Measuring Prosodic Richness in LLM-Generated Responses for Conversational Recommendation

Darshna Parmar, Pramit Mazumdar
Department of Computer Science and Engineering
Indian Institute of Information Technology Vadodara
{darshna.parmar, pramit.mazumdar}@iiitvadodara.ac.in

Abstract

This paper presents a novel framework for stylistic evaluation in conversational recommendation systems (CRS), focusing on the prosodic and expressive qualities of generated responses. While prior work has predominantly emphasized semantic relevance and recommendation accuracy, the stylistic fidelity of model outputs remains underexplored. We introduce the prosodic richness score (PRS), a composite metric that quantifies expressive variation through structural pauses, emphatic lexical usage, and rhythmic variability. Using PRS, we conduct both sentence-level and turn-level analyses across six contemporary large language models (LLMs) on two benchmark CRS datasets: ReDial, representing goal-oriented dialogue, and INSPIRED, which incorporates stylized social interaction. Empirical results reveal statistically significant differences ($p < 0.01$) in PRS between human and model-generated responses, highlighting the limitations of current LLMs in reproducing natural prosodic variation. Our findings advocate for broader evaluation of stylistic attributes in dialogue generation, offering a scalable approach to enhance expressive language modeling in CRS.

1 Introduction

Conversational Recommendation Systems (CRS) aim to provide personalized recommendations through natural language dialogue (Jannach et al., 2021). With the emergence of large language models (LLMs), such as LLaMA, Mistral, and Gemini, the quality of generated dialogue has significantly improved in terms of coherence, informativeness, and contextual relevance (Thoppilan et al., 2022; Li et al., 2018; Numaya et al., 2025). However, most existing CRS evaluation methods emphasize automatic metrics—such as BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), recommendation accuracy, and human evaluation (Liu et al., 2020)—while neglecting stylistic features, particularly prosodic aspects like rhythm, expressiveness, and tone.

Prosody—referring to rhythm, emphasis, and structural variation in language—is a key component of natural human communication (Ladd and Arvaniti, 2023). In text-based dialogue systems, prosodic features manifest through punctuation (e.g., pauses via commas or periods), lexical emphasis (e.g., adjectives, adverbs, interjections), sentence rhythm (e.g., variance in sentence length), and expressive markers (e.g., questions and exclamations). These cues contribute to the emotional resonance and naturalness of responses, making them especially important for CRS applications aiming to simulate engaging human-like dialogue or support voice interfaces.

Despite recent advances, current LLM-based systems often generate stylistically flat responses lacking variation in tone, structure, or emphasis—limiting engagement and diminishing perceived human-likeness. Existing evaluation frameworks fail to capture these nuanced yet important dimensions of expressiveness (Li et al., 2024; Park et al., 2024). To address this gap, we propose a novel method for analyzing the prosodic quality of LLM-generated responses in CRS. Specifically, we quantify expressiveness using three interpretable features: (i) structural pauses (e.g., punctuation frequency), (ii) lexical emphasis (e.g., adjectives, adverbs, interjections), and (iii) rhythm variation (variance in sentence lengths). PRS provides a normalized score that captures the richness of a response's structure and delivery style. To the best of our knowledge, this is the first systematic attempt to assess textual prosody in CRS responses.

We investigate the stylistic expressiveness of dialogue responses generated by six recent LLMs within a standard CRS framework. Our evaluation focuses on models such as LLaMA, Gemma,

Proceedings of the Workshop on Beyond English: NLP for all Languages in an Era of LLM from Texts associated with RANLP 2025, pages 120–130, Varna, Bulgaria, Sep 12, 2025.
https://doi.org/10.26615/978-954-452-105-9-014

Gemini, Qwen, and Mistral, assessing their ability to produce responses with natural prosodic qualities. The CRS model is trained and evaluated on the standard splits of two benchmark datasets: ReDial (Li et al., 2018), which is goal-oriented and centers on movie recommendations, and INSPIRED (Hayati et al., 2020), which emphasizes emotionally rich conversations. We analyze the LLM-generated responses using both sentence-level and turn-level PRS distributions, compare them with human-authored ground-truths, and examine stylistic variation across dialogue turns.

This enables a more holistic evaluation of LLMs in CRS by capturing stylistic expressiveness often overlooked by standard automatic and human evaluation methods.

Our key contributions are as follows:

- We introduce the Prosodic Richness Score (PRS), a novel metric designed to quantify stylistic and prosodic expressiveness in dialogue responses.
- We evaluate PRS on responses generated by six state-of-the-art LLMs, using a standard CRS framework applied to two benchmark datasets: ReDial and INSPIRED.
- Our analysis uncovers consistent prosodic gaps between LLM-generated and human-authored responses, highlighting the limitations of current models in producing naturally expressive dialogue.

This work offers a new lens for evaluating conversational agents, emphasizing not only what is generated but how it is said. Our findings underscore the need for more prosody-aware generation techniques to bridge the gap between human and machine dialogue.

2 Related Work

While LLMs have enhanced CRSs in terms of relevance and coherence, prior work has largely emphasized task-oriented metrics like recommendation accuracy. In contrast, stylistic and prosodic factors—crucial for naturalness and user engagement—remain underexplored. This section reviews related research on CRS evaluation, stylistic and affective language generation, and the role of prosody in natural language processing.

2.1 Evaluation of Conversational Recommender Systems

CRSs aim to provide personalized suggestions through dialogue (Christakopoulou et al., 2016; Li et al., 2018; Jannach et al., 2021). Traditional evaluation relies on task-specific metrics like BLEU, ROUGE, or Recall@K, which overlook stylistic richness and perceived naturalness. While human evaluations and learned reward models offer deeper insights (See et al., 2019; Ghazarian et al., 2022), they are costly and less interpretable. Recent reference-free evaluators, such as FACE (Chen et al., 2025), show strong alignment with human judgments, and others demonstrate robustness under adversarial settings (Vasselli et al., 2025). In contrast, our PRS provides a lightweight and interpretable measure of stylistic expressiveness.

2.2 Stylistic and Affective Generation in Dialogue

Stylistic elements such as tone, emotion, and personality play a vital role in enhancing conversational engagement(Qian et al., 2023; Ma et al., 2024). Dataset like Empathetic Dialogues (Rashkin et al., 2019) has highlighted the importance of generating affective and stylistically rich responses, particularly in CRS. Prior approaches to stylistic control in text generation have leveraged lexical constraints (Iso, 2024), persona-based latent attribute control (Lu et al., 2023), and decoding-time mechanisms via dynamic attribute graphs (Liang et al., 2024). However, evaluation metrics for stylistic expressiveness remain limited. We address this gap by introducing PRS, which facilitates cross-LLM comparisons of stylistic variation in CRS outputs, with a focus on the ReDial and INSPIRED dataset.

2.3 Prosody in Text and Speech Systems

Prosody—encompassing rhythm, emphasis, and structural variation—plays a crucial role in human communication and is extensively studied in speech synthesis(Li et al., 2025; Raitio et al., 2022). While TTS models incorporate prosodic control(Liu et al., 2024; Que and Ragni, 2025), textual dialogue evaluation rarely considers prosodic features. Some studies utilize prosodic cues for emotion recognition or discourse segmentation(Wei et al., 2023; Prévot and Wang, 2024), but these often rely on acoustic inputs. Our work is distinct in applying prosody-based analysis directly to text, enabling

stylistic evaluation of LLM-generated responses in CRS.

While previous work has made strides in affective dialogue generation and prosody modeling—particularly in speech applications—there remains a lack of principled, text-based analysis for quantifying prosodic expressiveness in generated responses. Our work bridges this gap by introducing a simple yet effective measure, the PRS, enabling scalable, interpretable evaluation of stylistic quality in LLM-generated responses across diverse CRS datasets.

3 Methodology

This section outlines our approach to evaluating the stylistic and prosodic expressiveness of responses generated by LLMs in conversational recommendation systems. We introduce a lightweight linguistic framework that extracts text-based prosodic features and computes a unified PRS to assess variation and naturalness across both human-authored ground-truths and model-generated responses.

3.1 Datasets

We use two publicly available CRS datasets. ReDial (Li et al., 2018) contains over 10,000 goal-driven movie recommendation dialogues with annotated movie mentions, emphasizing task performance. In contrast, INSPIRED (Hayati et al., 2020) includes 1,001 open-domain dialogues enriched with tone annotations (e.g., empathetic, humorous), supporting evaluation of affective and stylistic expressiveness. Together, these datasets enable both functional and stylistic analysis.

3.2 LLM-Generated Response Collection

To analyze stylistic variation in generated dialogue, we evaluate responses from six large language models: llama-3.1-8b-instant (Touvron et al., 2023), llama-3.2-3b-preview (Touvron et al., 2023), gemma2-9b-it (Anil et al., 2024), gemini-1.5-flash-8b (Google DeepMind, 2024), qwen-2.5-32b (Inc., 2024), and mistral-saba-24b (Jiang et al., 2024). The selected models span a broad spectrum of capacities, ranging from 3B to 32B parameters, and include both decoder-only and multimodal architectures. This range ensures coverage of both lightweight and high-capacity LLMs, allowing us to examine whether expressive richness in generated responses is consistently maintained across models of varying size and complexity. Each model is prompted to produce a single-turn response given a user utterance from either the ReDial or INSPIRED dataset. Each generated output is aligned with its corresponding human-authored ground-truth response, enabling direct comparison of prosodic and stylistic characteristics. For transparency and reproducibility, we include the exact prompts used during LLM-based response generation in Appendix A.

3.3 Text-Based Prosody Feature Extraction

To quantify the stylistic expressiveness, we compare responses generated by LLMs against human-written ground-truth responses from established CRS datasets. We extract a set of interpretable textual features that serve as proxies for prosodic expressiveness in dialogues generated via various LLMs. Specifically, we compute:

- Pause count: The number of punctuation-based structural markers (e.g., periods, commas, semicolons, question marks, and exclamation marks), which simulate natural pauses in spoken language.
- Emphasis count: The number of expressive lexical items—identified via part-of-speech tags such as adjectives, adverbs, and interjections—that often signal emotional tone or subjective emphasis.
- Rhythm variance: The variance in sentence length (in tokens), reflecting diversity in syntactic structure and rhythmic flow.
- Question and exclamation counts: The number of interrogative and exclamatory sentences, capturing tone variability and conversational dynamism.
- Sentence count: The number of distinct sentences in a response, offering a basic structural measure of length and complexity.

While PRS captures stylistic variation through pause, emphasis, and rhythm features, we acknowledge that overuse of these markers could artificially inflate scores. All features are extracted using a spaCy-based linguistic preprocessing pipeline, including sentence segmentation and part-of-speech tagging, which ensures consistent and linguistically informed analysis across both model-generated and human-authored responses.

3.3.1 Defining the Prosodic Richness Score

We define the PRS as a composite metric to capture the stylistic richness of a LLM generated response:

$$\text{PRS} = \frac{1}{10}\left(0.4 \cdot \text{pause} + 0.3 \cdot \text{emph} + 0.3 \cdot \text{var}\right) \tag{1}$$

The score is normalized between 0 and 1 to support direct comparisons across models and datasets. A higher score reflects greater stylistic diversity and perceived naturalness in generated responses. The weights in Equation 1 are assigned based on empirical observations of the contribution of each feature to stylistic expressiveness, with pause count receiving slightly greater weight due to its consistent role in conveying natural rhythm.

3.4 Prosody-Aware Evaluation Method

We propose a multi-level evaluation framework based on the PRS to assess the stylistic quality of LLM-generated dialogue. At the sentence level, PRS captures local prosodic expressiveness by comparing model outputs to human references. At the turn level, we analyze PRS evolution across dialogue turns to identify stylistic consistency or degradation over time. For model-wise comparison, we compute the average PRS for each LLM and benchmark it against human-authored baselines. To evaluate the statistical significance of stylistic differences, we conduct paired t-tests on the PRS distributions of model and human responses.

The final PRS is computed as defined in Equation 1. This score is calculated for both model-generated utterances (*model_PRS*) and the human-written ground truth responses (*gt_PRS*) at the sentence level. Although the reference sentences are drawn from the original ReDial and INSPIRED datasets—resources that may have been partially seen during pretraining—they serve as domain-aligned and affect-rich baselines for stylistic evaluation. This prosody-aware framework thus enables a linguistically grounded, fine-grained, and interpretable assessment of expressiveness in conversational recommendation systems.

4 Results & Discussion

To assess the stylistic expressiveness of LLMs in CRSs, we analyze PRS at both the sentence and turn levels. Our experiments aim to answer the following research questions:

- **RQ1:** How do different LLMs compare to human responses in terms of sentence-level stylistic expressiveness?
- **RQ2:** How does the prosodic richness of model-generated responses vary across successive dialogue turns?
- **RQ3:** To what extent do LLMs sustain stylistic diversity throughout an interaction compared to human-written dialogues?
- **RQ4:** What are the model-specific trends in stylistic degradation or consistency, and which models demonstrate stronger ability to retain prosodic richness?
- **RQ5:** Which LLM demonstrates the highest overall stylistic richness?

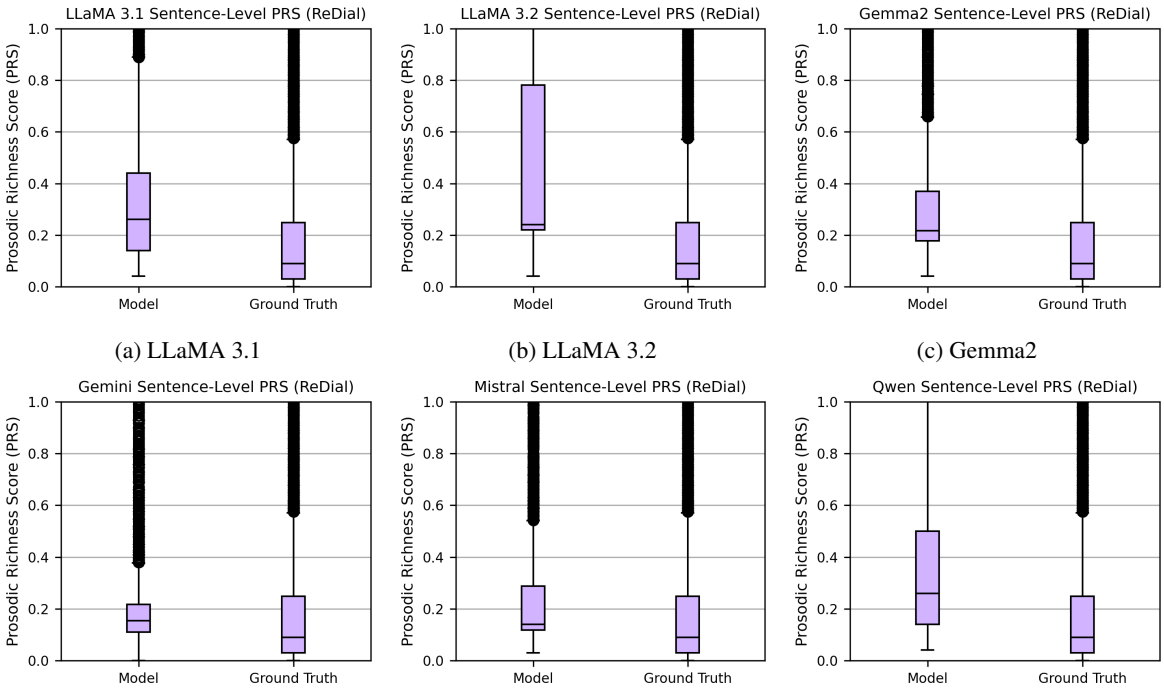
These questions guide our analysis of the expressive capacity of LLMs using PRS as a stylistic evaluation metric. We use both sentence-level and turn-level granularity to investigate how well models emulate human-like variation in tone, rhythm, and emphasis across CRS interactions.

4.1 RQ1: Sentence-Level Stylistic Expressiveness

To capture fine-grained expressiveness in CRS outputs, we compute PRS at the sentence level. Sentence-level analysis is essential, as individual utterances shape tone, emotional resonance, and user engagement—particularly in affect-rich dialogues. This granularity helps assess how well LLMs emulate human-like prosodic variation.

Figures 1 and 2 show PRS distributions across six LLMs. In both datasets, human responses consistently exhibit greater stylistic richness, with higher medians and broader variance. The gap is especially pronounced in INSPIRED, which contains emotionally expressive, tone-sensitive dialogue.

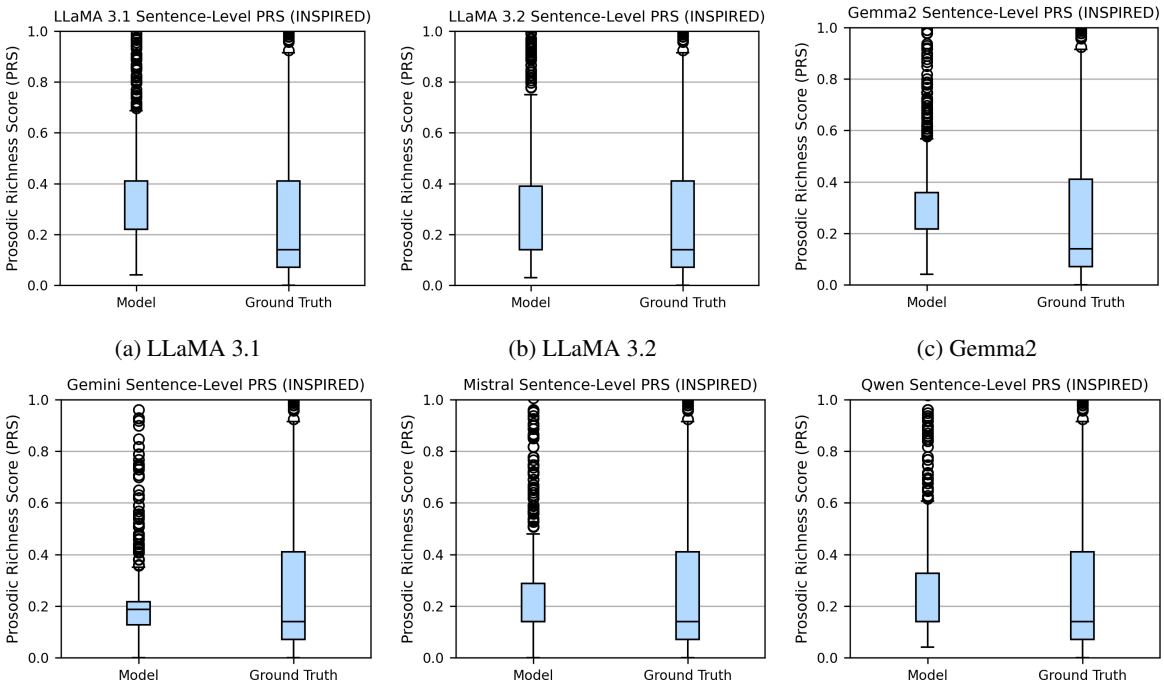
Among the models, LLaMA 3.1 and 3.2 display relatively higher prosodic variation, while Gemini, Mistral, and Qwen lag behind. Model outputs also show fewer outliers, highlighting their limited expressive variability compared to human responses. These results suggest that while LLMs produce contextually relevant outputs, they often lack the stylistic nuance found in human dialogue. Sentence-level PRS thus provides a valuable diagnostic for evaluating expressive quality and highlights the need for better stylistic modeling in CRS systems.



(a) LLaMA 3.1 (b) LLaMA 3.2 (c) Gemma2

(d) Gemini (e) Mistral (f) Qwen

Figure 1: Sentence-level PRS distributions for ReDial (RQ1), comparing prosodic variation in LLM-generated responses and human-authored references.



(a) LLaMA 3.1 (b) LLaMA 3.2 (c) Gemma2

(d) Gemini (e) Mistral (f) Qwen

Figure 2: Sentence-level PRS distributions for INSPIRED (RQ1), comparing prosodic variation in LLM-generated responses and human-authored references.

4.2 RQ2: PRS Across Dialogue Turns

To assess how stylistic expressiveness evolves throughout a conversation, we compute PRS at the turn level. Unlike sentence-level analysis, turn-level PRS captures the aggregate expressiveness of all utterances within a dialogue turn, revealing how

consistently an LLM maintains stylistic richness across interactions. Turn-level PRS is computed by averaging sentence-level PRS scores within each turn, yielding two PRS sequences per conversation—one for *model_PRS* and one for *gt_PRS*. These plots capture the progression of prosodic richness throughout the conversation.

Figures 3 and 4 illustrate a consistent decline in turn-level PRS for most LLMs, indicating diminished stylistic expressiveness over extended interactions. In contrast, human responses exhibit greater stability and variation, with the model-human gap widening in later turns. Among models, LLaMA 3.2 demonstrates stronger stylistic consistency than Gemini or Mistral, reflecting architectural and training differences. These results underscore the value of turn-level PRS in capturing temporal expressiveness—an essential dimension for affect-rich conversational systems—and motivate the development of style-aware, turn-sensitive generation strategies.

Case-wise Illustration: Appendix B present turn-level PRS patterns for single conversations from ReDial (ConvID: 22709) and INSPIRED (ConvID: 20191127-224739_530_live.pkl alise as 001), respectively. In both cases, LLaMA 3.2 and Gemma2 closely follow human PRS trends, while Gemini and Mistral display flatter or inconsistent profiles, reflecting reduced ability to sustain stylistic expressiveness across turns.

4.3 RQ3–RQ5: Model-Specific Stylistic Trends

To investigate the consistency of stylistic expression across dialogue progression (RQ3), patterns of degradation or stability (RQ4), and overall prosodic richness across models (RQ5), we conduct a comprehensive analysis of sentence- and turn-level PRS across six state-of-the-art LLMs using the ReDial and INSPIRED datasets.

Figure 5 presents a unified visualization of turn-wise evolution and model-level PRS gaps. The turn-wise plots reveal that human-authored responses exhibit consistently higher and more stable prosodic richness throughout the conversation. In contrast, most LLMs show a gradual decline in PRS, particularly in later turns, indicating a loss of stylistic variation over time. Notably, LLaMA 3.2 and Gemma2 demonstrate comparatively stronger stylistic consistency, while Gemini and Mistral show flatter or declining trends, reflecting limited ability to preserve expressive variation.

The model-wise comparison of aggregate PRS values relative to human ground truth. Across both datasets, even the most competitive models fall short of human-level prosodic expressiveness, underscoring persistent limitations in current LLMs. These findings highlight the need for prosody-aware generation strategies that explicitly model expressive diversity and sustain stylistic richness across dialogue turns.

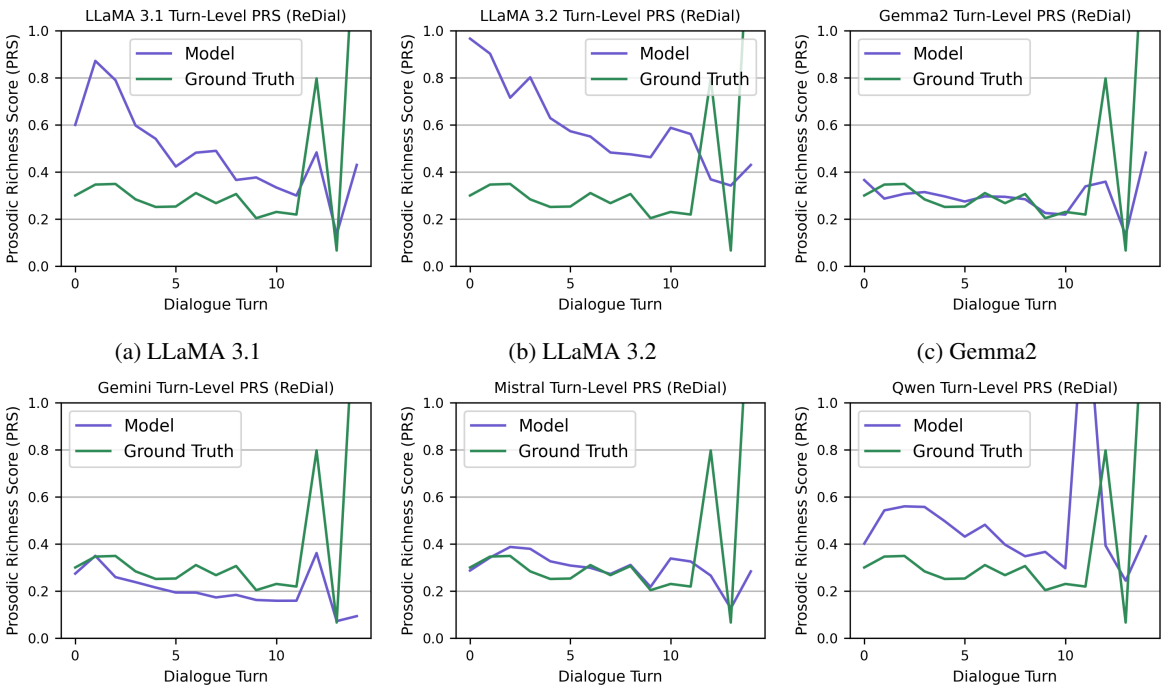
These results, consistent across both datasets, suggest that while LLMs achieve surface-level fluency, they often under utilize stylistic features such as lexical emphasis and rhythm variation, particularly in emotionally expressive contexts like INSPIRED.

5 Statistical Validation of Prosodic Richness Differences

To assess whether LLM-generated responses differ significantly in stylistic expressiveness compared to human-authored dialogue, we conducted a paired t-test on sentence-level PRS. For each model, PRS values were paired with corresponding ground-truth responses across both ReDial and INSPIRED datasets.

The results, summarized in Table 1, reveal several key findings:

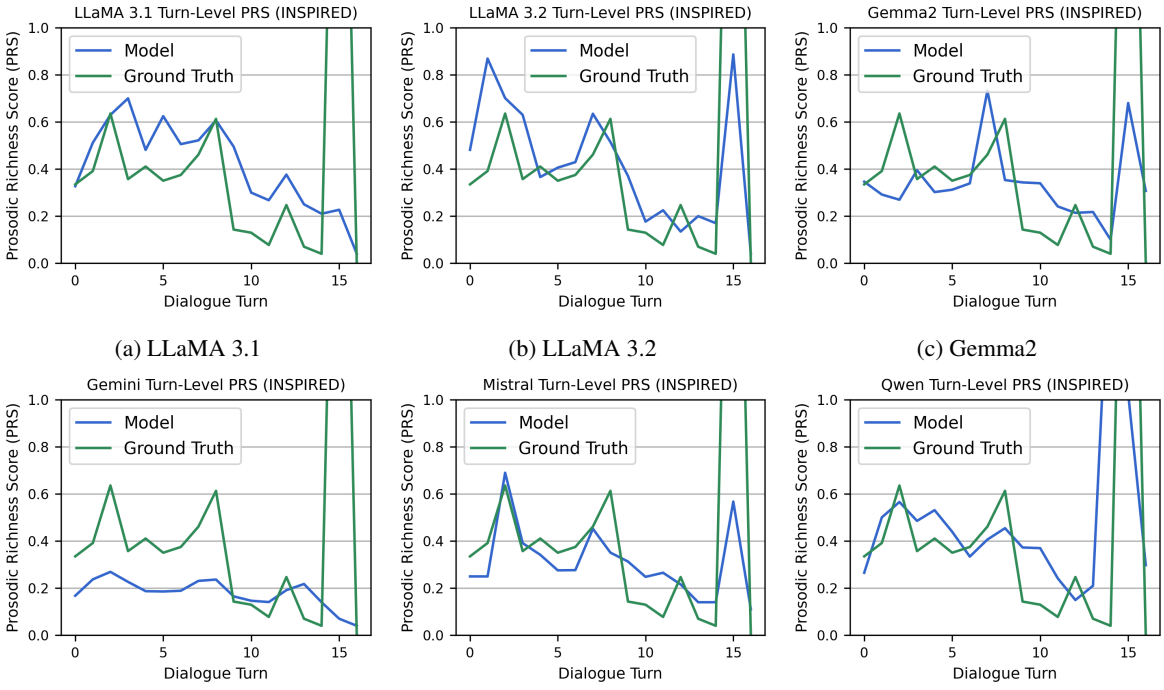
- Most models (e.g., LLaMA 3.1, LLaMA 3.2, Gemini, Qwen) show statistically significant differences ($p < 0.05$) from human responses on ReDial, indicating consistent stylistic divergence.
- On the INSPIRED dataset, however, fewer models show significance, suggesting either reduced expressiveness in the models or more variability in human references.
- LLaMA 3.2 shows significant differences across both datasets, indicating high stylistic deviation despite its strong median PRS.
- In contrast, models like Gemma 2 and Qwen do not differ significantly on INSPIRED, implying closer alignment to human style or reduced variance in outputs.
- Gemini consistently shows negative t-values (e.g., $t = -4.54$, $p < 0.0001$ on ReDial), indicating it underperforms in stylistic richness relative to human-written dialogue.



(a) LLaMA 3.1 (b) LLaMA 3.2 (c) Gemma2

(d) Gemini (e) Mistral (f) Qwen

Figure 3: Turn-level PRS comparisons for ReDial (RQ2), showing how stylistic expressiveness progresses across dialogue turns in LLM-generated and human-authored responses.

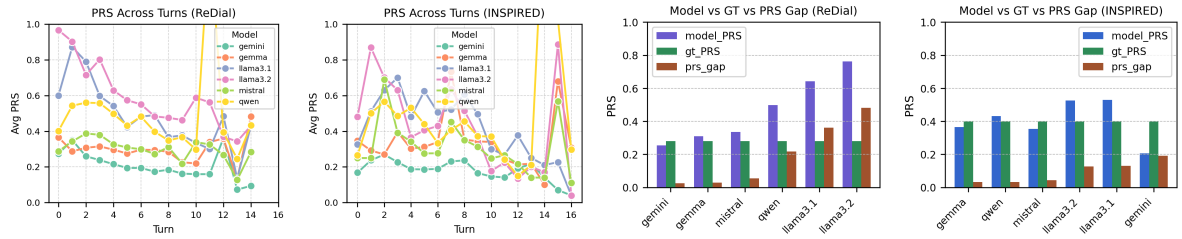


(a) LLaMA 3.1 (b) LLaMA 3.2 (c) Gemma2

(d) Gemini (e) Mistral (f) Qwen

Figure 4: Turn-level PRS comparisons for INSPIRED (RQ2), showing how stylistic expressiveness progresses across dialogue turns in LLM-generated and human-authored responses.

These findings statistically validate that while some LLMs approach human-level prosody, a measurable and significant expressiveness gap still exists. This supports the use of PRS as a diagnostic tool and highlights the need for prosody-aware training methods.



(a) (b) (c) (d)

Figure 5: Visualizations addressing RQ3–RQ5 over ReDial and INSPIRED datasets. (a–b) presents turn-level PRS across dialogue turns, capturing patterns of stylistic consistency and degradation (RQ3, RQ4). (c–d) presents aggregate model-wise PRS comparisons relative to human-authored ground truth, highlighting persistent gaps in expressive richness across LLMs (RQ5).

Table 1: Paired t-test results comparing sentence-level PRS between LLM-generated and human responses across ReDial and INSPIRED datasets.

Model	**ReDial**			**INSPIRED**		
	t-value	**p-value**	**Significance**	**t-value**	**p-value**	**Significance**
LLaMA 3.1	11.4	<0.0001	Yes	3.13	<0.0018	Yes
LLaMA 3.2	23.96	<0.0001	Yes	2.22	<0.0268	Yes
Gemma 2	1.19	<0.2321	No	-0.72	<0.4713	Yes
Gemini	-4.54	<0.0001	Yes	-6.2	<0.0001	Yes
Mistral	3.84	<0.0001	Yes	-0.91	<0.3611	No
Qwen	17.09	<0.0001	Yes	0.83	<0.4078	No

6 Conclusion

We introduced PRS as a stylistic metric to assess LLM-generated responses in conversational recommendation systems. Through sentence- and turn-level analysis on the ReDial and INSPIRED datasets, we found that while LLMs produce coherent responses, they often lack the stylistic richness and variation characteristic of human dialogue—particularly in extended interactions. Among the evaluated models, LLaMA 3.2 demonstrated the highest prosodic expressiveness, whereas Gemini and Mistral lagged behind. These findings underscore the importance of integrating prosodic and stylistic diversity into future CRS models to enable more engaging and human-like conversations.

Limitations

Our study evaluates prosodic expressiveness in CRS responses using PRS, but PRS is not integrated into generation or used for stylistic enhancement. No LLM is fine-tuned with PRS supervision, limiting its direct impact. While benchmarked against human responses, PRS is not yet validated with independent human ratings, and it relies on surface-level textual proxies. Incorporating higher-level prosodic cues or spoken responses could provide a more robust assessment of expressiveness in future work.

Ethics Statement

This work involves the analysis of model-generated and human-written conversational data using publicly available datasets: ReDial and INSPIRED. Both datasets are anonymized and curated for research use, and no personally identifiable information (PII) is processed. Our methodology does not involve human subjects or new data collection. However, we acknowledge that automatic evaluation of stylistic expressiveness may carry inherent biases based on dataset demographics and model training corpora. We urge caution in deploying stylistically expressive models in sensitive domains such as mental health or education, where unintended emotional tone may have real-world consequences.

References

Rohan Anil, Yiding Jiang, et al. 2024. Gemma: Lightweight, state-of-the-art open models. https://ai.google.dev/gemma.

Nuo Chen, Quanyu Dai, Xiaoyu Dong, Xiao-Ming Wu, and Zhenhua Dong. 2025. Face: A fine-grained reference free evaluator for conversational recommender systems. *arXiv preprint arXiv:2501.09493*.

Konstantina Christakopoulou, Filip Radlinski, and Katja Hofmann. 2016. Towards conversational recommender systems. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 815–824.

Sarik Ghazarian, Behnam Hedayatnia, Alexandros Papangelis, Yang Liu, and Dilek Hakkani-Tur. 2022. What is wrong with you?: Leveraging user sentiment for automatic dialog evaluation. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 4194–4204, Dublin, Ireland. Association for Computational Linguistics.

Google DeepMind. 2024. Gemini 1.5 technical report. arXiv preprint arXiv:2403.05530.

Shirley Anugrah Hayati, Dongyeop Kang, Qingxiaoyang Zhu, Weiyan Shi, and Zhou Yu. 2020. Inspired: Toward sociable recommendation dialog systems. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8142–8152. Association for Computational Linguistics.

Baidu Inc. 2024. Qwen2: The next-gen language model family. https://github.com/QwenLM/Qwen.

Hayate Iso. 2024. AutoTemplate: A simple recipe for lexically constrained text generation. In *Proceedings of the 17th International Natural Language Generation Conference*, pages 1–12, Tokyo, Japan. Association for Computational Linguistics.

Dietmar Jannach, Ahtsham Manzoor, Wanling Cai, and Li Chen. 2021. A survey on conversational recommender systems. *ACM Computing Surveys (CSUR)*, 54(5):1–36.

Yujia Jiang, Guillaume Lample, et al. 2024. Mistral 7b. https://mistral.ai/news/mistral-7b/.

D Robert Ladd and Amalia Arvaniti. 2023. Prosodic prominence across languages. *Annual Review of Linguistics*, 9(1):171–193.

Jinpeng Li, Zekai Zhang, Quan Tu, Xin Cheng, Dongyan Zhao, and Rui Yan. 2024. Stylechat: Learning recitation-augmented memory in llms for stylized dialogue generation. *arXiv preprint arXiv:2403.11439*.

Raymond Li, Samira Ebrahimi Kahou, Hannes Schulz, Vincent Michalski, Laurent Charlin, and Chris Pal. 2018. Towards deep conversational recommendations. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Yinghao Aaron Li, Cong Han, and Nima Mesgarani. 2025. Styletts: A style-based generative model for natural and diverse text-to-speech synthesis. *IEEE Journal of Selected Topics in Signal Processing*, 19(1):283–296.

Xun Liang, Hanyu Wang, Shichao Song, Mengting Hu, Xunzhi Wang, Zhiyu Li, Feiyu Xiong, and Bo Tang. 2024. Controlled text generation for large language model with dynamic attribute graphs. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 5797–5814, Bangkok, Thailand. Association for Computational Linguistics.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Jiaxuan Liu, Zhaoci Liu, Yajun Hu, Yingying Gao, Shilei Zhang, and Zhenhua Ling. 2024. Diffstyletts: Diffusion-based hierarchical prosody modeling for text-to-speech with diverse and controllable styles. *arXiv preprint arXiv:2412.03388*.

Zeming Liu, Haifeng Wang, Zheng-Yu Niu, Hua Wu, Wanxiang Che, and Ting Liu. 2020. Towards conversational recommendation over multi-type dialogs. *arXiv preprint arXiv:2005.03954*.

Zhenyi Lu, Wei Wei, Xiaoye Qu, Xian-Ling Mao, Dangyang Chen, and Jixiong Chen. 2023. Miracle: Towards personalized dialogue generation with latent-space multiple personal attribute control. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5933–5957, Singapore. Association for Computational Linguistics.

Zhiqiang Ma, Wenchao Jia, Yutong Zhou, Biqi Xu, Zhiqiang Liu, and Zhuoyi Wu. 2024. Personality enhanced emotion generation modeling for dialogue systems. *Cognitive Computation*, 16(1):293–304.

Ikumi Numaya, Shoji Moriya, Shiki Sato, Reina Akama, and Jun Suzuki. 2025. How stylistic similarity shapes preferences in dialogue dataset with user and third party evaluations. *arXiv preprint arXiv:2507.10918*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

ChaeHun Park, Minseok Choi, Dohyun Lee, and Jaegul Choo. 2024. Paireval: Open-domain dialogue evaluation with pairwise comparison. *arXiv preprint arXiv:2404.01015*.

Laurent Prévot and Sheng-Fu Wang. 2024. Investigating discourse segmentation in taiwan southern min spontaneous speech. In *5th Workshop on Computational Approaches to Discourse*, page 50.

Yushan Qian, Bo Wang, Shangzhao Ma, Wu Bin, Shuo Zhang, Dongming Zhao, Kun Huang, and Yuexian Hou. 2023. Think twice: A human-like two-stage conversational agent for emotional response generation. *arXiv preprint arXiv:2301.04907*.

Shumin Que and Anton Ragni. 2025. Visualspeech: Enhance prosody with visual context in tts. *arXiv preprint arXiv:2501.19258*.

Tuomo Raitio, Jiangchuan Li, and Shreyas Seshadri. 2022. Hierarchical prosody modeling and control in non-autoregressive parallel neural tts. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7587–7591.

Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2019. Towards empathetic open-domain conversation models: A new benchmark and dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5370–5381, Florence, Italy. Association for Computational Linguistics.

Abigail See, Stephen Roller, Douwe Kiela, and Jason Weston. 2019. What makes a good conversation? how controllable attributes affect human judgments. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1702–1723, Minneapolis, Minnesota. Association for Computational Linguistics.

Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, et al. 2022. Lamda: Language models for dialog applications. *arXiv preprint arXiv:2201.08239*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Y-Lan Boureau, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:2307.09288.

Justin Vasselli, Adam Nohejl, and Taro Watanabe. 2025. Measuring the robustness of reference-free dialogue evaluation systems. *arXiv preprint arXiv:2501.06728*.

Xianhao Wei, Jia Jia, Xiang Li, Zhiyong Wu, and Ziyi Wang. 2023. A discourse-level multi-scale prosodic model for fine-grained emotion analysis. *arXiv preprint arXiv:2309.11849*.

Appendix

A Prompts Used for LLM Response Generation

Case 1: No Recommendation Available

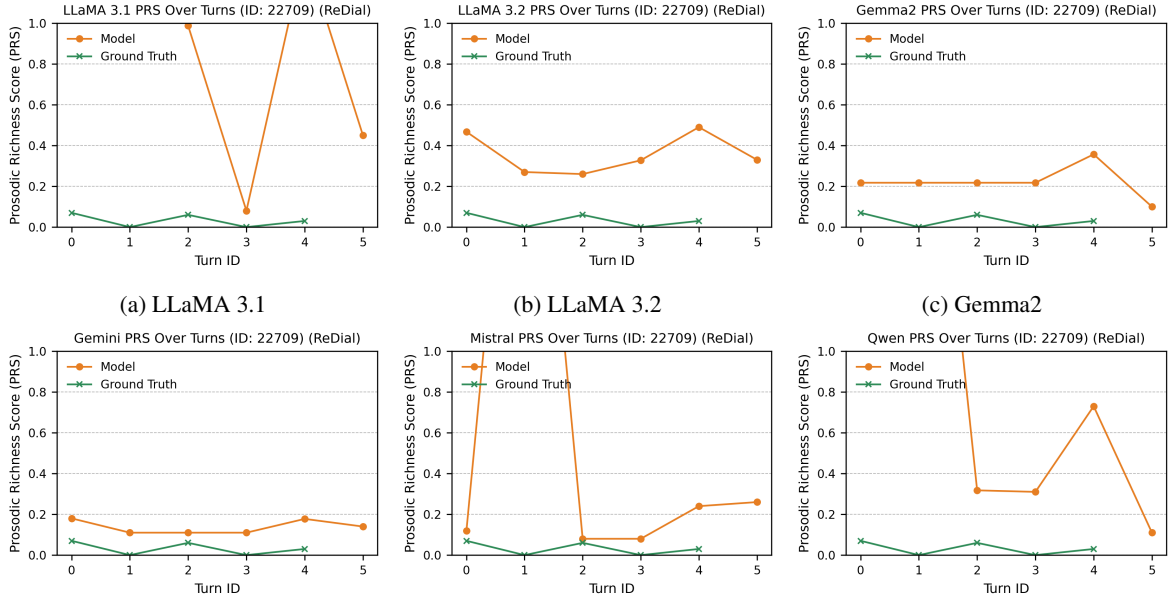
I will provide you with a user input that contains some sort of chit-chat or question. I want you to generate an output text that incorporates a sort of chit chat and then followed by some question related to movies, actors, genres etc.
Example 1: User Input: "Hi, how are you?" Output: "Hi! I'm doing well. What kind of movies are you looking for?" Now, do a similar task for the given user input.

Case 2: Recommendation Available

I will provide you with a user input that contains some movie names, actor names, cast, directors, genre, etc. Additionally, I will provide you with a recommendation that is relevant to the input. I want you to generate an output text that incorporates both the information from the user input and the recommendation.
Example 1: User Input: "I really liked Avengers and SpiderMan. They are both Thrillers and Tom Holland featured in both of them. Released in 2012 directed by Tarantino." Related Attributes: "Thor, Chris Hemsworth." Output: "You can watch Thor. It stars Chris Hemsworth and is similar to the Avengers."
Example 2: If user recommendation is empty then ask the user a relevant question about their likings regarding genres, casts etc and engage with the user.
Example 3: If the user input is present and some ambiguity is present regarding the recommendation generated then clarify it with the user by asking more specific questions regarding the cast, year of release etc. Now, do a similar task for the given user input and recommendation.

B Case-wise Turn-Level PRS Analysis over ReDial and INSPIRED Datasets

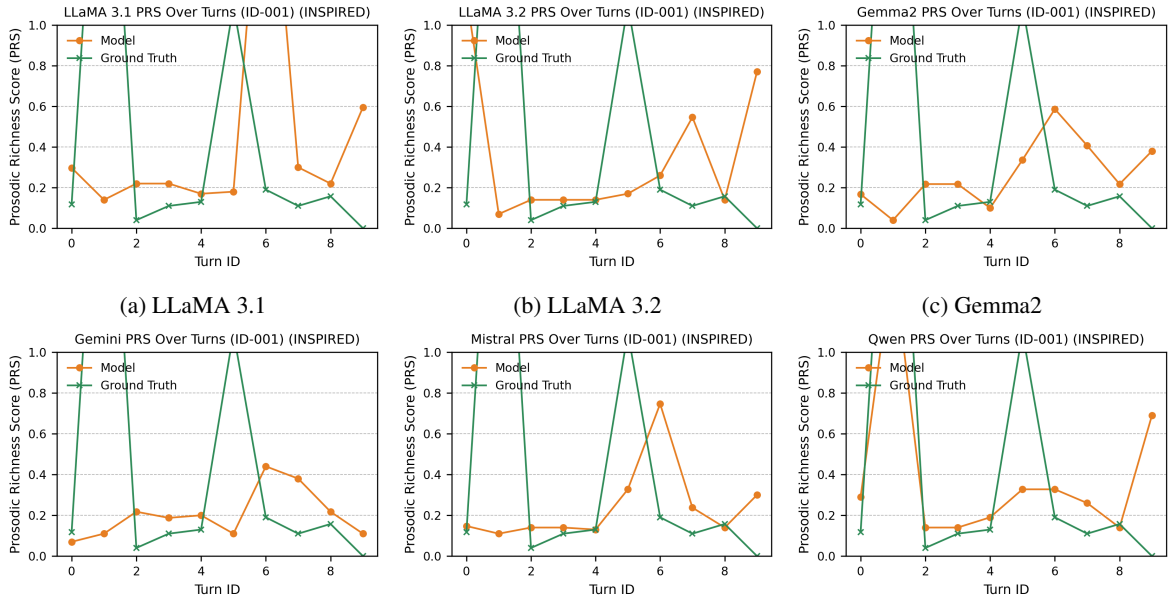
To support the main analysis in Section 4.2, we present turn-level PRS plots for individual conversations from each dataset. These visualizations illustrate how stylistic expressiveness evolves across dialogue turns for different LLMs in specific interaction contexts. Figures 6 and 7 correspond to ConvID 22709 (ReDial) and ConvID 001 (INSPIRED), respectively. Each subplot compares the model's PRS progression with ground-truth references, revealing variations in stylistic consistency.



(a) LLaMA 3.1 (b) LLaMA 3.2 (c) Gemma2

(d) Gemini (e) Mistral (f) Qwen

Figure 6: Turn-level PRS progression for a single ReDial conversation (ConvID: 22709), illustrating how stylistic expressiveness varies across dialogue turns for different LLMs. This per-conversation analysis highlights model-specific differences in maintaining prosodic richness.



(a) LLaMA 3.1 (b) LLaMA 3.2 (c) Gemma2

(d) Gemini (e) Mistral (f) Qwen

Figure 7: Turn-level PRS progression for a single INSPIRED conversation (ConvID: 001), illustrating how stylistic expressiveness evolves across dialogue turns for different LLMs. This case-specific analysis highlights model differences in preserving prosodic richness within emotionally expressive interactions.