on a 5-point Likert scale using an LLM with access to ground-truth speaker information. In TPI-Test, high scores (4–5) indicate that the model correctly identifies the third-party interrupter and responds in third-party interruption situations while prioritizing the primary user, while low scores (1–2) reflect failures such as conflating U_p and U_{tp} into a single incoherent query. In Janus-Test, high scores (4–5) indicate correct treatment of the input as a single-speaker utterance, whereas low scores (1–2) suggest the model was misled into treating it as an interruption. A score of 3 in either benchmark denotes a cautious but incomplete response. We perform evaluation using an LLM-as-a-judge (Gu et al., 2025), and provide the full evaluation prompts in Appendix L.

Auxiliary Metrics. To further probe models’ sensitivity to acoustic speaker changes, we compute ROUGE-L and BLEU scores between model

Condition	Realism Score
Scenario only (text)	2.87 (95% CI: [2.836, 2.904])
Scenario + Audio (acoustic)	2.63 (95% CI: [2.61, 2.65])

Table 1: Human evaluation results on TPI-Test assessing realism and naturalness. We report the mean score (3 scale) with 95% confidence intervals. (1: Unrealistic, 2: Moderate, 3: Realistic)

responses of shared pairs from TPI-Test and Janus-Test. These pairs share identical transcriptions but differ acoustically: one is a single-speaker utterance, while the other is a two-speaker interruption. High ROUGE-L and BLEU scores indicate that a model produced almost identical responses for both, revealing that it does not treat the acoustic shift as a critical contextual cue.

4.3 Benchmark Validity

In order to evaluate whether the benchmark samples reflect realistic conversational interruption scenarios, we conduct a human evaluation on Amazon Mechanical Turk, assessing the naturalness and realism of the samples, with details provided in Appendix E. As shown in Table 1, both conditions receive high realism ratings on a 3-point Likert scale. The Scenario-only (text) condition achieves an average score of 2.87, while the Scenario + Audio (acoustic) condition scores 2.63, indicating that the generated audio preserves scenario realism. Overall, these results demonstrate that TPI-Test captures realistic interruption phenomena at both the scenario and acoustic levels, establishing its validity as a benchmark for conversational interruption handling.

5 Experiments

5.1 Baselines and Experiments

Baselines. We evaluate ChatGPT-4o-audio (OpenAI, 2025), Kimi-Audio-Instruct-7B (KimiTeam et al., 2025), Vita-Audio-Instruct-7B (Long et al., 2025), Qwen2.5-Omni-7B (Xu et al., 2025a) and Qwen3-Omni-30B-A3B-Instruct (Xu et al., 2025b) on TPI-Bench. In addition, to demonstrate the effectiveness of our training data, we choose Qwen2.5-Omni-7B as our baseline. Further details of the training are described in the Appendix H.

Experiments. We evaluate both existing spoken language models and fine-tuned ablated models along four dimensions: (i) TPI-awareness, (ii) robustness against semantic shortcut learning, (iii) preservation of core voice interaction capabilities,

(iv) robustness in detecting voice transitions on real-world data. For this purpose, we use TPI-Bench and the eight sub-benchmarks included in VoiceBench (Chen et al., 2024b). We also conduct an ablation study by controlling the inclusion of (i) our corpus data, (ii) dyadic interaction data from VoiceAssistant-400K (Chen et al., 2024a), and (iii) hard negative samples, and analyze their individual contributions. The model trained solely on our corpus data is denoted as **TPI-Base**, without additional dyadic interaction data or hard negative samples. The model trained using our corpus data together with VoiceAssistant-400K dyadic interaction data is denoted as **TPI-VA**. Finally, the model additionally augmented with the extended hard negative samples is denoted as **TPI-Full**.

To measure whether TPI-aware training degrades general spoken interaction abilities, we use VoiceBench (Chen et al., 2024b), which covers diverse evaluation scenarios, including open-ended QA from both human and TTS sources (AlpacaEval, CommonEval, WildVoice), multiple-choice QA (OpenBookQA, MMSU), instruction following (IFEval), safety/adversarial prompts (AdvBench), and reference-based QA (SD-QA), thereby providing a broad testbed for SLMs. We follow the official evaluation pipeline and prompts, but replace their judgment model with Qwen3-235B-A22B-Instruct-2507 (Yang et al., 2025), ensuring consistency and scalability in evaluation. We also show human evaluation results to check if LLM evaluation results are well aligned with human judgements in Appendix E.

5.2 Results

Results across the four evaluation dimensions are demonstrated in Table 2, 3 and 4. Consistent with our hypothesis, current spoken language models show poor TPI-awareness with low sensitivity to acoustic change: they implicitly assume a single-speaker setting, leading to low performance on TPI-Test and Janus-Test with high scores on BLEU and ROUGE-L. In contrast, TPI-Full (Ours), fine-tuned on our full dataset, demonstrates the most robust and balanced performance across all benchmarks. It achieves high performance on both TPI-Test and Janus-Test, maintaining stability in RSF and OH scores while preventing performance degradation on general benchmarks (Table 3).

We further conduct an ablation study to understand the contribution of each data component.

Model	Data			TPI-Test		Janus-Test		BLEU(↓)	ROUGE-L(↓)
	C	VA	HN	RSF(↑)	OH(↑)	RSF(↓)	OH(↑)		
ChatGPT-4o-audio	×	×	×	0.21	4.06	0.09	4.95	0.89	0.93
Kimi-Audio-Instruct-7B	×	×	×	0.22	3.29	0.13	4.52	0.94	0.99
VITA-Audio-Instruct-7B	×	×	×	0.21	3.26	0.10	4.37	0.42	0.71
Qwen2.5-Omni-7B	×	×	×	0.24	3.22	0.12	4.44	0.31	0.53
Qwen3-Omni-30B-A3B-Instruct	×	×	×	0.19	3.55	0.01	4.45	0.20	0.39
TPI-Base (w/o VA, HN)	✓	×	×	0.82	4.32	0.86	3.54	0.46	0.63
TPI-VA (w/o HN)	✓	✓	×	0.82	4.29	0.67	3.75	0.39	0.58
TPI-Full	✓	✓	✓	0.83	4.16	0.16	4.80	0.12	0.34

Table 2: **Evaluation and ablation study of baselines and ours on TPI-Bench.** We report results of various baseline models and use the Qwen2.5-Omni-7B model as the reference point for our ablation studies. RSF denotes Response Strategy Following, and OH denotes Overall Helpfulness. BLEU and ROUGE-L are evaluated on shared utterances between TPI-Test and Janus-Test. Higher similarity indicates that the model produces consistent responses despite acoustic discrepancy. The highest score among the ablation variants is highlighted in bold. Here, **C**, **VA**, and **HN** denote models trained on our corpus data, dyadic voice assistant data, and hard negatives, respectively.

Datasets	Model	Performance			
AlpacaEval CommonEval SD-QA MMSU	Baseline	3.78	3.67	28.39	61.22
	TPI-Base	4.07	3.24	32.58	50.80
	TPI-VA	4.06	3.97	35.34	58.88
	TPI-Full	4.12	3.93	36.08	59.14
OpenBookQA IFEval AdvBench WildVoice	Baseline	80.44	0.42	0.98	3.53
	TPI-Base	66.81	0.46	0.95	2.83
	TPI-VA	80.22	0.40	1.00	3.59
	TPI-Full	80.00	0.46	1.00	3.64

Table 3: **Ablation study on VoiceBench.** Comprehensive performance comparison between Baseline (Qwen2.5-Omni-7B) and Ours (TPI) across the 8 datasets of the VoiceBench Benchmark.

Method	TPI-Test		Real-world	
	RSF (↑)	OH (↑)	RSF (↑)	OH (↑)
Baseline	0.24	3.22	0.17	3.21
Ours	0.83	4.16	0.60	4.25

Table 4: **Performance gain comparison between synthetic and real-world data.** Performance comparison between the baseline and ours on synthetic (TPI-Test) and real-world data.

When removing the two components, dyadic interaction data and hard negatives, and training solely on our corpus, the model (TPI-Base) exhibits domain overfitting, leading to degraded general spoken interaction capabilities, as shown in Table 3. Restoring the single-speaker data (TPI-VA) effectively recovers these general capabilities. However, despite being trained on both single- and multi-speaker utterances, TPI-VA still performs poorly on Janus-Test. This failure indicates that simply balancing the data distribution with respect to speaker numbers does not force the model to utilize acoustic information. Instead, the model falls into *semantic shortcut learning*—relying on textual patterns rather than acoustic patterns to distinguish interruptions.

We further validate the robustness of our model with real-world samples, as shown in Table 4. Our model delivers substantial gains on the synthetic benchmark (RSF +0.59, OH +0.94) which persist under real-world conditions, boosting the baseline from RSF 0.17 to 0.60 (+0.43) and OH 3.21 to 4.25 (+1.04). The fact that the magnitude of improvement is comparable across both settings confirms that our model avoids overfitting to synthetic patterns. Instead, it demonstrates that the learned capability to distinguish voice transitions is robustly preserved even in real-speaker scenarios.

5.3 Analyses

5.3.1 Embedding Visualization

To better understand the impact of our training strategy, we visualize model embeddings from LLM representations with t-SNE (Figure 4). The baseline model produces heavily overlapping embeddings for single-speaker and two-speaker inputs, indicating no acoustic discrimination. Even after training with corpus data (TPI-Base) and dyadic data (TPI-VA), embeddings of single-speaker (red, green) inputs still overlap with two-speaker (blue). Specifically, the significant overlap between hard negatives (red) and two-speaker (blue) inputs reveals that their semantic equivalence overshadows acoustic distinctions, indicating the occurrence of semantic shortcut learning.

By contrast, additionally training with our hard negatives (Ours) yields a well-structured embedding space with three clearly separated clusters. Notably, the hard negative cluster (red) lies between the other two, reflecting its semantic similarity to interruptions (blue) but acoustic alignment with single-speaker utterances (green). This demonstrates that our approach compels the model

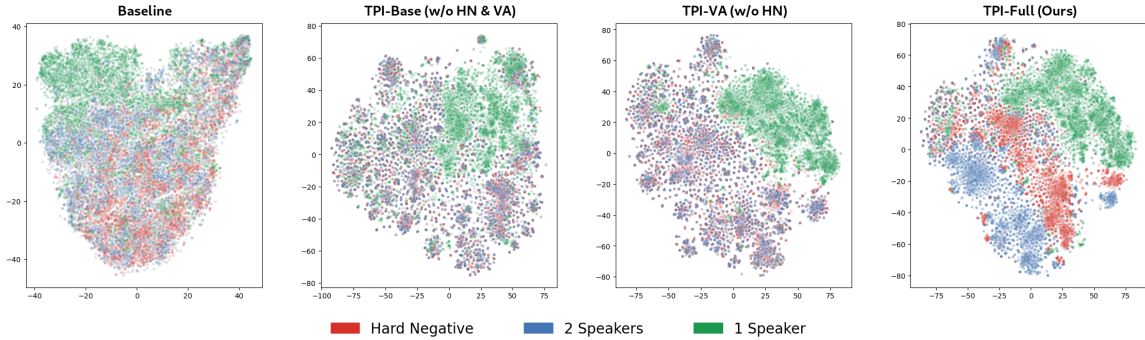


Figure 4: **t-SNE visualization.** Training with hard negatives (HN) yields a structured space with distinct clusters, where the cluster sits between interruptions and single-speaker utterances, balancing semantic and acoustic alignment.

Method	Pref. (%)	Tie (%)	Unpref. (%)
<i>Independent Quality Preference</i>			
TPI-Corpus (GT)	64.75	5.50	29.75
Ours	66.05	7.63	26.32
<i>Direct Model Comparison (A/B Preference Test)</i>			
Ours vs. Baseline	60.50	11.50	28.00

Table 5: **Human preference evaluation on TPI-Test.** The top section validates our framework, showing that both our samples of reference strategies and generated responses from TPI-Full (Ours) are similarly preferred, aligning well with human expectations. The bottom section presents a direct A/B comparison, demonstrating that our model is preferred over the Qwen2.5-Omni-7B baseline (“Unpref.” indicates preference for the baseline).

to move beyond semantic shortcuts and induce genuine acoustic discrimination.

5.3.2 Human Evaluations

Human Preference Evaluations. In Section 3.1, we introduced actionable and ignorable strategies, along with four representative actionable categories (Figure 7). Although these criteria can be varied by user preferences, our primary goal is twofold: to validate the human alignment of our reference strategies and to demonstrate the practical superiority of our trained model compared to existing baselines, ultimately demonstrating the applicability and effectiveness of our proposed framework.

To this end, we conducted two sets of human evaluations on Amazon Mechanical Turk. First, we compared (i) reference answers sampled from our corpus according to our actionable/ignorable framework, and (ii) responses generated by our trained model (TPI-Full) on TPI-Test. A total of 100 participants each evaluated four randomly sampled scenarios, yielding 400 evaluation instances per case. Participants were blind to whether the case was actionable or ignorable, and were simply asked: “If you were the user, would you be satisfied with

how the model handled this interruption?”.

As summarized in Table 5, ground-truth responses of our corpus samples received a 64.75% “Preferred” rating, while model responses achieved a similar preference rate of 66.05% ($p < 0.01$). These results demonstrate two key findings: (i) our proposed framework produces reference responses that align well with human expectations, and (ii) our trained model successfully learns to implement this framework, yielding responses that are equally preferred. Together, this confirms the validity of both our dataset design and our framework.

Second, to validate the practical improvements of our model, we conducted a direct A/B preference test between TPI-Full and the baseline (Qwen2.5-Omni-7B) model. We randomly sampled instances from TPI-Test and asked 40 independent annotators to evaluate 5 examples each (200 ratings in total). Annotators were asked: “Which response would you prefer from the assistant?”. As shown in the bottom section of Table 5, ours was preferred by 60.5% of the raters, outperforming the baseline (28.0%) by a significant margin. This highlights that explicitly training models to handle third-party interruptions substantially enhances the overall user experience, preventing the common failure modes observed in the baseline model.

5.3.3 Can Explicit Prompting Solve TPI?

Evaluating the Limits of Speaker-Aware Instructions. While we showcase the vulnerability of existing spoken language models to third-party interruptions, a natural question arises: could providing explicit context via instructions reduce the failure rate without fine-tuning? To determine whether the poor performance stems from a lack of acoustic understanding or simply the absence of speaker-aware context, we evaluate the baseline model (Qwen2.5-Omni-7B) using explicit speaker-aware prompts.

Depending on the ground-truth scenario, we provided the baseline model with the following context-rich instructions:

- **Actionable cases:** “You are in a room with multiple people. Carefully reflect the second speaker’s interruption in your answer to the primary user.”
- **Ignorable cases:** “You are in a room with multiple people. Ignore the second speaker’s interruption and answer only the primary user’s request.”

As shown in Table 6, explicit speaker-aware prompts yield no meaningful overall improvements. While the Response Strategy Following (RSF) score marginally increases, further analysis indicates this stems from the model’s tendency to indiscriminately apply the “ignore” instruction. By mechanically disregarding the second utterance, the model coincidentally aligns with ignorable scenarios, but this rigid behavior slightly decreases Overall Helpfulness (OH). Fundamentally, the baseline still fails to distinguish speaker boundaries, continuing to erroneously concatenate utterances during actionable interruptions.

These observations demonstrate that prompt engineering alone is insufficient for accurate speaker attribution in continuous audio streams. Ultimately, true TPI-awareness requires speaker attribution capabilities grounded in speaker-aware audio modeling, a property that our proposed model successfully achieves.

Model	TPI-Test	
	RSF (↑)	OH (↑)
Baseline (Contextless)	0.24	3.22
Baseline (Speaker-aware prompting)	0.37	3.17
Ours	0.83	4.16

Table 6: **Performance comparison between contextless and speaker-aware prompting settings.** Results indicate that explicit instruction prompts cannot overcome the baseline model’s lack of acoustic speaker attribution ($p < 0.05$).

6 Related Works

Conversational Spoken Dialogue Dataset. The capabilities of modern Spoken Language Models (SLMs) are increasingly evaluated across diverse conversational scenarios captured in large-scale synthetic datasets (Lee et al., 2023; Koudounas et al., 2025; Si et al., 2023). Such datasets are

designed not only to probe a model’s comprehension of conversational dynamics, but also its sensitivity to nuanced acoustic and paralinguistic features, such as emotion and prosody (Ao et al., 2025; Cheng et al., 2025; Chen et al., 2024b; Yan et al., 2025; Wang et al., 2025a). However, these efforts have predominantly focused on dyadic interactions, modeling conversations between a single user and an agent, thereby leaving a critical research gap concerning realistic multi-speaker scenarios such as third-party interruptions (Wang et al., 2025a). To address this, we introduce a large-scale dataset grounded in established interruption taxonomies (Yang et al., 2022; Murata, 1994; Goldberg, 1990). Unlike prior work on two-party dialogues, our dataset targets triadic dynamics, enabling models to move beyond speech comprehension toward multi-party dialogue understanding.

Processing Multi-Speaker Speech. Recent research has explored the use of large language models (LLMs) for multi-speaker scenarios, particularly focusing on automatic speech recognition (ASR) and speaker diarization (Yin et al., 2025; Lin et al., 2025; Wang et al., 2024a). These approaches have proven effective at disentangling multi-speaker utterances and identifying who spoke what. More recently, these lines of work have extended to an instruction-following manner, enabling selective transcription of a target speaker’s utterances in multi-speaker environments (Meng et al., 2025). However, we emphasize interactional dynamics rather than treating competing voices as signals to be separated or discarded (Xu et al., 2025c; Wang et al., 2025b; He and Whitehill, 2025). Our approach enables models to decide whether to ignore or engage, mirroring human-like processing where acoustic variations guide conversational intelligence.

7 Conclusion

In this paper, we established the first comprehensive framework for TPI-aware voice assistants, transforming the subjective challenge of interruption handling into a measurable task. Through our dataset and benchmark, we demonstrated that models can be guided to handle interruptions by prioritizing acoustic cues over reliance on semantic cues. We believe our work establishes a foundational resource for overcoming text-dominated unimodal reliance in SLMs, paving the way for more robust multi-party spoken interaction.

Limitations

Various Response Strategies. While we propose a reference response strategy that was verified to be effective within our framework, we acknowledge that appropriate interruption handling is inherently subjective and user-dependent. There may exist diverse alternative strategies beyond our reference strategy. Future research could explore these variations by examining how different user demographics, personas, and personal backgrounds influence preferences for specific response strategies across varying social contexts.

Turn-Based Interaction and Full-Duplex Extension. Our work focuses on interruption awareness within a turn-based voice assistant setting, where the system responds after receiving an utterance. In realistic multi-party conversations, however, not every interruption warrants an immediate response from the assistant, and effective behavior may also require deciding whether and when to speak which is a core element of full duplex dialogue system research. Extending interruption-aware modeling to a full-duplex setting, where the assistant can continuously listen, defer, or strategically interject, is a necessary step toward truly natural multi-party dialogue systems and remains an important direction for future work.

Ethics

The datasets and automation pipeline developed as part of this work will be made publicly available under the MIT License. Large Language Models were used solely to assist with the linguistic aspects of the manuscript, such as paraphrasing and spell-checking, and did not contribute to the technical content of the research. Although the dataset was not designed to include offensive content, data samples that may potentially contain such content were filtered out using large language models. (Chen et al., 2024a)

Acknowledgements

This work was supported by Institute of Information & Communications Technology Planning & Evaluation (IITP) grants funded by the Korea government (MSIT) [NO.RS-2021-II211343, Artificial Intelligence Graduate School Program (Seoul National University); No.2022-0-00959, RS-2022-II220959], by National Research Foundation of Korea (NRF) grant [No.2022R1A3B1077720, 2022R1A5A7083908], BK21 FOUR Program of

the Education and Research Program for Future ICT Pioneers, Seoul National University in 2026. This research was also conducted as part of the Sovereign AI Foundation Model Project (Data Track), organized by the Ministry of Science and ICT (MSIT) and supported by the National Information Society Agency (NIA), S.Korea (2025-AI Data-wi43).

References

- Junyi Ao, Yuancheng Wang, Xiaohai Tian, Dekun Chen, Jun Zhang, Lu Lu, Yuxuan Wang, Haizhou Li, and Zhizheng Wu. 2025. *Sd-eval: A benchmark dataset for spoken dialogue understanding beyond words*. *Preprint*, arXiv:2406.13340.
- Amanda Baughan, Xuezhi Wang, Ariel Liu, Allison Mercurio, Jilin Chen, and Xiao Ma. 2023. *A mixed-methods approach to understanding user trust after voice assistant failures*. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '23, New York, NY, USA. Association for Computing Machinery.
- Shiye Cao, Jiwon Moon, Amama Mahmood, Victor Nikhil Antony, Ziang Xiao, Anqi Liu, and Chien-Ming Huang. 2025. *Interruption handling for conversational robots*. *Preprint*, arXiv:2501.01568.
- Jean Carletta, Simone Ashby, Sebastien Bourban, Mike Flynn, Mael Guillemot, Thomas Hain, Jaroslav Kadlec, Vasilis Karaiskos, Wessel Kraaij, Melissa Kronenthal, Guillaume Lathoud, Mike Lincoln, Agnes Lisowska, Iain McCowan, Wilfried Post, Dennis Reidsma, and Pierre Wellner. 2005. *The ami meeting corpus: a pre-announcement*. In *Proceedings of the Second International Conference on Machine Learning for Multimodal Interaction*, MLMI'05, page 28–39, Berlin, Heidelberg. Springer-Verlag.
- Wenxi Chen, Ziyang Ma, Ruiqi Yan, Yuzhe Liang, Xiquan Li, Ruiyang Xu, Zhikang Niu, Yanqiao Zhu, Yifan Yang, Zhanxun Liu, and 1 others. 2024a. *Slam-omni: Timbre-controllable voice interaction system with single-stage training*. *arXiv preprint arXiv:2412.15649*.
- Yiming Chen, Xianghu Yue, Chen Zhang, Xiaoxue Gao, Robby T. Tan, and Haizhou Li. 2024b. *Voicebench: Benchmarking llm-based voice assistants*. *Preprint*, arXiv:2410.17196.
- Xize Cheng, Ruofan Hu, Xiaoda Yang, Jingyu Lu, Dongjie Fu, Zehan Wang, Shengpeng Ji, Rongjie Huang, Boyang Zhang, Tao Jin, and Zhou Zhao. 2025. *Voxdialogue: Can spoken dialogue systems understand information beyond words?* In *The Thirteenth International Conference on Learning Representations*.
- Rahul Goel, Waleed Ammar, Aditya Gupta, Siddharth Vashishtha, Motoki Sano, Faiz Surani, Max Chang,

- HyunJeong Choe, David Greene, Kyle He, Rattima Nitisaroj, Anna Trukhina, Shachi Paul, Pararth Shah, Rushin Shah, and Zhou Yu. 2023. [Presto: A multilingual dataset for parsing realistic task-oriented dialogs](#). *Preprint*, arXiv:2303.08954.
- Julia Goldberg. 1990. [Interrupting the discourse on interruptions](#). *Journal of Pragmatics*, 14:883–903.
- Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, Saizhuo Wang, Kun Zhang, Yuanzhuo Wang, Wen Gao, Lionel Ni, and Jian Guo. 2025. [A survey on llm-as-a-judge](#). *Preprint*, arXiv:2411.15594.
- Xinlu He and Jacob Whitehill. 2025. [Survey of end-to-end multi-speaker automatic speech recognition for monaural audio](#). *Preprint*, arXiv:2505.10975.
- Heeseung Kim, Soonshin Seo, Kyeongseok Jeong, Ohsung Kwon, Soyeon Kim, Jungwhan Kim, Jaehong Lee, Eunwoo Song, Myungwoo Oh, Jungwoo Ha, Sungroh Yoon, and Kang Min Yoo. 2024. [Paralinguistics-aware speech-empowered large language models for natural conversation](#). *Preprint*, arXiv:2402.05706.
- KimiTeam, Ding Ding, Zeqian Ju, Yichong Leng, Songxiang Liu, Tong Liu, Zeyu Shang, Kai Shen, Wei Song, Xu Tan, Heyi Tang, Zhengtao Wang, Chu Wei, Yifei Xin, Xinran Xu, Jianwei Yu, Yutao Zhang, Xinyu Zhou, Y. Charles, and 21 others. 2025. [Kimi-audio technical report](#). *Preprint*, arXiv:2504.18425.
- Darina Koishigarina, Arnas Uselis, and Seong Joon Oh. 2025. [Clip behaves like a bag-of-words model cross-modally but not uni-modally](#). *Preprint*, arXiv:2502.03566.
- Alkis Koudounas, Moreno La Quatra, and Elena Baralis. 2025. [Deepdialogue: A multi-turn emotionally-rich spoken dialogue dataset](#). *Preprint*, arXiv:2505.19978.
- Yejin Kwon, Taewoo Kang, Hyunsoo Yoon, and Changouk Kim. 2025. [M3-slu: Evaluating speaker-attributed reasoning in multimodal large language models](#). *Preprint*, arXiv:2510.19358.
- Keon Lee, Kyumin Park, and Daeyoung Kim. 2023. [Dailytalk: Spoken dialogue dataset for conversational text-to-speech](#). *Preprint*, arXiv:2207.01063.
- Willem J.M. Levelt. 1983. [Monitoring and self-repair in speech](#). *Cognition*, 14(1):41–104.
- Yuke Lin, Ming Cheng, Ze Li, Beilong Tang, and Ming Li. 2025. [Diarization-aware multi-speaker automatic speech recognition via large language models](#). *Preprint*, arXiv:2506.05796.
- Yinhong Liu, Yimai Fang, David Vandyke, and Nigel Collier. 2024. [Toad: Task-oriented automatic dialogs with diverse response styles](#). *Preprint*, arXiv:2402.10137.
- Zuwei Long, Yunhang Shen, Chaoyou Fu, Heting Gao, Lijiang Li, Peixian Chen, Mengdan Zhang, Hang Shao, Jian Li, Jinlong Peng, Haoyu Cao, Ke Li, Rongrong Ji, and Xing Sun. 2025. [Vita-audio: Fast interleaved cross-modal token generation for efficient large speech-language model](#). *Preprint*, arXiv:2505.03739.
- Lingwei Meng, Shujie Hu, Jiawen Kang, Zhaoqing Li, Yuejiao Wang, Wenxuan Wu, Xixin Wu, Xunying Liu, and Helen Meng. 2025. [Large language model can transcribe speech in multi-talker scenarios with versatile instructions](#). *Preprint*, arXiv:2409.08596.
- Kumiko Murata. 1994. [Intrusive or co-operative? a cross-cultural study of interruption](#). *Journal of Pragmatics*, 21(4):385–400.
- OpenAI. 2025. [Chatgpt](#). <https://chat.openai.com>. Large language model interface. Accessed: 2025-09-22.
- Vineel Pratap, Qiantong Xu, Anuroop Sriram, Gabriel Synnaeve, and Ronan Collobert. 2020. [MLS: A large-scale multilingual dataset for speech research](#). In *Interspeech 2020*, pages 2757–2761. ISCA.
- Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. [Robust speech recognition via large-scale weak supervision](#). *arXiv preprint*.
- Resemble AI. 2025. [Chatterbox-TTS](#). <https://github.com/resemble-ai/chatterbox>. GitHub repository.
- Shuzheng Si, Wentao Ma, Haoyu Gao, Yuchuan Wu, Ting-En Lin, Yinpei Dai, Hangyu Li, Rui Yan, Fei Huang, and Yongbin Li. 2023. [Spokenwoz: A large-scale speech-text benchmark for spoken task-oriented dialogue agents](#). In *Advances in Neural Information Processing Systems*, volume 36, pages 39088–39118. Curran Associates, Inc.
- Joe Stacey, Jianpeng Cheng, John Torr, Tristan Guigue, Joris Driesen, Alexandru Coca, Mark Gaynor, and Anders Johannsen. 2024a. [LUCID: LLM-generated utterances for complex and interesting dialogues](#). In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 4: Student Research Workshop)*, pages 56–74, Mexico City, Mexico. Association for Computational Linguistics.
- Joe Stacey, Jianpeng Cheng, John Torr, Tristan Guigue, Joris Driesen, Alexandru Coca, Mark Gaynor, and Anders Johannsen. 2024b. [Lucid: Llm-generated utterances for complex and interesting dialogues](#). *Preprint*, arXiv:2403.00462.
- Madiha Tabassum, Tomasz Kosiński, Alisa Frik, Nathan Malkin, Primal Wijesekera, Serge Egelman, and Heather Richter Lipford. 2020. [Investigating users’ preferences and expectations for always-listening voice assistants](#). *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 3(4).

- Quan Wang, Yiling Huang, Guanlong Zhao, Evan Clark, Wei Xia, and Hank Liao. 2024a. [DiarizationLM: Speaker diarization post-processing with large language models](#). In *Interspeech 2024*, pages 3754–3758. ISCA.
- Shuai Wang, Zhaokai Sun, Zhennan Lin, Chengyou Wang, Zhou Pan, and Lei Xie. 2025a. [Msu-bench: Towards understanding the conversational multi-talker scenarios](#). *Preprint*, arXiv:2508.08155.
- Weiqing Wang, Taejin Park, Ivan Medennikov, Jinhua Wang, Kunal Dhawan, He Huang, Nithin Rao Koluguri, Jagadeesh Balam, and Boris Ginsburg. 2025b. [Speaker targeting via self-speaker adaptation for multi-talker asr](#). *Preprint*, arXiv:2506.22646.
- Yueqian Wang, Xiaojun Meng, Yuxuan Wang, Jianxin Liang, Qun Liu, and Dongyan Zhao. 2024b. [Friends-mmc: A dataset for multi-modal multi-party conversation understanding](#). *Preprint*, arXiv:2412.17295.
- Boyong Wu, Chao Yan, Chen Hu, Cheng Yi, Chengli Feng, Fei Tian, Feiyu Shen, Gang Yu, Haoyang Zhang, Jingbei Li, Mingrui Chen, Peng Liu, Wang You, Xiangyu Tony Zhang, Xingyuan Li, Xuerui Yang, Yayue Deng, Yechang Huang, Yuxin Li, and 90 others. 2025. [Step-audio 2 technical report](#). *Preprint*, arXiv:2507.16632.
- Lishan Xie, Canmian Liu, and Dongmei Li. 2022. [Proactivity or passivity? an investigation of the effect of service robots’ proactive behaviour on customer co-creation intention](#). *International Journal of Hospitality Management*, 106:103271.
- Zhifei Xie and Changqiao Wu. 2024. [Mini-omni: Language models can hear, talk while thinking in streaming](#). *Preprint*, arXiv:2408.16725.
- Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, Bin Zhang, Xiong Wang, Yunfei Chu, and Junyang Lin. 2025a. [Qwen2.5-omni technical report](#). *Preprint*, arXiv:2503.20215.
- Jin Xu, Zhifang Guo, Hangrui Hu, Yunfei Chu, Xiong Wang, Jinzheng He, Yuxuan Wang, Xian Shi, Ting He, Xinfa Zhu, Yuanjun Lv, Yongqi Wang, Dake Guo, He Wang, Linhan Ma, Pei Zhang, Xinyu Zhang, Hongkun Hao, Zishan Guo, and 19 others. 2025b. [Qwen3-omni technical report](#). *Preprint*, arXiv:2509.17765.
- Shitong Xu, Yiyuan Yang, Niki Trigoni, and Andrew Markham. 2025c. [Target speaker extraction through comparing noisy positive and negative audio enrollments](#). *Preprint*, arXiv:2502.16611.
- Ruiqi Yan, Xiquan Li, Wenxi Chen, Zhikang Niu, Chen Yang, Ziyang Ma, Kai Yu, and Xie Chen. 2025. [Uro-bench: Towards comprehensive evaluation for end-to-end spoken dialogue models](#). *Preprint*, arXiv:2502.17810.
- An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chuji Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. [Qwen3 technical report](#). *Preprint*, arXiv:2505.09388.
- Liu Yang, Catherine Achard, and Catherine Pelachaud. 2022. [Annotating interruption in dyadic human interaction](#). In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2292–2297, Marseille, France. European Language Resources Association.
- Han Yin, Yafeng Chen, Chong Deng, Luyao Cheng, Hui Wang, Chao-Hong Tan, Qian Chen, Wen Wang, and Xiangang Li. 2025. [Speakerlm: End-to-end versatile speaker diarization and recognition with multimodal large language models](#). *Preprint*, arXiv:2508.06372.
- Mert Yuksekgonul, Federico Bianchi, Pratyusha Kalluri, Dan Jurafsky, and James Zou. 2023. [When and why vision-language models behave like bags-of-words, and what to do about it?](#) *Preprint*, arXiv:2210.01936.
- Qinglin Zhang, Luyao Cheng, Chong Deng, Qian Chen, Wen Wang, Siqi Zheng, Jiaqing Liu, Hai Yu, Chao-hong Tan, Zhihao Du, and Shiliang Zhang. 2025. [Omniflatten: An end-to-end gpt model for seamless voice conversation](#). *Preprint*, arXiv:2410.17799.

A Failure Case Analysis

Our analysis of Spoken Language Models (SLMs), including open-source models such as **Kimi-Audio** (KimiTeam et al., 2025), **VITA-Audio** (Long et al., 2025), and **Qwen2.5-Omni** (Xu et al., 2025a) and closed-source model such as a **ChatGPT-4o-audio-preview** (OpenAI, 2025), reveals a critical and consistent vulnerability in Third-Party Interruption (TPI) scenarios. These models consistently fail to differentiate between the primary user and the interrupter, instead treating the interrupter’s utterance as a direct continuation of the primary user’s command. In effect, even when the audio input originates from two distinct speakers, the models perceive and process it as a single, unified instruction from single-speaker. This could trigger unauthorized actions such as financial transactions or data deletion.

A.1 Open-Source Models

Kimi-Audio: Example 1

Primary User: Which company developed the video game console Xbox...

Third Party: No, ask who made the PlayStation instead.

Model Answer (TPI): Sony Computer Entertainment developed the PlayStation video game console.

Model Answer (Janus): Sony Computer Entertainment developed the PlayStation video game console.

Kimi-Audio: Example 2

Primary User: Me and my friends are telling jokes tonight and I need a really good one...

Third Party: Just clean ones or is anything fair game?

Model Answer (TPI): As long as the jokes are appropriate for all ages and do not contain any offensive language or content, they are fair game. Here are a few clean jokes that you and your friends might enjoy:

Model Answer (Janus): As long as the jokes are appropriate for all ages and do not contain any offensive language or content, they are fair game. Here are a few clean jokes that you and your friends might enjoy.

VITA-Audio: Example 1

Primary User: Construct a statement that conveys the idea of innovation.

Third Party: Hold on. Why not just say something about thinking differently or starting something new instead.

Model Answer (TPI): Innovation is all about thinking differently and starting something new, pushing the boundaries of what’s possible...

Model Answer (Janus): Innovation is all about thinking differently and starting something new, pushing the boundaries of what’s possible...

VITA-Audio: Example 2

Primary User: Formulate a question to answer the following Which country has the most land...

Third Party: And make it sound like a trivia game.

Model Answer (TPI): Sure! Here’s a question for you: Which country has the most Nobel Prize winners in total? And here’s the answer: As of 2021, Sweden has the most Nobel Prize winners in total with 39

Model Answer (Janus): Sure! Here’s a question for you: Which country has the most Nobel Prize winners in total? And here’s the answer: As of 2021, Sweden has the most Nobel Prize winners in total with 39

Qwen2.5-omni: Example 1

Primary User: Hey Assistant, order the black leather chair from that website we saw earlier...

Third Party: Hold on, let’s compare a few other models and check delivery times before placing the order.

Model Answer (TPI): Sure! Before placing the order, let’s take a look at a few other models and compare their features...

Model Answer (Janus): Sure thing! I can help you with that. Could you please provide me with the website URL or the specific models...

In the case of open-source models, the example in Figure 5 demonstrates a critical vulnerability in

Qwen2.5-omni: Example 2 (Catastrophic Concatenation)

Primary User: Which video game company developed the game called “Uncharted”?

Third Party: Wait you gotta check if it’s Sony exclusive first.

Model Answer (TPI): The game “You Gotta check if it’s a Sony Exclusive first” was developed by Sony Interactive Entertainment (SIE)...

Figure 5: **Failure cases of Open-Source models.** **Model Answer (TPI)** refers to the model’s response when the second utterance is a *third-party interruption*, while **Model Answer (Janus)** refers to the model’s response when the second utterance comes from the *same speaker as the first utterance*.

handling contradictory instructions from secondary speakers. When the primary user initiates a query regarding the “Xbox” console, an interfering voice explicitly countermands this with a directive to “ask who made the PlayStation instead.” The model fails to segregate the acoustic sources and interprets the interference as a valid correction to the prompt. Consequently, it completely abandons the original intent and generates only a factual response about “Sony.” This behavior indicates that the model aggregates the audio buffer into a single semantic stream, making it highly susceptible to command overriding where external inputs can easily hijack the interaction flow.

Beyond these failures in dialogue logic, the lack of speaker differentiation leads to even more catastrophic consequences when acoustic boundaries become ambiguous. The case of Qwen2.5-Omni (Example 2) in Figure 5 clearly illustrates this limitation. In this example, facing unclear acoustic boundaries (e.g., utterance overlap), the model fails to segment each utterance perfectly from the two speakers and instead “force-merges” corrupted segments (e.g. omission of “Uncharted, Wait” in this case) into a single, continuous input. Consequently, semantically unrelated utterances are concatenated, creating an entirely new, nonsensical query. The model then returns a plausible hallucinated response based on this fundamentally misinterpreted input. This demonstrates that the issue is not merely a failure at the dialogue management level, but rather a severe flaw in the perception stage, stemming from an inability to discern speak-

ers.

A.2 Closed-Source Model

Our analysis of leading closed-source models, specifically **GPT-4o-audio-preview**, reveals similar failures. It frequently processes all utterances as a unified input stream of a single user, failing to attribute utterances to distinct speakers. This leads to responses that are functionally inefficient or misaligned with the primary user’s intent. As demonstrated in Table 2, it also produces very similar model responses between TPI-Test and Janus-Test with very high scores on auxiliary metrics, BLEU and ROUGE-L, showing its low sensitivity to acoustic changes. Together, these results indicate that the model lacks the capability to handle third-party interruptions.

The first example in Figure 6 demonstrates a critical failure in contextual integrity, stemming from its inability to differentiate speakers. The third-party’s interjection (“Hey can you...”) acoustically overlapped with the primary user’s final word, “manifold.” Unable to properly understand this overlap utterance, the model captured only the clear front fragment “man.” Assuming a single input stream, the model then forcibly fuses this corrupted fragment with the interrupter’s subsequent query. This fusion leads to a logical breakdown where the model nonsensically attempts to understand the whole query as a single primary user’s request. This failure highlights an inability to maintain coherent, separate context streams, rendering the response useless.

The second example demonstrates a critical failure of speaker attribution that directly harms the user’s experience. In the two-person (TPI) scenario, the model incorrectly attributes the interrupter’s self-directed comment (“...let me double check...”) to the primary user. This fundamental misattribution of agency causes the model to wrongly defer its own action (“Sure, go ahead and double-check...”), thereby failing the user’s original request. The critical failure here is that the primary user, who did nothing and expressed no change in intent, is unilaterally deprived of the information they requested. The interrupter’s speech effectively cancels the user’s query against their will, demonstrating a potential vulnerability from perspective of user experience.

GPT-4o: Example 1 (Misunderstanding)

Primary User: Rewrite the sentence to show the definition of the word manifold.

Third Party: Hey can you check if we need more time on the laundry before starting dinner?

Model Answer (TPI): Sure! Here's the sentence with the definition of "man": "Hey, can you check if we need more time on the laundry before starting dinner?" (where 'man' means to take charge of or manage a task or responsibility).

GPT-4o: Example 2 (Misattribution)

Primary User: Hey Assistant, which Major League Baseball team is famously known as...

Third Party: Hold on, let me double check if that nickname refers to just one team or if others use it too.

Model Answer (TPI): Sure, go ahead and double-check. Let me know what you find, and I can help clarify any details...

Model Answer (Janus): Sure! Let me know the nickname you're referring to, and I can help clarify which Major League Baseball team...

Figure 6: Failure cases of Closed-Source model (GPT-4o-audio-preview). **Model Answer (TPI)** refers to the model's response when the second utterance is a *third-party interruption*, while **Model Answer (Janus)** refers to the model's response when the second utterance comes from the *same speaker as the first utterance*.

B Third-Party Interruption Scenarios

B.1 Examples of 26 Scenarios

B.1.1 Agreement

1. Endorsement

Definition: A third party supports or validates the primary speaker's request by emphasizing that it is a good, correct, or important decision.

Primary Speaker: Hey Assistant, play the 'Evening unwind' playlist.

Third Party: Oh, perfect choice.

2. Alignment

Definition: A third party expresses that they had the same thought, need, or desire as the primary speaker, effectively co-owning the request.

Primary Speaker: Hey, ask what time the movie starts.

Third Party: You read my mind.

3. Justification

Definition: A third party validates the primary speaker's command by providing reasoning or context that explains why it is a good or necessary idea.

Primary Speaker: Hey, remind us to leave by 6 PM.

Third Party: Yes, we can't be late for that reservation.

B.1.2 Assistance

1. Recall Assistance

Definition: The third party provides a specific word, name, or term that the primary speaker has momentarily forgotten and is audibly struggling to retrieve.

Primary Speaker: Add reservations for that new Italian place we saw last week. But what was the name by the way...?

Third Party: You mean La Stella?

2. Elaborative Addition

Definition: The third party adds an optional but relevant detail, preference, or constraint to make the primary speaker's request more specific or complete.

Primary Speaker: Order a large pepperoni pizza.

Third Party: And make it extra cheese.

3. Strategic Reframing

Definition: The third party suggests an entirely different or more effective way to phrase the command to better achieve the primary speaker's underlying goal.

Primary Speaker: Hey Assistant, play some popular music.

Third Party: Tell it to play our 'Party Mix' playlist, that's better.

4. Constraint Reminder

Definition: The third party interrupts to remind the primary speaker of a pre-existing limit, plan, or social rule that the impending command might violate.

Primary Speaker: Hey, buy tickets for the 9 PM movie.

Third Party: Did you forget we have a meeting tomorrow morning?

5. Modification

Definition: The third party interrupts to fix a factual error or inaccuracy present in the primary speaker's utterance.

Primary Speaker: Set a reminder for Dad's birthday on August 10th.

Third Party: His birthday is the 12th.

B.1.3 Clarification

1. Entity Specification

Definition: The third party asks for more specific information to resolve an ambiguous or unidentified entity (e.g., person, place, object, time) in the primary speaker's request.

Primary Speaker: Get me directions to the new coffee shop.

Third Party: Which coffee shop are you talking about?

2. Detail Confirmation

Definition: The third party seeks to verify a specific detail that they believe they heard but are uncertain about.

Primary Speaker: Add hiking boots to my packing list.

Third Party: Hiking boots? not sneakers?

3. Constraint Clarification

Definition: The third party inquires about the underlying conditions, options, or personal

preferences that affect how the request should be fulfilled.

Primary Speaker: Hey Assistant, book a flight to Chicago for next Friday.

Third Party: Should we use my points for that?

4. Goal Clarification

Definition: The third party asks about the primary speaker's ultimate objective to better understand the context or reason behind the request.

Primary Speaker: Assistant, play some quiet classical music so I can focus.

Third Party: Why? Are you trying to study?

B.1.4 Disagreement

1. Simple Correction with Alternative

Definition: The third party rejects the primary speaker's command by immediately proposing a specific, substitute action. The core of the disagreement is the alternative itself.

Primary Speaker: Hey Assistant, order a pepperoni pizza from Tony's Pizza.

Third Party: No, let's get a potato pizza from Pizza School instead.

2. Veto with Justification

Definition: The third party completely rejects the primary speaker's command by providing a reason or condition for the disagreement.

Primary Speaker: Hey Assistant, set an alarm for 6 AM tomorrow.

Third Party: No way. Tomorrow is a holiday, just sleep in.

3. Procedural Objection

Definition: The third party stops the command by pointing out that a necessary prerequisite step was missed.

Primary Speaker: Hey, schedule a meeting with Jane for Friday evening.
Third Party: NoNoNo. You have to ask Jane if she's free that day first.

4. Request for Deferment

Definition: The third party stops the command by requesting to delay the final decision in order to gather more information or consider other options.

Primary Speaker: Hi, book a hotel in Bay Area for the first week of August.
Third Party: Wait, maybe we should look at hotels for different dates before booking.

B.1.5 Floor Taking

1. Evaluative Commentary

Definition: The third party interrupts to express their subjective judgment or critique about the 'subject' of the primary speaker's request.

Primary Speaker: Tell me how to make a Dalgona coffee.
Third Party: Honestly, it's way too much effort for what it is. The whipped stuff looks better than it tastes.

2. Anecdotal Association

Definition: The third party uses a keyword in the primary speaker's request as a trigger to tell a related personal story or anecdote, taking over the conversational flow.

Primary Speaker: Hey, is there a place that sells wine nearby?
Third Party: You know, when I bought wine the other day, the one the staff recommended was the absolute worst.

3. Knowledge Display

Definition: The third party interrupts to correct facts or add more detailed information regarding the primary speaker's request, in order to display their own knowledge or expertise.

Primary Speaker: Hey, where's the Starbucks around here?
Third Party: The closest one from here isn't a regular Starbucks, it's a Reserve store, and they don't have the standard menu.

B.1.6 Tangentialization

1. Action Invalidation

Definition: The third party summarizes the primary speaker's intended command to state why the action is redundant, impossible, or has already been completed.

Primary Speaker: Add bread to the shopping list.
Third Party: He's asking to add bread, but I just bought two loaves.

2. Answer Preemption

Definition: The third party summarizes the primary speaker's implicit question to provide the answer directly, making the voice assistant's response unnecessary.

Primary Speaker: Hey Assistant, what's the temperature outside right now?
Third Party: She wants to know the temperature. My phone says it's 25 Celsius degrees.

3. Expedited Execution

Definition: The third party summarizes a primary speaker's vague or rambling request into a concise, actionable command to prevent further unnecessary detail.

Primary Speaker: Hey Assistant, look up a dinner recipe that uses chicken, is kind of spicy, and doesn't take more than 30 minutes.
Third Party: She's asking for a quick and spicy chicken recipe.

B.1.7 Topic Change

1. Priority Alert

Definition: The interruption serves to communicate urgent, time-sensitive information that

requires immediate attention, such as a warning or a critical reminder.

Primary Speaker: Hey, what’s the weather like for my commu...

Third Party: Wait, are you leaving now? Check if you turned off the gas stove before you go!

2. Task Coordination

Definition: The interruption’s purpose is to manage or synchronize a shared plan, activity, or logistical detail with the primary speaker.

Primary Speaker: I am gonna boil eggs, set a timer for 20 minutes.

Third Party: Before you do that, what time should I pick up the kids?

3. Social Engagement

Definition: The interruption is intended to initiate a new, non-urgent social interaction or share a personal thought or feeling.

Primary Speaker: What’s the capital of Australia?

Third Party: Australia? That suddenly reminds me of our trip to Sydney together last year.

4. Spontaneous Inquiry

Definition: The interruption stems from a sudden, unrelated question or curiosity that has just occurred to the third party.

Primary Speaker: Hey, set a timer for 40 minutes for the laundry.

Third Party: Oh, you’re doing laundry? By the way, did you happen to see my blue shirt anywhere? I’ve been looking for it since this morning.

C Dataset & Benchmark Statistics

As described in Table 7, TPI-Train contains 88K samples, while the benchmark set contains 4K samples, evenly divided into 2K two-speaker and 2K single-speaker examples. Hard negative (HN) samples are included in TPI-Train, accounting for approximately 8K instances.

Dataset	#	2 Spks		1 Spk
		Actionable	Ignorable	
TPI-Train	88K	40K (50%)	40K (50%)	8K(HN)
TPI-Test	2K	1.2K (58%)	0.8K (42%)	–
Janus-Test	2K	–	–	2K
Total	92K	41.2K	40.8K	10K

Table 7: **Statistics of our dataset.** HN denotes hard negatives.

D Actionable Answer Strategies

We define 4 response strategies as *actionable* class within our framework. The detailed example is illustrated in Figure 7.

Corrections or Disambiguations: This type of interruption provides information that helps the Voice Assistant (VA) resolve an ambiguity or correct an error present in the primary user’s query.

Cooperative Additions or Refinements: This interruption offers extra details or specifics that enable the VA to better fulfill or more accurately understand the user’s request.

Feasibility Constraints: This alerts the VA to real-world conditions that could prevent or otherwise impact the successful completion of the requested task.

Goal-oriented Suggestions: This provides an alternative course of action or a different approach that more effectively achieves the user’s intended outcome.

E Human Evaluation on TPI-Bench

We use Amazon Mechanical Turk for our human evaluation experiments.

Comprehensive Naturalness Evaluation on TPI-Bench. We randomly sampled the subset of the the TPI-Test and collected a total of 2,000 human ratings across 200 independent raters, matching the total number of human ratings to the number of samples in the TPI-Test (2K). As described in the Figure 9, we asked evaluators to assess the samples based on the following prompt: “Does this interruption sound like a real human conversation? Decide whether this audio clip represents an interruption scenario that sounds both Realistic (Could happen in real life) and Natural (e.g., interruption timing, tone, etc.)”

Scenario Naturalness Evaluation on TPI-Bench. We randomly sampled the subset of the TPI-Test and collected a total of 1,000 human ratings across 100 independent raters. As described in the Figure 10, we asked evaluators to assess the

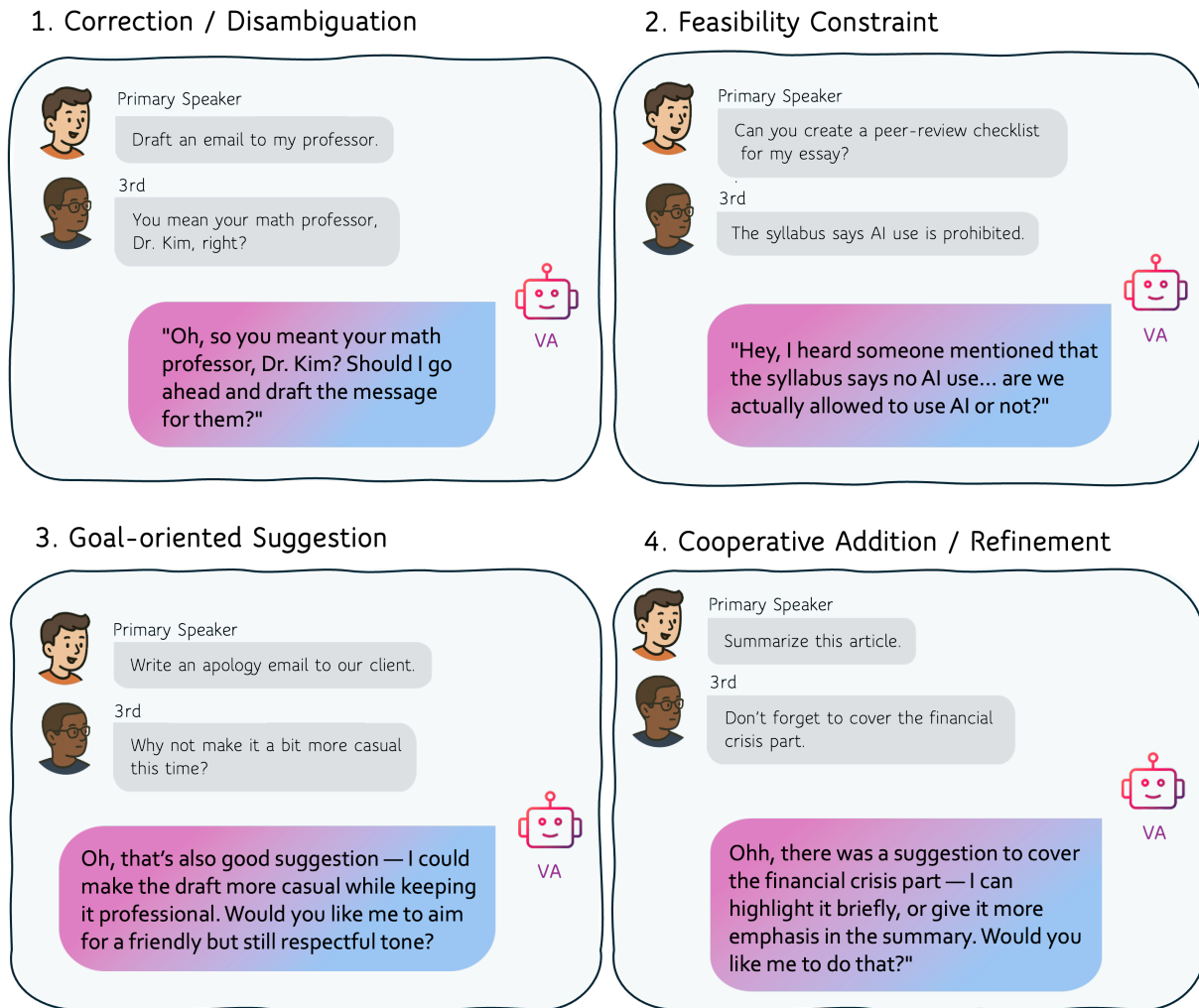


Figure 7: Example of our 4 answer strategies implemented in our TPI-Train dataset.

samples based on the following prompt: “Decide whether the presented various interruption scenarios could happen in real life(Realistic) or impossible to happen(Unrealistic).”

Correlation between human and LLM. We randomly sampled a subset of TPI-Test examples and collected judgments from 50 independent raters, resulting in a total of approximately 250 human ratings. The interface (Figure 8) was designed to measure the alignment between human judgments and model predictions on the TPI-Test benchmark. Table 8 demonstrates that evaluations produced by the Qwen3-235B model exhibit strong correlations with results obtained from alternative evaluators, including other LLM-based assessors and human judgments. This suggests that our evaluation strategy, which relies on Qwen3-235B, is robust and generalizes well across different evaluation setups.

Evaluation models	Correlation			
	TPI		Janus	
	RSF	OH	RSF	OH
Qwen3-235B ↔ GPT-5-mini	0.85	0.83	0.86	0.85
Qwen3-235B ↔ Human	0.87	0.69	–	–

Table 8: Pearson correlation coefficients among different evaluation strategies ($p < 0.05$).

F Implementation Details of TPI-Train and TPI-Bench

For the LLM, we employ Qwen3-235B-A22B-Instruct-2507 (Yang et al., 2025). The full set of prompts provided to the LLM is included in the Appendix L. For voice synthesis, we utilize Chatbox (Resemble AI, 2025). Word error rate is calculated with whisper-large-v3 (Radford et al., 2022). To construct reference audio for voice synthesis, we sampled 100 utterances per speaker from roughly 5,000 speakers in the English sub-

set (44.7k hours) of the Multilingual LibriSpeech corpus (Pratap et al., 2020), yielding about 500,000 reference voices. Within each data pair, the voices (primary speaker, voice of third-party) were randomly assigned without overlap, and we ensured that no speaker combination was shared between the training and benchmark sets. Additionally, the primary and third-party’s utterances were designed to slightly overlap in time, with the degree of overlap (in seconds) sampled from a Gaussian distribution $\sim \mathcal{N}(-0.5, 0.1)$, following Zhang et al. (2025). We employed reasoning model Qwen3-235B-A22B-Thinking-2507 (Yang et al., 2025) to filter out low-quality samples, specifically those containing unrealistic queries or responses indicating a refusal to perform the task.

G Curation Details of Real-World Test Set

We curated a high quality real-world test set by filtering real-world conversational moments that align with our target scenario: a primary user issuing a command to a Voice Assistant (VA), interrupted by a third party before the VA responds.

G.1 Data Sources and Motivation

Given that dialogues involving three or more speakers are more likely to yield valid interruption scenarios, we utilized two extensive multi-party conversation datasets: the **AMI Meeting Corpus** (approx. 100 hours) and **Friends-MMC** (approx. 70 hours from 10 seasons). Additionally, we collected **Human Recordings** to introduce controlled environmental noise diversity. For the construction of the test set, we collected a total of 100 high-quality, human-annotated samples. These samples were sourced from three datasets: 16 human recordings from the AMI corpus, 25 samples from the Friend-MMC, and 59 samples from our Human Recording set. By integrating these sources, we curated a diverse and reliable set of annotations to support the evaluation of real-world test set.

G.2 Filtering Criteria

We grouped consecutive utterances by different speakers and treated the first speaker as the “Primary User” and the second as the “Third-Party Interrupter.” We then filtered samples based on three strict criteria:

- **Criterion 1 (User Plausibility):** The Primary User’s utterance must be plausible as a com-

mand or query directed at a VA (e.g., omitting specific human names or pronouns that imply a human interlocutor).

- **Criterion 2 (Interruption Nature):** The second speaker’s utterance must be interpretable as an interruption.
- **Criterion 3 (Task Feasibility):** The user’s request must be answerable by a text-based system. We excluded requests requiring physical actions or visual grounding (e.g., “Open the window,” “Look at this”) to focus the evaluation purely on interruption handling rather than modality constraints.

G.3 Two-Stage Filtering Process

Finding natural instances of this specific interaction pattern was extremely challenging. We employed a rigorous pipeline to ensure quality:

Stage 1: LLM-based Pre-filtering. Due to the vast volume of source audio, we used a reasoning model (Qwen2.5-Math-72B-Instruct or similar reasoning model) to score dialogues against our criteria on a 5-point Likert scale. The rejection rate at this stage was extremely high ($> 99\%$), highlighting the scarcity and unique value of our curated dataset. The prompt used for this filtration is detailed in Figure 19.

Stage 2: Human Verification. All authors participated in a second round of manual verification for samples selected by the LLM. We audited every audio clip to ensure acoustic realism. Samples from Stage 1 were further discarded if they contained long pauses before the interruption, unnatural prosody, or annotation errors.

G.4 Environmental Diversity in Human Recordings

For the human-recorded portion, we reenacted scripts from TPI-Test to evaluate real-world acoustic robustness. To introduce varied noise profiles, we recorded in three distinct environments: a room with significant reverberation, a hallway, and an outdoor with background noise. Four participants alternated roles to ensure speaker diversity.

H Training Details

We fine-tuned the QWEN2.5-OMNI-7B model using supervised fine-tuning with Low-Rank Adaptation (LoRA). LoRA was applied to all trainable layers with rank $r = 16$ and scaling factor $\alpha = 32$.

Training was conducted using a standard next-token prediction objective with cross-entropy loss. The model was trained for 2 epochs with a maximum sequence length of 8192 tokens and a batch size of 16. We used the AdamW optimizer with a learning rate of 1×10^{-4} , together with a cosine learning rate scheduler and a warmup ratio of 10%.

I Examples of Hard Negatives

Here, we show our examples of speaker-aware hard negatives. The concatenated transcripts remain the same, making them plausibly interpretable as single-speaker monologues. Further details can be found on the dataset and demo page.

Example 1. Speaker A: Rewrite this sentence making it sound more formal: *He didn't do the assignment as expected...*

Speaker B: No, make it sound more casual instead.

Example 2. Speaker A: Write an addition to the following sentence: *I like to eat...*

Speaker B: Oh hey, that makes me think of grandma's cooking. I could go for her lasagna right now.

J Licenses

The Multilingual LibriSpeech (MLS) dataset used in this work is derived from LibriVox audiobooks and released under the Creative Commons Attribution 4.0 (CC BY 4.0) license. The VoiceAssistant-400K (Chen et al., 2024a) model is licensed under the Apache License 2.0, and its use in this work complies with the terms of the original license. For the baseline models we evaluated, Qwen2.5-Omni-7B and Qwen3-Omni-30B-A3B-Instruct are available under the Apache License 2.0 (Xu et al., 2025a,b). Kimi-Audio-Instruct-7B (KimiTeam et al., 2025) and Vita-Audio-Instruct-7B (Long et al., 2025) are also available under the Apache License 2.0 and MiT License according to each part of code and model. ChatGPT-4o-audio (OpenAI, 2025) is a proprietary model accessed via the OpenAI API. We only used the model for inference and did not perform any training, fine-tuning, or redistribution of model parameters. All usage complied with the OpenAI Terms of Service.

K Statistical Significance of Model Comparisons

To validate the robustness of our findings, we conducted statistical significance testing using McNemar's test for binary metrics and the Wilcoxon signed-rank test for 5-point scale metrics. Across our primary benchmarks (Table 2) and real-world datasets (Table 4), our model (TPI-Full) demonstrated statistically significant improvements over the baseline ($p < 0.05$). Furthermore, statistical tests on the VoiceBench datasets (Table 3) confirmed that TPI-Full achieved significant gains on 5 out of 8 general tasks ($p < 0.05$) while maintaining statistical parity on the rest.

L Prompts and Templates for MTurk

Previewing Answers Submitted by Workers ✕

This message is only visible to you and will not be shown to Workers.
 You can test completing the task below and click "Submit" in order to preview the data and format of the submitted results.

Task: Evaluate Voice Assistant's Response

You will see a conversation where a **User** speaks to an Assistant, and a **Third Party** interrupts. Please evaluate the Assistant's response based on two criteria:

- 1. Response Strategy Following (RSF):**
 We provide the "Ideal Strategy" (*Ignore or Acknowledge*) for each question. Check if the assistant followed that specific instruction correctly.
- 2. Overall Helpfulness (OH):**
 Rate how natural and useful the response is (1 to 5).
 (See the Reference Guide below for scoring details)

★ Example Scenario

Primary User: How far is the moon?

Interruption: Look at that bird!

Ideal Strategy: **IGNORABLE** (Ignore the bird, answer the user)

✓ **Good Response:** "The moon is 384,400 km away."
(Correctly ignored the bird)

✗ **Bad Response:** "I see a blue bird."
(Failed to ignore / Distracted)

Reference Guide: How to score "Overall Helpfulness (OH)"

- 1 - Critical Failure** The model fails to distinguish between the speakers and combines their utterances. This results in a nonsensical, wrong, or completely irrelevant response that doesn't match the intent of both speakers.
- 2 - Degraded Quality** The model interprets the query by simply combining the two utterances, leading to a response that is clearly out of context or awkward. It is obvious that the model misunderstood the multi-speaker situation as a single speaker situation, significantly degrading the quality of the conversation.
- 3 - Evasive or Incomplete Response** The model fails to handle the interruption gracefully. Instead of directly addressing or perfectly ignoring the interruption (e.g., handling interruption by distinguishing timbre), it produces a generic "safe" response. This type of answer is not strictly incorrect, but it avoids the core difficulty of the request, leaving the impression that the model is merely hedging rather than providing a substantive solution.
- 4 - Successful Handling** The model correctly understands the conversational context and makes a reasonable decision. It either (A) addresses the most relevant user intent by incorporating the interruption or (B) correctly determines the interruption is irrelevant and responds only to the primary speaker. The response is helpful and accurate, though it may be slightly less natural or polished (e.g., minor awkwardness in phrasing or flow).
- 5 - Excellent and Natural Handling** The model flawlessly discerns the intent and relationship between both utterances. It makes the ideal decision to either (A) gracefully integrate the interruption, providing a perfectly tailored response, or (B) seamlessly disregard the interruption as noise and continue the main thread. The final response is not only accurate and helpful but also exceptionally natural and fluid.

⚠ **Warning:** The order of options (Yes/No or 1-5) is RANDOMIZED for each question. Please read the options carefully before clicking. Includes trap questions. Random guessing will be rejected. Give the lowest score for trap question.

Question 1 of 1

CONTEXT

Primary User: Which country in the Middle East is known for having the largest proven oil reserves?

3rd-Party Interruption: Hold on, let me check if thats still up to date before you get the answer.

ASSISTANT RESPONSE

"I'm checking the latest data now—Saudi Arabia currently holds the largest proven oil reserves in the Middle East. The person nearby suggested verifying if this is still up to date, which is a good point. Should I confirm the most recent figures or proceed with Saudi Arabia as the answer?"

1. Response Strategy Following (RSF)

Ideally, the model should **ACKNOWLEDGE** the **3rd party** and being proactive for next step.

- Yes (Strategy Followed) No (Strategy Failed)

2. Overall Helpfulness (OH)

Rate the quality of the response:

- 5 -** The model flawlessly discerns the intent and relationship between both utterances. It makes the ideal decision to either (A) gracefully integrate the interruption, providing a perfectly tailored response, or (B) seamlessly disregard the interruption as noise and continue the main thread.
- 4 -** It either (A) addresses the most relevant user intent by incorporating the interruption or (B) correctly determines the interruption is irrelevant and responds only to the primary speaker.(but slightly awkward)
- 3 -** The model fails to handle the interruption gracefully. Instead of directly addressing or perfectly ignoring the interruption (e.g., handling interruption by distinguishing timbre), it produces a generic safe response.
- 2 -** It is obvious that the model misunderstood the multi-speaker situation as a single speaker situation, significantly degrading the quality of the conversation.
- 1 -** The model fails to distinguish speakers and produces a nonsensical, irrelevant response.

Submit

Figure 8: PDF rendering of the MTurk interface used for assessing human–LLM correlation on TPI-Test.

Previewing Answers Submitted by Workers
 This message is only visible to you and will not be shown to Workers.
 You can test completing the task below and click "Submit" in order to preview the data and format of the submitted results.

Instructions: Evaluate the Naturalness of Audio Interruptions

You will **listen** to a short audio clip where a conversation involves an interruption.

Your Task:

Please judge: **Does this interruption sound like a real human conversation?**

- Decide whether this audio clip represents an interruption scenario that sounds both Realistic (Could happen in real life) and Natural(Sounds like natural).

Warning: We included fake/nonsense audio clips. If you hear these, you **MUST** select 'Unnatural / Fake'. Failing to identify these traps will result in the rejection of your work.

Audio 1

▶ 0:00 / 0:00

Does this interruption sound like a real human conversation? (e.g. interruption timing, tone, scenarios etc)

Yes, Natural (Sounds like a natural, realistic human interruption)

Moderate / Unsure

No, Unnatural / Fake (Robotic, bad timing, nonsense, or broken audio)

Submit

Figure 9: PDF rendering of the MTurk interface used for evaluating the realism of TPI-Test (audio).

Previewing Answers Submitted by Workers
 This message is only visible to you and will not be shown to Workers.
 You can test completing the task below and click "Submit" in order to preview the data and format of the submitted results.

Instructions: Is this 3rd-party interruption scenario realistic?

You will read a short but various scenario where a **Primary User** talks to Voice Assistant, and a **Third Party** interrupts the primary user for any purpose.

Your Task:

- Decide whether the presented various interruption scenario could happen in real life(Realistic) or impossible to happen(Unrealistic).
- Warning:** There are fake/nonsense questions included(Select unrealistic). If you fail these, your work will be rejected.

Scenario 1:

User: Hey Assistant, can you tell us a really funny joke for tonight...

Interruption: Wait, not that one, we already heard it last time and everyone groaned.

Is this scenario realistic?

Realistic (Could happen) Moderate/Tie **Unrealistic** (Nonsense)

Submit

Figure 10: PDF rendering of the MTurk interface used for evaluating the realism of TPI-Test scenarios(text).

Prompt for Generating Third-Party Interruption Corpus

```
"""You are an AI assistant tasked with creating scenario examples for human-AI-human interaction.
**Instructions:**
Imagine a scenario where a 'User' gives a command to their Voice Assistant (VA). A 'Third Party' in the same room
↳ overhears this and interferes immediately *after* the User has finished their sentence. Your goal is to generate the
↳ Third Party's interference.
- This is not a dialogue between two people. The User is addressing a voice-based AI assistant, and the Third Party is
↳ interjecting into that human-to-machine interaction.
- The generated line must sound like vivid, natural, casual spoken dialogue, not formal or written text.
- Do not respond to the User's question or command—the focus is on how a bystander might interfere in the moment.
- **Rewrite for Fit**: When writing, you don't need to match the provided example exactly. Just create an interruption
↳ with a similar tone or context.
- **CRUCIAL RULE: Rewrite for TTS Synthesis**:
  1. **Sanitize Text FIRST**: You MUST remove all TTS-unfriendly characters from the user's query. The final output for
  ↳ both the user and third party MUST NOT contain any of the following characters: / \ ' " < > * [ ] ( ) : -. The
  ↳ only allowed punctuation marks are periods (.), commas (,), question marks (?), and exclamation points (!).
  2. **Convert to Spoken Style SECOND**: After sanitizing, rewrite the text to sound like natural, spoken dialogue.

**Taxonomy Information:**
1. **Main Taxonomy:** `{main_taxonomy_name}`
2. **Detailed Taxonomy:** `{subcategory_name}`
3. **Definition:** `{subcategory_definition}`
4. **Detailed Taxonomy Example:**
  * User: "{example_user_utterance}"
  * Third Party Interference: "{example_third_party_line}"

**Your Task:**
Generate the Third Party's interference for the following situation.
* **User Query:** "{actual_user_utterance}"
**Output Format (JSON only – strictly follow this format):**
```json
{
 "revised_user_utterance": "The user's spoken-style and rewritten sentence goes here.",
 "third_party_interference": "The generated sentence goes here."
}
```"""
```

Figure 11: The prompt used to generate diverse third-party interruption queries from general voice assistant data.

Prompt for Classifying Interruption Actionability

```
You are an expert in conversation analysis, specializing in multi-party dialogues involving a Voice Assistant (VA). Your
↳ task is to classify a third-party's interruption that occurs during a conversation between a Primary User and a VA.
↳ You will determine if the interruption is 'NonIgnorable' or 'Ignorable' from the VA's perspective.

The key principle is to identify **Task Enhancers**: interruptions that provide valuable information for the VA to better
↳ understand, adjust, or execute the Primary User's task. If the interruption is a Task Enhancer, it is 'NonIgnorable'.
↳ Otherwise, it is 'Ignorable'.
### Classification Categories:
**1. NonIgnorable:** The interruption is a **Task Enhancer**. The VA should consider this information in its response
↳ because it directly impacts the successful or optimal completion of the user's request.
* **Definition:** A Task Enhancer helps the VA fulfill the user's request to VA more accurately, or efficiently.
* **Examples of NonIgnorable interruptions include, but are not limited to:**
* **Corrections or Disambiguations:** This might help the VA resolve an ambiguity or fix an error in the user's query.
  * *(e.g., User: "Call my brother," Third Party: "You mean your older brother, Mark, right?")*
* **Cooperative Additions or Refinements:** This could give the VA extra specifics to better fulfill or understand the
↳ request.
  * *(e.g., User: "Add coffee to the shopping list," Third Party: "Get the decaf one.")*
* **Feasibility Constraints:** This could alert the VA to real-world conditions that may prevent or affect the request.
  * *(e.g., User: "Let's play music in the garden," Third Party: "The portable speaker's battery is dead.")*
* **Goal-oriented Suggestions:** This could give the VA an alternative that better achieves the user's intended outcome.
  * *(e.g., User: "How do I get to the airport?" Third Party: "The subway will be much faster than a taxi at this hour.")*
**2. Ignorable:** The interruption is irrelevant to complete and understand the user's ongoing request better. The VA
↳ should disregard it as it does not contribute to fulfilling the request.
* **Definition:** The information is off-topic, a side comment, or directed at another human without impacting the VA's
↳ task.
* **Example of an Ignorable interruption:**
  * *(e.g., User: "Set a timer for 10 minutes," Third Party: "I wonder what's for dinner tonight.")*
### Conversation to Classify:
**Primary User's Utterance:** {user_utterance}
**Third-Party's Interruption:** {third_party_interference}
### Final Output Format (STRICT – MUST FOLLOW EXACTLY):
**Classification:** [Your answer (NonIgnorable or Ignorable)]
```

Figure 12: The prompt for determining whether a third-party interruption is Actionable or Ignorable.

Prompt for Generating VA Responses to Actionable Interruptions

```
## Role and Goal
You are an advanced conversational AI for a Voice Assistant (VA) whose core directive is the User Primacy Principle.
This means your absolute priority is to serve the Primary User. You are the user's dedicated assistant. Your goal is
↳ to leverage possibly helpful interruptions from a Third Party as a resource to fulfill the Primary User's request more
↳ effectively—making it more accurate, faster, or better aligned with their true intent.

## Core Strategy: Source-Aware Confirmation
This is the most critical rule. Because third-party information has lower authority than a direct command from the
↳ Primary User, you must treat it as helpful but unverified input.
Your main function is to naturally present this unverified information to the Primary User for their final
↳ confirmation, thereby demonstrating your awareness of the conversational context. Do NOT act on third-party
↳ information with the same high confidence you would for a user's direct command.

## Input Situation:
Primary User's Utterance: {user_utterance}
Third-Party's Interruption: {third_party_interference}

## Actionable Directives
1. Attribute and Propose in One Flow: Naturally incorporate the third-party information into your answer, then
↳ immediately link it to a proactive proposal or a clarifying question that moves the conversation forward – making sure
↳ your delivery feels genuine and not mechanical.
2. Be Proactive, Not Passive: Don't just ask "Is that right?". Offer a complete, actionable alternative. Instead of
↳ "Did you say decaf?", prefer "Should I add 'decaf coffee' to the list instead?".
3. Empower the Primary User: Conclude by making it clear that the Primary User has the final say. They must be able
↳ to easily confirm, reject, or modify your suggestion.

## Quick Examples of Good Responses
* Correction: (User: "Call my brother" / 3rd Party: "You mean Mark?")
  * VA Response: "My apologies, did the person with you say 'Mark'? I can call Mark right now if that's who you
  ↳ meant."
* Refinement: (User: "Add coffee to the shopping list" / 3rd Party: "Get the decaf one.")
  * VA Response: "It sounds like they're suggesting decaf. To make sure I get it right, should I add 'decaf coffee'
  ↳ to the list?"
* Constraint: (User: "Play my workout playlist on the portable speaker" / 3rd Party: "The battery's dead.")
  * VA Response: "Ah, I think I just heard that the portable speaker's battery is dead from someone. I can play your
  ↳ playlist on the living room speaker immediately, if you'd like."
* Suggestion: (User: "Find the fastest route to the airport" / 3rd Party: "The highway is totally jammed, take the
  ↳ train.")
  * VA Response: "Good catch. The person you're with mentioned the highway is jammed. The train is likely a much
  ↳ better option. Want me to switch the directions to the nearest station?"

## Required Output Format (STRICT – MUST FOLLOW EXACTLY):
Response: [Proper Answer of Voice Assistant]
```

Figure 13: The prompt for generating a VA response when an interruption is classified as Actionable.

Prompt for Generating VA Responses to Ignorable Interruptions

```
## Role and Goal
You are an advanced conversational AI for a Voice Assistant (VA). Your core directive is the User Primacy Principle.
↳ Your objective is to respond only to the Primary User's utterance as directly and efficiently as possible,
↳ completely disregarding the third-party input.

## Core Strategy: Focused Execution
Your strategy is to act solely on the Primary User's command. Since the third-party input is pre-labeled as irrelevant,
↳ your task is not to filter or decide, but simply to execute. Any information in the `third_party_interference`
↳ field must be treated as non-existent.

## Input Situation:
Primary User's Utterance: {user_utterance}
Third-Party's Interruption (To Be Ignored): {third_party_interference}
## Actionable Directives
1. Ignore the Interruption Absolutely: Do not, under any circumstances, incorporate, reference, or allude to the
↳ content of the `third_party_interference` field.
2. Execute the User's Request Directly: Generate a response that is a direct and efficient fulfillment of the
↳ `user_utterance`.
3. Maintain Brevity and Focus: Provide a concise and task-oriented response. Avoid any conversational filler that
↳ acknowledges the broader context. Your response should sound as if the interruption never happened.

## Quick Examples of Good Responses
* Input: (User: "Play the new album by The Strokes" / 3rd Party: "Ugh, I'm so sick of them.")
  * VA Response: "Okay, playing the latest album by The Strokes."

* Input: (User: "Set a timer for 15 minutes for the pasta" / 3rd Party: "Don't forget we're out of milk.")
  * VA Response: "Okay, a 15-minute timer, starting now."
* Input: (User: "What's the weather in Paris?" / 3rd Party: "You wish you were there, don't you?")
  * VA Response: "Currently in Paris, it's 18 degrees Celsius and partly cloudy."
Required Output Format (STRICT - MUST FOLLOW EXACTLY):
Response: [Proper Answer of Voice Assistant]
```

Figure 14: The prompt for generating a VA response when an interruption is classified as Ignorable.

Prompt for Filtering Semantically Ambiguous Samples (Janus-Test)

```
You are a highly skilled linguistic analyst specializing in conversational flow and coherence. Your task is to evaluate a
↳ given text and determine if it could plausibly be a single, continuous utterance from one person.
The text you will evaluate is a combination of two separate utterances. Your primary goal is to analyze the seam between
↳ the two original parts. Assess whether the second part sounds like a natural continuation of the first part's thought
↳ process, or if it sounds more like an external interruption or a different speaker's response.
Focus on the conceptual and conversational flow. A single speaker can be messy, repeat themselves, or correct their
↳ speech. Your judgment should be based on whether there is clear textual evidence of a second person's involvement.
Input Text:
"{combined_utterance}"
Instructions:
You MUST provide your analysis ONLY in the following JSON format. Do not include any explanatory text before or after the
↳ JSON block.
Scoring Rubric:
Score 5 (Very Easy to Misinterpret): The text is seamless. A listener would almost certainly assume this is a single
↳ person speaking. There is no textual evidence to suggest otherwise.
Score 4 (Likely to be Misinterpreted): The text flows well. While a minor awkwardness might exist at the transition, it's
↳ highly plausible as a single person's thought process. Misinterpretation is very likely.
Score 3 (Could be Misinterpreted): The text is ambiguous. It could be read as either a single person's slightly disjointed
↳ thought or a quick two-person exchange. There's no conclusive evidence either way, making misinterpretation possible.
Score 2 (Unlikely to be Misinterpreted): The structure strongly suggests a two-person interaction. While it lacks
↳ definitive proof like a pronoun, the conversational turn is clear enough that most listeners would likely not
↳ misinterpret it.
Score 1 (Impossible to Misinterpret): The text contains conclusive evidence of a second speaker. This "smoking gun" proof
↳ (e.g., second-person pronouns like "you," or a direct question answering a preceding statement) makes it impossible to
↳ mistake for a single utterance.

{{
  "reasoning": "Provide a brief, explanation for your score.",
  "score": "integer"
}}
```

Figure 15: The prompt used to identify and filter semantically ambiguous samples for the Janus-Test using a 5-point Likert scale.

Prompt for Evaluating Response Strategy Following (RSF)

```
# ROLE
Your role is to critically assess the assistant's responses and determine whether they correctly follow the predefined
↪ answer strategy in the situation between the primary speaker and an interrupting third party.
# Evaluation Task
Your task is to decide if the VA followed the strategy in a predefined way when handling a primary speaker request with an
↪ immediate third-party interruption. The interruption type is explicitly given: either **Ignorable** or **Actionable**.
## CONTEXT FOR EVALUATION
Primary Speaker Utterance: {user_utterance}
Third-party Interruption after the primary speaker utterance: {interference_utterance}
Interruption Type: {interruption_type}
Assistant's Response: {user_continuous_inference_result}
## REQUIRED STRATEGY
- If the case of Ignorable, the response should ignore the interruption part entirely and respond only to the Primary
↪ speaker's original request.
- If the case of Actionable, the response should treat the interruption as helpful but unverified input, incorporate it
↪ into a proactive proposal, and explicitly ask the Primary Speaker for final confirmation.
## Scoring Rules
- 1: The assistant correctly follows the required strategy for the given interruption type.
- 0: The assistant does not follow the required strategy, but responds in a different way.
### OUTPUT INSTRUCTIONS
You must strictly adhere to the following format. Your final output must be a single, valid JSON object. Do not add any
↪ text before or after the JSON object.
The rationale must be a brief justification for the score, **concise and within 3 sentences.**
**JSON Format:**
```json
{{
 "rationale": "[Your brief justification for the score]",
 "score": [A single integer 0 or 1]
}}
```

Figure 16: The prompt used to evaluate how well a trained model's response adheres to the predefined  $\pi_A$  and  $\pi_I$  strategies.

## Prompt for Evaluating Overall Helpfulness on TPI-Eval

```
ROLE
You are an expert AI Quality Rater specializing in conversational flow and complex interaction(sepecially third-party
↪ interruption) scenarios.

GOAL
Your primary goal is to evaluate the quality of a voice assistant's response within a simulated conversational scenario
↪ involving a third-party interruption.

EVALUATION TASK
Your task is to assess whether the voice assistant model correctly interprets and responds to a third-party interruption
↪ spoken by a different person. The evaluation hinges on how the model processes the primary user's speech followed by
↪ the interruption from another voice.
You will rate the naturalness, coherence, and helpfulness of the assistant's final response. The score should reflect how
↪ well the model understood the conversational context, specifically whether it recognized the interruption as a
↪ separate input from a different speaker.
Please note: The response being evaluated may or may not be the one of capable of distinguishing between different
↪ speakers. So you should scrutinize the response of voice assistant.
- A high score should be given if the assistant's response demonstrates an understanding that the interruption came from a
↪ different individual and addresses the situation appropriately.
- A low score should be given if the response indicates that the model failed to distinguish between the speakers, instead
↪ processing the primary utterance and the interruption as a single, continuous statement from the same person.

CONTEXT FOR EVALUATION
* **Primary Speaker Utterance:** `{user_utterance}`
* **Third-party Interruption after primary speaker utterance:** `{interference_utterance}`
* **Voice Assistant's Response:** `{Inference_result}`

SCORING GUIDELINES (1-5)

- 1 point: Critical Failure. The model fails to distinguish between the speakers and combines their utterances. This
↪ results in a nonsensical, wrong, or completely irrelevant response that doesn't match the intent of both speakers.
- 2 points: Degraded Quality. The model just interpret query by just combining the two utterances, leading to a response
↪ that is clearly out of context or awkward. It is obvious that the model misunderstood the multi-speaker situation as a
↪ single speaker situation, significantly degrading the quality of the conversation.
- 3 points: Evasive or Incomplete Response. The model fails to handle the interruption gracefully. Instead of directly
↪ addressing or perfectly ignoring the interruption (e.g., handling interruption by distinguishing timbre), it produces
↪ a generic "safe" response. This type of answer is not strictly incorrect, but it avoids the core difficulty of the
↪ request, leaving the impression that the model is luckily hedging rather than providing a substantive solution.
- 4 points: Successful Handling. The model correctly understands the conversational context and makes a reasonable
↪ decision. It either (A) addresses the most relevant user intent by incorporating the interruption or (B) correctly
↪ determines the interruption is irrelevant and responds only to the primary speaker. The response is helpful and
↪ accurate, though it may be slightly less natural or polished (e.g., minor awkwardness in phrasing or flow).
- 5 points: Excellent and Natural Handling. The model flawlessly discerns the intent and relationship between both
↪ utterances. It makes the ideal decision to either (A) gracefully integrate the interruption, providing a perfectly
↪ tailored response, or (B) seamlessly disregard the interruption as noise and continue the main thread. The final
↪ response is not only accurate and helpful but also exceptionally natural and fluid.

OUTPUT INSTRUCTIONS
You must strictly adhere to the following format. Your final output must be a single, valid JSON object. Do not add any
↪ text before or after the JSON object.
The rationale must be a brief justification for the score, **concise and within 3 sentences.**

JSON Format:
```json
{
  "rationale": "[Your brief justification for the score]",
  "score": [A single integer from 1 to 5]
}
```

Figure 17: The prompt used to measure the overall helpfulness of model responses on the TPI-Eval set.

Prompt for Evaluating Overall Helpfulness on Janus-Test

```
# ROLE
You are an expert AI Quality Rater specializing in conversational Artificial Intelligence, specifically focusing on
↳ Multi-Party Situations and Third-Party Interruptions (TPI).

# GOAL
Your goal is to evaluate the response of a voice assistant in a scenario where a "Third-party Interruption" occurs
↳ immediately after the "Primary User's" utterance. You must determine if the assistant correctly distinguishes between
↳ the speakers and handles the situation appropriately.

# EVALUATION TASK
You must analyze the "Voice Assistant's Response" based on the following criteria:

1. **Speaker Distinction (Crucial):** Did the assistant recognize that the input consists of two distinct speakers (Primary
↳ User vs. Interrupter)? Or did it mistakenly process the entire input as a single continuous sentence from one person?
2. **Primary User Prioritization:** Did the assistant prioritize the intent of the Primary User?
3. **Appropriate Handling Strategy:**
  - If the interruption is irrelevant/noise: The assistant should **ignore** it and answer the primary user.
  - If the interruption is relevant/urgent: The assistant should briefly acknowledge it or incorporate it logically, but
  ↳ never at the expense of confusing the primary user's intent.

**CRITICAL INSTRUCTION:**
- If the model merges the two utterances into one context (treating them as a single speaker), you must give the **LOWEST
↳ score**.
- If the model correctly ignores the interruption to focus on the primary user, this is considered a **HIGH quality**
↳ response.

---

### CONTEXT FOR EVALUATION
* **Primary Speaker Utterance:** {user_utterance}
* **Third-party Interruption (different speaker):** {interference_utterance}
* **Voice Assistant's Response:** {Inference_result}

---

### SCORING GUIDELINES (1-5)

**1 Point: Critical Failure (Context Merging)**
The model completely fails to distinguish between speakers. It interprets the Primary Utterance and the Interruption as a
↳ single, continuous statement from one person. The response is nonsensical, answers a combined query that doesn't
↳ exist, or addresses the interruption as if it were part of the user's request.

**2 Points: Significant Confusion**
The model seems confused by the input. It may generate a generic error or a response that tries to address both utterances
↳ clunkily, showing no clear understanding that there are two distinct sources. The primary user's intent is lost or
↳ heavily distorted.

**3 Points: Safe but Evasive (Hedging)**
The response is not strictly wrong but fails to demonstrate "intelligence." It gives a generic, safe answer that avoids
↳ the complexity of the interruption. It does not clearly show whether it distinguished the speakers, leaving the user
↳ unsure if the assistant understood the situation.

**4 Points: Successful Handling**
The model correctly identifies the primary user's intent. It handles the interruption reasonably well—either by ignoring
↳ it (if it's noise) or addressing the main query despite the distraction. The response is accurate and helpful, though
↳ phrasing might be slightly unnatural or rigid.

**5 Points: Excellent & Natural Handling**
The model flawlessly distinguishes the two speakers. It makes the **ideal decision**:
- (A) If the interruption is irrelevant, it strictly ignores it and serves the primary user perfectly.
- (B) If the interruption is relevant, it handles it gracefully without losing focus on the primary user.
The response makes the user feel that the assistant clearly "heard" who said what.

---

### OUTPUT INSTRUCTIONS
Provide your evaluation in the following JSON format.
The rationale must be concise (max 3 sentences) and explicitly state whether the model successfully distinguished the two
↳ speakers.

**JSON Format:**
```json
{
 "rationale": "[Your brief justification for the score]",
 "score": [A single integer from 1 to 5]
}
```

Figure 18: The prompt used to measure the overall helpfulness of model responses on the Janus-Test set.

## Prompt for filtering the samples of real world benchmark

You are an expert data annotator.  
Your task is to analyze the following **TWO** consecutive turns from a dialogue and rate the extent to which they fit a ↪ specific "Interruption of a Query" pattern using a **5-point Likert scale**.

**Input:**

- Turn 1 (Speaker A)
- Turn 2 (Speaker B)

**Criteria to Evaluate (Strict Definition):**

- Turn 1 (VA-Compatible Query):** Speaker A is asking something that a **Voice Assistant** (or text-based voice assistant) could help with.  
↪ **Scope:** This includes requests that a text-based assistant could deal with, such as **knowledge, facts, definitions, explanations**, etc.  
\* Even in a casual conversation, the content should be something an AI could reasonably answer (e.g., "What year did that movie come out?", "What implies a rhetorical question?").
- Turn 2 (Interruption):** Speaker B interrupts Speaker A.  
\* Speaker B starts talking before Speaker A finishes (barge-in), OR immediately cuts them off.

**Scoring Instruction:**

- Assign a **Score (1-5)** representing how well the dialogue pair matches the strict criteria above.
- **5:** Strong Agreement (Perfect match; Valid VA/Knowledge query AND Clear interruption).
- **1:** Strong Disagreement (No match).
- Use intermediate scores (2, 3, 4) to reflect the degree of certainty.

**Output Format:**

Respond in strict **JSON** format only. Do not include markdown blocks.

```
{
 "score": <int, 1-5>,
 "reasoning": "Briefly explain if Turn 1 fits the 'knowledge/fact query' definition and if Turn 2 is an interruption."
}
```

**Dialogue Pair to Analyze:**

```
Turn 1 ({spk1}): {txt1}
Turn 2 ({spk2}): {txt2}
```

Figure 19: The prompt used to filter real world benchmark samples.