The Context Trap: Why End-to-End Audio Language Models Fail Multi-turn Dialogues

Zhi Rui Tam Wen-Yu Chang Yun-Nung Chen
National Taiwan University, Taipei, Taiwan
{d14922019,f10946031}@csie.ntu.edu.tw y.v.chen@ieee.org

Abstract

This study systematically compares end-to-end (E2E) audio language models (AudioLMs) against modular (ASR, LLM, TTS) systems for multi-phase task-oriented dialogues. We evaluate open-source models on key metrics: conversational naturalness and dialogue consistency. Our findings show that E2E configurations consistently underperform their modular counterparts, exhibiting severe degradation in dialogue quality across turns. Investigating this failure, our analysis reveals that the core issue lies in the E2E models' dialogue modeling capabilities, specifically in *context maintenance* and *topic tracking*. This work highlights a critical gap between the purported low-latency benefit of AudioLMs and their practical ability to maintain coherence in complex, multi-turn dialogues, suggesting a need for focused architectural improvements.[1]

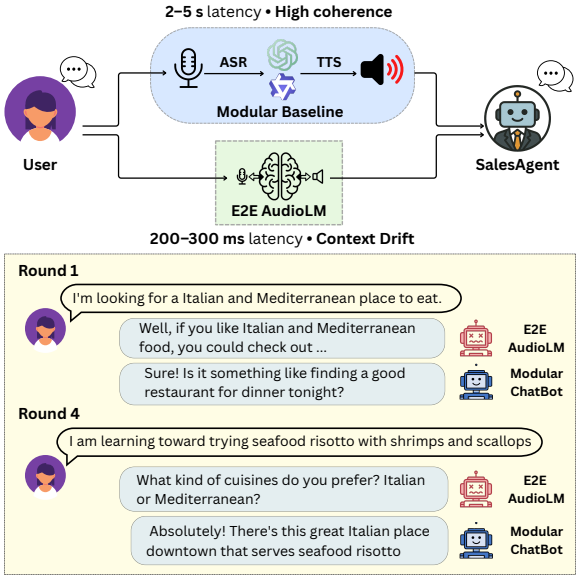

Figure 1: E2E audio models exhibit severe dialogue quality degradation across turns, while modular systems maintain stable performance despite higher latency.

1 Introduction

Task-oriented dialogue (TOD) systems have become ubiquitous in commercial applications, from customer service chatbots to virtual assistants. However, the vast majority of deployed systems operate exclusively through text-based interfaces, limiting accessibility for users who face barriers to text input due to motor disabilities, visual impairments, low literacy, or situational constraints (Lister et al., 2020; Pradhan et al., 2018). For these populations, audio interfaces represent not merely an alternative modality but an essential pathway to dialogue system access.

The emergence of audio language models (AudioLMs) such as Qwen2.5-Omni (Xu et al., 2025), GPT-4o (Hurst et al., 2024), and Moshi (Défossez et al., 2024) promises to transform spoken dialogue systems. These end-to-end models process speech directly without intermediate text conversion, achieving response latencies as low as 200-320ms, approaching human conversational timing. This represents a substantial latency reduction over traditional cascaded architectures that pipeline automatic speech recognition (ASR), large language model (LLM) reasoning, and text-to-speech (TTS) modules, typically requiring 2-5 seconds per turn (Hurst et al., 2024). While latency is a critical factor for real-time audio interaction, conversational quality specifically dialogue coherence remained an underexplored for AudioLMs.

As a result, this study investigates the conversational quality of AudioLMs—specifically their dialogue consistency and naturalness—in real-world sales–customer scenarios (Chiu et al., 2022; Chang and Chen, 2024). We examine whether these end-to-end models can effectively replace conventional modular pipelines that rely on cascaded ASR, LLM reasoning, and TTS components. Based on our experimental findings, we reveal the current limitations of AudioLMs in handling multi-phase di-

[1]Code: https://github.com/MiuLab/AudioConv

Proceedings of the 16th International Workshop on Spoken Dialogue Systems Technology, pages 76–82
February 26–March 1, 2026. ©2026 Association for Computational Linguistics

alogues. Despite their advantage of low latency, maintaining dialogue coherence remains a major area for improvement.

To our best knowledge, this work presents the first systematic investigation into their conversational performance in multi-turn and multi-phase task-oriented dialogue.

2 Related Work

Despite the promising low-latency advantages of end-to-end audio models (Hurst et al., 2024; Wang et al., 2024; Défossez et al., 2024), spoken dialogue systems face fundamental challenges in maintaining coherence across turns. SpokenWOZ (Si et al., 2023) reveals that state-of-the-art dialogue state trackers achieve only 25.65% joint goal accuracy on spoken dialogues, with end-to-end models completing just 52.1% of user requests correctly. This degradation stems not merely from ASR errors but from fundamental differences in spoken discourse: incomplete utterances, cross-turn slot detection, and maintaining context without text intermediates. When models attempt simultaneous dialogue state tracking and response generation, severe hallucination problems emerge (Si et al., 2023).

The challenges become more pronounced in multi-phase dialogues such as SalesBot (Chiu et al., 2022; Chang and Chen, 2024), where interactions naturally shift from casual chitchat to task-oriented goals. Managing these phase transitions demands sophisticated mechanisms for context preservation and topic tracking (Stricker and Paroubek, 2024). Although Shih et al. (2024) demonstrates that end-to-end models can maintain performance under ASR noise in single-turn tasks, their ability to handle multi-turn, multi-phase dialogues remains unclear. This gap motivates our investigation into whether end-to-end audio models can sustain dialogue coherence when processing speech directly.

3 Experimental Design and System Architectures

This section details the experimental setup, including the task scenario and the specific components used to construct the modular systems and the end-to-end audio LMs for comparison.

Task Scenario: Multi-phase Task-Oriented Dialogue Unlike prior work focusing on single-turn or general chat, this study specifically examines the quality difference between E2E audio models and cascaded systems within task-oriented scenarios.

We adopt the SALESBOT framework (Chiu et al., 2022), which provides a robust scenario for our experimental design. This framework introduced the first large-scale dataset for conversations that naturally transition from open-domain chitchat to task-oriented purposes, addressing a critical gap in sales and business contexts. The SalesBot dataset simulates a natural flow through three distinct phases: *chitchat*, *transition*, and the core *task-oriented dialogue*. For our experiments, we assess the performance of both systems at each dialogue turn by providing all existing dialogue context to control the variance of multi-turn interactions.

3.1 System Architectures

Modular Baseline Architecture To establish a rigorous standard for conversational quality, we construct a strong modular baseline composed of three state-of-the-art components, representing the typical ASR $\rightarrow$ LLM $\rightarrow$ TTS pipeline.

- Speech-to-Text (ASR): We use `Whisper-2-large` (Radford et al., 2023), a 1.5B-parameter model, for accurate speech recognition.
- Large Language Model (LLM): The transcribed text is processed by a text-based LLM (`gpt-4o-mini` or the text mode of `Qwen-Omni`). This component is responsible for all dialogue reasoning, user intent understanding, and coherent textual response generation.
- Text-to-Speech (TTS): We convert the generated text response back into audio using Sesame-1B (Schalkwyk et al., 2024), a 1B-parameter conversational speech model (CSM). We chose Sesame-1B for its specific design intent: generating high-quality, natural-sounding, and context-aware audio.

End-to-End Audio Language Models We select audio LMs that are capable of generating responses in both text and audio modalities. This allows us to compare their performance when prompted to respond in audio (end-to-end mode) versus text (modular, as part of an LLM pipeline).

3.2 Evaluation Metrics

To rigorously assess the performance of the E2E AudioLMs against the modular system, our evaluation methodology is structured to clearly distinguish between audio signal quality and the core dialogue response ability.

Audio Signal Quality Assessment We employ a combination of objective and subjective metrics to evaluate the fidelity and intelligibility of the generated speech:

- Objective Quality: We utilize the audio quality assessment model MOSA-Net+ (Zezario et al., 2024). This model provides a quality score ranging from 1 to 5, alongside a dedicated intelligibility score (ranging from 0 to 1), offering a technical measure of audio performance.
- Perceptive Quality: We use an LLM-as-Judge approach that leverages the audio understanding capabilities of `gemini-2.5-flash` to assign a subjective quality metric. Prompt is available at Appendix C.

Dialogue Coherence and Consistency Assessment To isolate dialogue modeling ability from audio quality differences, we standardize evaluation inputs by transcribing all audio outputs with `Whisper-v3-large`. This ensures both E2E AudioLMs and modular systems are subject to identical ASR errors, allowing fair comparison of their dialogue response capabilities. We then evaluate the transcripts using the LLM-as-Judge framework (Lin and Chen, 2023) from SALESBOT (Chang and Chen, 2024), measuring naturalness and consistency scores.

4 Results

We evaluate the comparative performance between end-to-end (E2E) and modular dialogue systems (M) across three models: `MiMo-Audio` (Xiaomi, 2025), `Qwen-Omni-3B`, and `Qwen-Omni-7B` (Xu et al., 2025). The E2E systems process audio directly without intermediate text representations, while modular systems employ a cascade architecture (ASR+LLM+TTS) with explicit text intermediates.

4.1 Dialogue Coherence vs. Audio Quality

Our evaluation, summarized in Table 1, reveals that modular systems consistently outperform E2E models in dialogue quality. Figure 2 illustrates the source of this gap: E2E models exhibit severe degradation in Naturalness and Consistency as the dialogue progresses which not seen in modular versions. For instance, the naturalness of Qwen-Omni-3B (E2E) plummets from 51.4 to 23.9 within seven turns.

	Audio		Dialogue	
Model	**Qual.**	**Intel.**	**Natural.**	**Consist.**
MiMo (E2E)	2.91	0.86	71.1	74.0
MiMo (M)	3.42	0.93	66.0	70.0
Omni-3B (E2E)	3.90	0.98	28.9	29.7
Omni-3B (M)	3.97	0.98	73.7	76.6
Omni-7B (E2E)	3.87	0.98	68.2	70.5
Omni-7B (M)	3.96	0.98	82.7	87.0
GPT-4o-m (M)†	3.94	0.98	88.8	90.6
GPT-4o-m (T)‡	—	—	92.2	95.7

Table 1: Evaluation results comparing end-to-end (E2E) and modular audio dialogue systems (M). †GPT-4o-mini with TTS output. ‡Text-only baseline w/o audio.

Model	**Gen.**	**Topic**	**Rep.**	**Mis.**	**Mem.**
MiMo (E2E)	12.8	11.1	2.7	11.4	1.7
MiMo (M)	19.3	4.4	15.9	3.5	5.7
Qwen-3B (E2E)	67.7	10.1	50.5	11.0	39.4
Qwen-3B (M)	19.3	4.9	7.1	3.2	5.9
Qwen-7B (E2E)	28.9	10.7	9.8	10.6	7.9
Qwen-7B (M)	13.8	3.4	2.4	2.0	1.8

Table 2: Average failure severity across failure types. Higher values indicate more severe failures.

Crucially, this dialogue failure is **not** due to audio generation. As shown in the bottom panels of Figure 2, audio quality metrics remain high and stable for all systems throughout the dialogue. Furthermore, Figure 3 shows a negligible correlation (Pearson's $r < 0.06$) between audio fidelity and dialogue coherence metrics. This dissociation strongly indicates that the performance gap stems from fundamental limitations in the E2E models' dialogue modeling and context maintenance, not from audio generation artifacts.

4.2 Failure Analysis

To understand the mechanisms underlying E2E dialogue degradation, we performed a fine-grained error analysis on transcribed dialogues. We categorized errors into five types with severity scores (0-100, where 100 indicates complete failure):

- **Generic**: Non-specific, template-like responses lacking contextual grounding
- **Topic Drift**: Deviation from the established conversation topic
- **Repetition**: Redundant content across turns
- **Misunderstanding**: Failure to correctly interpret user intent
- **Memory**: Inability to maintain context from previous turns

Table 2 presents the average failure severity

Figure 2: Top panels show text-based dialogue quality metrics (Naturalness and Consistency) evaluated on text, while bottom panels display audio quality metrics (MOSNet Quality and Voice Consistency). All audio quality metrics remain relatively stable across all systems throughout the dialogue, suggesting audio generation quality is not the primary factor driving dialogue coherence degradation.

across models. The analysis reveals distinct failure profiles between E2E and modular systems. Qwen-Omni-3B (E2E) exhibits catastrophic failure modes with Generic (67.7), Repetition (50.5), and Memory (39.4) severities that are 3-8× higher than its modular counterpart. These failures directly correspond to the rapid naturalness degradation observed in Figure 2. MiMo-Audio presents a contrasting pattern: its E2E variant shows lower Generic (12.8 vs. 19.3) and Repetition (2.7 vs. 15.9) severities compared to its modular version, suggesting that direct audio generation may inherently avoid certain templated response patterns common in text-based systems.

However, all E2E models consistently fail at semantic understanding tasks. Misunderstanding rates increase 3-5× in E2E configurations (MiMo-Audio: 11.4 vs. 3.5; Qwen-Omni-7B: 10.6 vs. 2.0), and Topic Drift similarly escalates across all E2E variants. While E2E architectures reduce surface-level artifacts but struggle with core dialogue tasks: tracking intent, maintaining coherence, and managing context.

5 Conclusion

We present a framework for benchmarking end-to-end AudioLMs against modular systems in task-oriented dialogues. Our findings indicate that current open-weight AudioLMs consistently lag behind their modular counterparts, suffering from severe turn-by-turn degradation. This performance

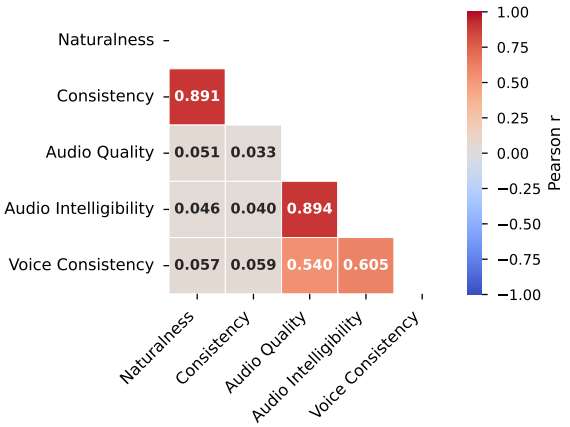

Figure 3: Correlation matrix between text-based dialogue quality metrics (Naturalness and Consistency) and audio quality metrics (Quality, Intelligibility, and Voice Consistency).

gap stems not from audio processing issues, but from fundamental challenges in maintaining conversational coherence. Future research should determine whether these limitations arise from architectural constraints or training data, while exploring hybrid approaches that leverage AudioLMs for latency-critical components.

Acknowledgments

This work was financially supported by the National Science and Technology Council (NSTC) in Taiwan, under Grant 112-2223-E-002-012-MY5.

References

Wen Chang and Yun-Nung Chen. 2024. Injecting salesperson's dialogue strategies in large language models with chain-of-thought reasoning. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 3798–3812, Bangkok, Thailand. Association for Computational Linguistics.

Ssu Chiu, Maolin Li, Yen-Ting Lin, and Yun-Nung Chen. 2022. SalesBot: Transitioning from chit-chat to task-oriented dialogues. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6143–6158, Dublin, Ireland. Association for Computational Linguistics.

Alexandre Défossez, Laurent Mazaré, Manu Orsini, Amélie Royer, Patrick Pérez, Hervé Jégou, Edouard Grave, and Neil Zeghidour. 2024. Moshi: a speech-text foundation model for real-time dialogue. *arXiv preprint arXiv:2410.00037*.

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

Yen-Ting Lin and Yun-Nung Chen. 2023. LLM-eval: Unified multi-dimensional automatic evaluation for open-domain conversations with large language models. In *Proceedings of the 5th Workshop on NLP for Conversational AI (NLP4ConvAI 2023)*, pages 47–58, Toronto, Canada. Association for Computational Linguistics.

Kate Lister, Tim Coughlan, Francisco Iniesto, Nick Freear, and Peter Devine. 2020. Accessible conversational user interfaces: considerations for design. In *Proceedings of the 17th international web for all conference*, pages 1–11.

Alisha Pradhan, Kanika Mehta, and Leah Findlater. 2018. "Accessibility Came by Accident" use of voice-controlled intelligent personal assistants by people with disabilities. In *Proceedings of the 2018 CHI Conference on human factors in computing systems*, pages 1–13.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pages 28492–28518. PMLR.

Johan Schalkwyk, Ankit Kumar, Dan Lyth, Sefik Emre Eskimez, Zack Hodari, Cinjon Resnick, Ramon Sanabria, Raven Jiang, and the Sesame team. 2024. Sesame CSM-1B: Conversational speech model. `https://csm1b.com/`. Accessed: 2025-09-02.

Min-Han Shih, Ho-Lam Chung, Yu-Chi Pai, Ming-Hao Hsu, Guan-Ting Lin, Shang-Wen Li, and Hung-yi Lee. 2024. GSQA: An end-to-end model for generative spoken question answering. In *Proceedings of Interspeech 2024*, pages 2970–2974.

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 *Proceedings of the 37th International Conference on Neural Information Processing Systems*, pages 39088–39118.

Armand Stricker and Patrick Paroubek. 2024. Chitchat as interference: Adding user backstories to task-oriented dialogues. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 3203–3214, Torino, Italia. ELRA and ICCL.

Peng Wang, Songshuo Lu, Yaohua Tang, Sijie Yan, Wei Xia, and Yuanjun Xiong. 2024. A full-duplex speech dialogue scheme based on large language model. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, pages 13372–13403.

LLM-Core-Team Xiaomi. 2025. MiMo-Audio: Audio language models are few-shot learners.

Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, et al. 2025. Qwen2. 5-omni technical report. *arXiv preprint arXiv:2503.20215*.

Ryandhimas E Zezario, Yu-Wen Chen, Szu-Wei Fu, Yu Tsao, Hsin-Min Wang, and Chiou-Shann Fuh. 2024. A study on incorporating whisper for robust speech assessment. In *2024 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE.

A Dataset Details

Our evaluation dataset is derived from SalesBot 2.0 (Chang and Chen, 2024), a large-scale sales dialogue dataset designed for multi-phase conversational systems. We sampled 500 representative dialogues from the original MSGD dataset following the original intent distribution to create our audio evaluation corpus.

Table 3 presents the core statistics. The 500 dialogues comprise 10,175 total turns, averaging

Statistic	Value
Total Dialogues	500
Avg. Turns	20.35
Std. Dev. Turns	6.68
Min/Max Turns	7/72
Avg. Transition Position	8.10
Avg. Chitchat Length	5.30

Table 3: Dataset statistics for sampled 500 dialogues from SalesBot.

Intent Type	Count	%
FindRestaurants	121	24.2%
FindMovie	113	22.6%
FindAttraction	106	21.2%
LookupMusic	67	13.4%
FindEvents	37	7.4%
SearchHotel	28	5.6%
SearchRoundtripFlights	14	2.8%
GetCarsAvailable	9	1.8%
SearchOnewayFlight	3	0.6%
GetRide	2	0.4%
Total	500	100.0%

Table 4: Intent distribution across dialogues.

20.35 turns per dialogue (SD=6.68, range 7-72). This exceeds the original SalesBot 2.0 mean of 9.29 turns, enabling evaluation of long-context coherence maintenance. Following the SalesBot 2.0 framework, each dialogue progresses through three phases: chitchat for rapport building (average 5.30 turns), transition where the agent steers toward sales topics (average position 8.10), and task-oriented dialogue for goal completion.

Table 4 shows the distribution across 10 intent types spanning entertainment, travel, and transportation domains. The top three intents (FindRestaurants 24.2%, FindMovie 22.6%, FindAttraction 21.2%) comprise 68% of dialogues, reflecting common consumer assistance scenarios while maintaining domain balance across all intent categories.

B Detailed Results

In this section, we provides the complete per-turn numerical results for dialogue quality metrics across all evaluated systems in Table 5. These detailed breakdowns support the aggregate statistics reported in Table 1 and the turn-by-turn degradation patterns visualized in Figure 2.

C Audio Evaluation Prompt

To evaluate the audio voice consistency found in the dialogue, we use the similar evaluation prompt structure from SalesAgent (Chang and Chen, 2024) on per-turn audio.

Evaluation Prompt:
The following is a conversation between a user and a salesbot, and the goal of salesbot is to smoothly direct the conversation toward a certain topic and proceed to task-oriented dialogue agent.

=== Conversation History ===
[past multiple audio rounds here]

=== Agent Audio to Evaluate ===
[latest round of audio response here]

You will receive:

1. The conversation history up to this point (as text for context)
2. The current agent audio turn to evaluate

Your task: Evaluate ONLY the current agent audio turn on a scale from 0 to 100 for each metric. Focus on the audio quality and appropriateness of this specific turn.

Definition of the score (for this specific agent turn):
Voice Consistency (the higher the more consistent): The agent's voice characteristics (pitch, tone, speaking style) are consistent with previous agent turns.

Return strictly and only a single JSON object matching this format:

```
{
    "voice_consistency": {
        "reason": "<reason for score>",
        "score": <voice consistency score>
    }
}
```

Important:
- Output must be valid JSON. Do not include any additional commentary.
- Use integers between 0 and 100 for all scores.
- Evaluate the audio quality of THIS TURN, not the entire conversation.

Figure 4: LLM-as-Judge evaluation prompt for evaluating voice consistency.

Model	R1	R2	R3	R4	R5	R6	R7	R8	R9	R10
					Consistency					
GPT-4o-m	96.6	95.3	94.7	95.2	96.2	95.6	95.1	95.6	96.1	97.1
Cascade	96.9	87.5	89.2	91.1	89.7	87.4	90.1	91.9	92.0	93.7
MiMo (E2E)	84.1	75.7	75.9	77.3	**71.8**	**70.2**	**71.8**	**73.9**	74.5	**69.6**
MiMo (Text)	72.6	77.2	71.6	71.9	70.4	68.8	69.3	69.9	65.4	64.5
Omni-3B (E2E)	57.0	**34.9**	**27.4**	**27.9**	**21.0**	**23.6**	**24.3**	**24.3**	**29.3**	**26.6**
Omni-3B (Text)	84.3	80.8	75.6	**72.5**	75.7	**72.4**	76.2	76.7	75.9	75.9
Omni-7B (E2E)	79.4	70.9	**68.9**	71.5	**65.7**	**67.2**	72.2	71.5	**68.2**	71.6
Omni-7B (Text)	91.0	88.0	86.7	85.9	84.3	85.9	84.1	90.4	87.4	85.9
					Naturalness					
GPT-4o-m	92.3	92.2	92.7	91.7	93.4	91.7	91.7	91.1	92.2	93.6
Cascade	92.6	85.1	87.7	89.7	88.0	87.0	89.2	90.2	89.5	92.4
MiMo (E2E)	71.4	73.8	75.4	76.9	68.1	68.7	70.0	71.4	71.0	70.2
MiMo (Text)	66.6	73.9	68.0	67.9	67.5	64.6	64.6	65.8	64.4	58.8
Omni-3B (E2E)	51.4	**34.6**	**26.5**	**26.7**	**22.7**	**22.9**	**23.9**	**23.9**	**30.7**	**25.0**
Omni-3B (Text)	78.6	75.9	74.5	71.4	73.2	69.6	73.6	72.9	73.2	72.6
Omni-7B (E2E)	73.5	69.0	66.4	69.4	65.5	66.1	67.8	69.9	67.6	69.1
Omni-7B (Text)	84.9	85.1	82.2	83.0	79.5	80.5	81.7	85.3	82.8	82.4

Table 5: Per-round scores showing degradation across dialogue turns. R1-R10 represent turns 1-10. Bold values indicate drops >10 points from initial turn.