Emotionally Aware or Tone-Deaf? Evaluating Emotional Alignment in LLM-Based Conversational Recommendation Systems

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

Abstract

Recent advances in Large Language Models (LLMs) have enhanced the fluency and coherence of Conversational Recommendation Systems (CRSs), yet emotional intelligence remains a critical gap. In this study, we systematically evaluate the emotional behavior of six state-of-the-art LLMs in CRS settings using the ReDial and INSPIRED datasets. We propose an emotion-aware evaluation framework incorporating metrics such as Emotion Alignment, Emotion Flatness, and per-emotion F1-scores. Our analysis shows that most models frequently default to emotionally flat or mismatched responses, often misaligning with user affect (e.g., joy misread as neutral). We further examine patterns of emotional misalignment and their impact on user-centric qualities such as personalization, justification, and satisfaction. Through qualitative analysis, we demonstrate that emotionally aligned responses enhance user experience, while misalignments lead to loss of trust and relevance. This work highlights the need for emotion-aware design in CRS and provides actionable insights for improving affective sensitivity in LLM-generated recommendations.

1 Introduction

Conversational Recommendation Systems (CRSs) aim to provide personalized recommendations through interactive dialogue (Jannach and Chen, 2022), but often lack emotional intelligence—the ability to understand and respond to user emotions. While LLMs have improved CRS fluency and contextual relevance (Zheng et al., 2023; Zhang et al., 2024), their affective awareness remains underexplored. Emotional alignment is vital for user trust and satisfaction (Pezenka et al., 2024), yet LLM-generated responses frequently exhibit emotional flatness or mismatches (Lechner et al., 2023), reducing conversational quality. Prior work has focused on intent and personalization (Lu et al., 2021; Zhou et al., 2022), with limited attention to emotional grounding. With the advent of LLMs, recent CRS architectures have adopted generative paradigms (Zhang et al., 2024; Feng et al., 2023). While systems like ChatCRS (Li et al., 2025) improve task goal guidance in LLM-based CRS, they do not explicitly evaluate emotional alignment in response generation. To address this gap, we conduct a systematic study of emotional behavior in LLM-generated CRS responses. We investigate how well LLMs align their emotional tone with that of the user, and whether their responses demonstrate sufficient emotional variability across conversation turns. Our contributions are threefold: (1) We propose an evaluation framework for emotional alignment, flatness, and diversity in CRS; (2) We apply it across six LLMs on ReDial and INSPIRED to systematically assess affective behavior; (3) We analyze misalignment cases and their impact on personalization, justification, and user satisfaction.

2 Related Work

Emotion modeling in dialogue systems has gained importance with the rise of LLMs, yet remains underexplored in CRSs. Early CRSs used modular pipelines with template-based responses (Li et al., 2018; Chen et al., 2019), while recent LLM-based systems like ChatCRS (Zhang et al., 2024; Feng et al., 2023; Li et al., 2025) offer greater fluency but often rely on synthetic supervision and shallow emotion alignment, and thus still lack deep emotional grounding. In open-domain dialogue, datasets like EmpatheticDialogues (Rashkin et al., 2019) and models such as MoEL (Lin et al., 2019), MIME (Majumder et al., 2020), and EmpDG (Li et al., 2020) emphasize generating affective responses, yet such approaches are rarely applied in CRS settings. Recent studies reveal that even advanced LLMs struggle with emotional alignment and flatness (Wang et al., 2023; Li et al., 2024),

Proceedings of the 9th Widening NLP Workshop, pages 167–174

especially in task-oriented contexts. Our work builds on these insights by evaluating emotional alignment and flatness in LLM-generated CRS responses using the ReDial dataset, addressing a critical gap in affect-aware recommendation dialogue research.

3 Conversational Recommendation Task Definition

We define a conversational recommendation system (CRS) as a dialogue agent that interacts naturally with users and provides recommendations during conversation. Formally, given multi-type context data—including the conversation history $\mathcal{C}_i = \{u_1, s_1, \ldots, u_t, s_t\}$ for user i and a knowledge graph $G = (V, E)$ consisting of entities $v \in V$ (e.g., movies, actors, genres) and relationships $(v_i, r, v_j) \in E$ connecting entities via relation types r—the CRS iteratively performs the following at each turn $t+1$: (1) Recommend a set of items $\mathcal{I}_{t+1} \subseteq V$ for user u_i, and (2) Generate a contextually coherent system response s_{t+1} based on the conversation history $\mathcal{C}_i$ and the knowledge graph G. The generated response s_{t+1} is appended to the conversation history $\mathcal{C}_i$ for subsequent turns, enabling iterative recommendation and dialogue generation.

4 Methodology

We propose an evaluation framework to assess emotional intelligence in LLM-based CRS responses, focusing on alignment, flatness, and diversity. All user, ground truth, and model-generated utterances are annotated using a transformer-based classifier fine-tuned on GoEmotions (Demszky et al., 2020) and EmpatheticDialogues (Rashkin et al., 2019), assigning one of seven emotion labels: `joy`, `sadness`, `anger`, `fear`, `surprise`, `disgust`, or `neutral` based on Ekman's taxonomy (Ekman, 1992). The reliability of these annotations was confirmed via manual verification on a random subset of utterances, showing substantial agreement with the programmatically assigned labels. To assess the emotional behavior of LLMs in CRS settings, we conduct the following evaluations:

1) Emotion Alignment: Measures how well the model's response emotion matches the user's expressed emotion:

$$\frac{\#\text{Emotion Matches}}{\#\text{Total Turns}} \times 100 \tag{1}$$

2) Emotional Flatness: Measures the variability in the model's emotional expressions using Shannon entropy H, computed as:

$$H = -\sum_{i=1}^{n} p(e_i) \log p(e_i) \tag{2}$$

where $p(e_i)$ is the proportion of responses labeled with emotion e_i, and n is the number of distinct emotion classes. We normalize H by dividing it by $\log_2(n)$.

3) Emotion Diversity: Per-emotion F1 scores measure how accurately the model generates responses that reflect each emotion category. F1 is the harmonic mean of precision and recall between the ground-truth emotion labels (from ReDial/INSPIRED) and the predicted labels assigned to LLM responses. High per-emotion F1 indicates that the model not only aligns with human references but also covers a range of emotions effectively.

4) llustrative Cases of Emotion Divergence: We also examine representative misalignment cases to interpret affective breakdowns in interaction quality.

User Emotion	Model Emotion	Comment
Joy	Neutral	Missed positive sentiment.
Surprise	Fear	Misread as threat or anxiety.
Sadness	Joy	Feels insensitive or dismissive.
Anger	Neutral	Lacks empathetic tone.
Fear	Disgust	Misinterprets concern.
Disgust	Sadness	Softens user's frustration.
Neutral	Joy	Overly enthusiastic tone.

Table 1: Common emotional misalignment patterns between user and model responses.

Thus, our evaluation framework offers a systematic approach to quantifying and analyzing emotional intelligence in LLM-driven CRS, revealing key affective gaps and guiding future improvements in empathetic conversational recommendation.

5 Datasets and Experimental Setup

To evaluate the emotional and conversational capabilities of LLMs in a CRS context, we describe the task, datasets, model integration, and inference settings used in our experiments.

5.1 Datasets

We evaluate the CRS on two benchmark datasets: (1) ReDial (Li et al., 2018), which contains 11,348

movie recommendation dialogues, split into 10,006 for training and 1,342 for testing; and (2) INSPIRED (Hayati et al., 2020), consisting of 1,001 emotionally rich, persona-driven dialogues, split into 801 for training and 200 for testing.

5.2 Model Integration

The evaluated models[1] are integrated into a unified CRS pipeline by replacing the Natural Language Generation (NLG) module. Each dialogue turn is processed as a triplet consisting of the user query, ground-truth response, and LLM-generated response for emotional evaluation. While the current experiments cover models from different series, exploring multiple sizes within the same series could provide additional insights into the impact of model scale.

5.3 Inference Settings

All models are accessed via official APIs through HuggingFace implementations. To ensure comparability, we fix decoding parameters across all six LLMs: greedy decoding (temperature = 0.0) and maximum output length = 128 tokens. No additional sampling strategies were applied. We adapt the prompt templates from our earlier study (Parmar and Mazumdar, 2025) on LLM response generation for conversational recommendation. While we use the same prompts and setup for consistency, the evaluation metrics and analyses in this work are different. Full prompt templates are provided in Appendix A. While LLM responses can vary with different prompt formulations, we keep the prompts consistent across all models to focus on differences arising from model behavior rather than input variations.

6 Emotional Performance Analysis of LLMs in CRS

We systematically evaluate the emotional intelligence of LLM-generated CRS responses across five key questions, focusing on alignment, diversity, flatness, misalignment, and impact on response quality. We use the ground-truth responses from ReDial and INSPIRED as human baselines to contextualize LLM performance. Exact numerical values for these baselines are not reported, as our focus is on comparing relative trends across models. Nonetheless, they represent realistic human conversational behavior, allowing qualitative interpretation of alignment, diversity, and flatness. The reliability of the emotion labels was confirmed via manual verification on a random subset of 150 user utterances from ReDial and INSPIRED, showing substantial agreement with the programmatically assigned labels (accuracy = 0.79, macro-F1 = 0.77), ensuring the annotations are suitable for analysis. Our analysis assesses affective sensitivity and expressive range.

6.1 RQ1: How accurately do LLMs align their emotional tone with user emotions?

To assess the emotional sensitivity of LLMs in conversational recommendation, we compute Emotion Alignment Accuracy (Equation 1) using two complementary criteria: Exact Match, requiring a direct match between the user's and model's emotions, and Coarse Match, which categorizes emotions into broader affective groups (positive, negative, neutral). These metrics quantify the extent to which model responses appropriately reflect the user's emotional state. As shown in Figure 1 (page 4), Mistral achieves the highest coarse emotion alignment, followed by LLaMA 3.2 and Gemini, with Qwen performing moderately.

6.2 RQ2: Do LLMs exhibit emotional flatness in their generated responses?

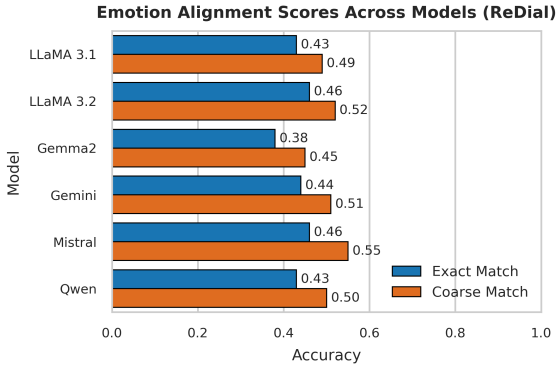
Flat emotional expression can make conversations feel dull, robotic, and disconnected. To assess this, we compute the Emotion Flatness Score (Equation 2), based on Shannon Entropy. Higher scores indicate richer affective variation, while lower scores suggest emotional flatness—often due to repetitive or default use of *neutral*. This metric reveals whether models sustain affective dynamics or lapse into monotonous tones.

As shown in Table 2, Gemma2 exhibits the highest normalized flatness scores, indicating greater emotional diversity. In contrast, LLaMA 3.1 and Gemini show the lowest scores across both datasets, suggesting flatter and more monotonous emotional distributions. For detailed distribution patterns, refer to Figure 2.

6.3 RQ3: How do LLMs differ in emotional expressiveness and alignment?

Capturing a wide range of user emotions is essential for creating engaging and empathetic con-

[1] `llama-3.1-8b-instant`, `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)



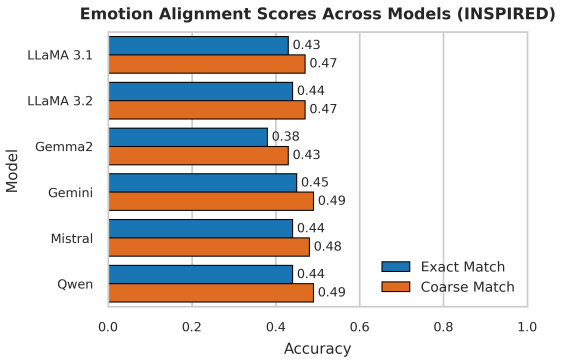


Figure 1: Emotion Alignment accuracy (%) of six LLMs on **ReDial** (left) and **INSPIRED** (right). Higher values indicate stronger alignment between user and model response emotions.

Model	Entropy (H)	Normalized Score
LLaMA 3.1 (R)	1.325	0.472
LLaMA 3.1 (I)	1.273	0.453
LLaMA 3.2 (R)	1.432	0.510
LLaMA 3.2 (I)	1.253	0.485
Gemma2 (R)	1.648	0.587
Gemma2 (I)	1.567	0.558
Gemini (R)	1.521	0.542
Gemini (I)	1.328	0.473
Mistral (R)	1.423	0.507
Mistral (I)	1.441	0.513
Qwen (R)	1.512	0.539
Qwen (I)	1.290	0.460

Table 2: Emotion flatness scores (R: ReDial, I: INSPIRED) range from 0 to 1, where lower values indicate less emotional variation (flatness) and higher values reflect greater emotional diversity.

versational experiences. Despite strong language capabilities, LLMs vary in emotional sensitivity—some models respond empathetically, while others default to neutral tones. To assess affective breadth, we compute per-emotion F1 scores using a fine-tuned emotion classifier. As shown in Figure 2, models like Qwen and LLaMA 3.2 achieve higher scores across emotions such as *joy* and *fear*, whereas Gemini and Mistral show limited expressiveness, especially for intense emotions. These differences highlight which models better preserve emotional nuance in CRS.

6.4 RQ4: What common misalignment patterns occur in LLM responses?

To improve affective grounding in CRS, it is crucial to understand where LLMs fail in emotion interpretation. Emotion confusion matrices (Figures 3 and 4 in Appendix B) reveal systematic misalignment patterns, such as frequent mapping of *joy* to *neutral* or *surprise* to *fear*, reflecting limitations in handling subtle or infrequent emotions. Here, misclassification refers to incorrectly recognizing the user's emotion, while misalignment refers to generating a response that does not appropriately reflect the recognized user emotion. Notably, models like LLaMA 3.1 and Gemini exhibit the most frequent misclassifications, whereas Gemma2 and Qwen achieve higher emotional alignment by more consistently generating responses that match the user's emotional state. These observations underscore the importance of model-specific sensitivity for affect-aware CRS design.

6.5 RQ5: Does emotional alignment enhance personalization, justification, and satisfaction?

Emotionally appropriate responses contribute to enhanced personalization, contextual justification, and user engagement in conversational recommendation. To understand the consequences of emotional misalignment, we conduct a qualitative analysis of dialogue turns exhibiting affective mismatches.

#	Emotion (User → Model)	User Utterance	Model Response
1	neutral → fear	"Any movie suggestions?"	"You might not like it, but it's intense."
2	surprise → neutral	"I didn't expect to love that movie!"	"Okay. I'll keep that in mind."
3	joy → neutral	"That movie made me so happy!"	"Thanks for letting me know."
4	neutral → joy	"Can you suggest something with good reviews?"	"Oh wow, you'll absolutely *love* this one!"
5	joy → neutral	"Such a fun thriller! Loved every second."	"I found it kind of disturbing and dark…"

Table 3: Examples of emotional misalignment in CRS dialogues. Note: These examples are constructed hypothetically to illustrate typical misalignments.

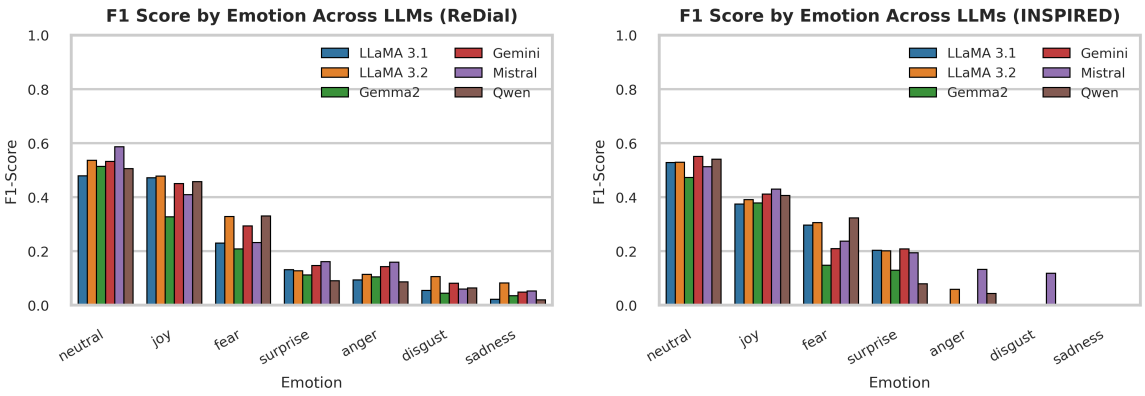


Figure 2: Emotion Diversity (F1 Score) across six LLMs on **ReDial** (left) and **INSPIRED** (right).

As shown in Table 3 and discussed in RQ4, mismatches such as *surprise* to *neutral* or *joy* to *fear* reduce perceived relevance, trust, and empathetic quality in CRS responses. These affective mismatches ultimately impair the overall interaction quality.

Discussion and Insights An interesting insight emerges when comparing emotional flatness (Table 2) and per-emotion F1 scores (Figure 2): there exists a non-trivial relationship between emotional diversity (RQ2) and emotion-specific expressiveness (RQ3). For instance, Gemma exhibits the highest emotional expressiveness according to flatness metrics, while LLaMA 3.1 and Gemini appear emotionally monotonous, as they are biased towards *neutral* or *joy* (not so emotionally distributed). However, this trend is not consistently reflected in Figure 2. LLaMA 3.2, for example, demonstrates strong F1 scores across multiple emotion categories, suggesting high emotional expressiveness, which seems at odds with its flatness score. Conversely, despite its high flatness score, Gemma achieves relatively lower F1 scores across several emotions. This divergence indicates that emotional diversity, as captured by entropy, does not always translate to accurate or contextually appropriate emotional expression. These findings highlight the need to jointly interpret flatness and emotion-specific performance metrics when evaluating affective behavior in LLM-based CRS systems.

In summary, our study reveals key affective limitations in LLM-driven CRS. Models frequently struggle with emotion alignment (RQ1) and exhibit flat emotional profiles (RQ2), with notable variation in affective sensitivity across models (RQ3). Systematic misalignments (RQ4) and their adverse impact on response quality (RQ5) underscore the need for improved emotional grounding in future CRSs.

7 Limitations

This work focuses on evaluating emotional alignment in LLM-generated CRS responses using automatic analyses. Human evaluation of user experience has not been conducted yet and is left for future studies.

8 Conclusion

This work presents an emotion-aware evaluation framework for analyzing LLM responses in conversational recommendation settings. By annotating model and user utterances with emotion labels, we assessed emotional alignment, diversity, and misalignment patterns across ReDial and INSPIRED datasets. Our findings reveal that while some models demonstrate moderate emotional sensitivity, many tend to default to neutral responses, resulting in flat and affectively misaligned conversations. These results underscore the need for integrating emotional intelligence into CRSs to foster more engaging and empathetic user experiences.

Acknowledgment

This work is supported by the Anusandhan National Research Foundation (ANRF), Department of Science and Technology, Government of India under project number CRG/2023/003741.

References

Rohan Anil, Yiding Jiang, and 1 others. 2024. Gemma: Lightweight, state-of-the-art open models. https://ai.google.dev/gemma.

Qibin Chen, Junyang Lin, Yichang Zhang, Ming Ding, Yukuo Cen, Hongxia Yang, and Jie Tang. 2019. Towards knowledge-based recommender dialog system. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1803–1813, Hong Kong, China. Association for Computational Linguistics.

Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. 2020. Goemotions: A dataset of fine-grained emotions. *arXiv preprint arXiv:2005.00547*.

Paul Ekman. 1992. An argument for basic emotions. *Cognition & emotion*, 6(3-4):169–200.

Yue Feng, Shuchang Liu, Zhenghai Xue, Qingpeng Cai, Lantao Hu, Peng Jiang, Kun Gai, and Fei Sun. 2023. A large language model enhanced conversational recommender system. *arXiv preprint arXiv:2308.06212*.

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.

Dietmar Jannach and Li Chen. 2022. Conversational recommendation: A grand ai challenge. *AI Magazine*, 43(2):151–163.

Yujia Jiang, Guillaume Lample, and 1 others. 2024. Mistral 7b. https://mistral.ai/news/mistral-7b/.

Fabian Lechner, Allison Lahnala, Charles Welch, and Lucie Flek. 2023. Challenges of gpt-3-based conversational agents for healthcare. *arXiv preprint arXiv:2308.14641*.

Chuang Li, Yang Deng, Hengchang Hu, Min-Yen Kan, and Haizhou Li. 2025. ChatCRS: Incorporating external knowledge and goal guidance for LLM-based conversational recommender systems. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 295–312, Albuquerque, New Mexico. Association for Computational Linguistics.

Ming Li, Yusheng Su, Hsiu-Yuan Huang, Jiali Cheng, Xin Hu, Xinmiao Zhang, Huadong Wang, Yujia Qin, Xiaozhi Wang, Kristen A Lindquist, and 1 others. 2024. Language-specific representation of emotion-concept knowledge causally supports emotion inference. *iScience*, 27(12).

Qintong Li, Hongshen Chen, Zhaochun Ren, Pengjie Ren, Zhaopeng Tu, and Zhumin Chen. 2020. EmpDG: Multi-resolution interactive empathetic dialogue generation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4454–4466, Barcelona, Spain (Online). International Committee on Computational Linguistics.

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)*.

Zhaojiang Lin, Andrea Madotto, Jamin Shin, Peng Xu, and Pascale Fung. 2019. MoEL: Mixture of empathetic listeners. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 121–132, Hong Kong, China. Association for Computational Linguistics.

Yu Lu, Junwei Bao, Yan Song, Zichen Ma, Shuguang Cui, Youzheng Wu, and Xiaodong He. 2021. RevCore: Review-augmented conversational recommendation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1161–1173, Online. Association for Computational Linguistics.

Navonil Majumder, Pengfei Hong, Shanshan Peng, Jiankun Lu, Deepanway Ghosal, Alexander Gelbukh, Rada Mihalcea, and Soujanya Poria. 2020. MIME: MIMicking emotions for empathetic response generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8968–8979, Online. Association for Computational Linguistics.

Darshna Parmar and Pramit Mazumdar. 2025. Measuring prosodic richness in llm-generated responses for conversational recommendation. In *Proceedings of GlobalNLP 2025: Beyond English—Natural Language Processing for All Languages in an Era of Large Language Models (RANLP 2025 Workshop)*, 12th September 2025. Workshop paper, not available online.

Ilona Pezenka, Lili Aunimo, Gerald Janous, and David Dobrowsky. 2024. Emotionality in task-oriented chatbots–the effect of emotion expression on chatbot perception. *Communication Studies*, 75(6):825–843.

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.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Y-Lan Boureau, and 1 others. 2023. Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:2307.09288.

Xuena Wang, Xueting Li, Zi Yin, Yue Wu, and Jia Liu. 2023. Emotional intelligence of large language models. *Journal of Pacific Rim Psychology*, 17:18344909231213958.

Xiaoyu Zhang, Ruobing Xie, Yougang Lyu, Xin Xin, Pengjie Ren, Mingfei Liang, Bo Zhang, Zhanhui Kang, Maarten de Rijke, and Zhaochun Ren. 2024. Towards empathetic conversational recommender systems. In *Proceedings of the 18th ACM Conference on Recommender Systems*, pages 84–93.

Zhonghua Zheng, Lizi Liao, Yang Deng, and Liqiang Nie. 2023. Building emotional support chatbots in the era of llms (2023). *arXiv preprint arXiv:2308.11584*.

Yuanhang Zhou, Kun Zhou, Wayne Xin Zhao, Cheng Wang, Peng Jiang, and He Hu. 2022. C^2-crs: Coarse-to-fine contrastive learning for conversational recommender system. In *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining*, pages 1488–1496.

Appendix

A Prompts Used for LLM Response Generation

A.1 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.

A.2 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 Emotion Confusion Matrices (RQ4)

To identify patterns of emotional misalignment, we present emotion confusion matrices for all six LLMs on ReDial (Figure 3) and INSPIRED (Figure 4). These visualizations provide a fine-grained diagnostic view of model behavior, highlighting systematic confusions that are not captured by aggregate alignment scores. Notably, several models struggle with subtle or context-dependent emotions, pointing to limitations in affective reasoning that warrant further attention.

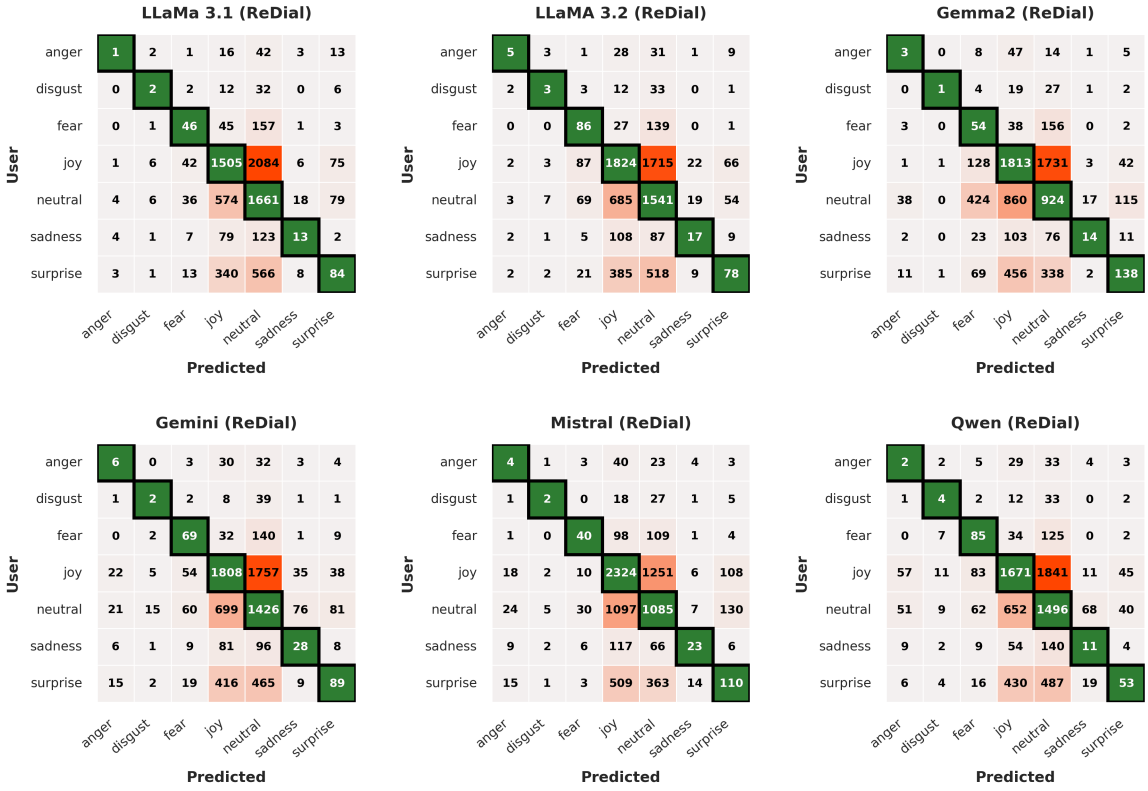


Figure 3: Emotion confusion matrices for **ReDial**. Rows denote user-expressed emotions; columns denote model-predicted emotions. Diagonal entries indicate correct alignment.

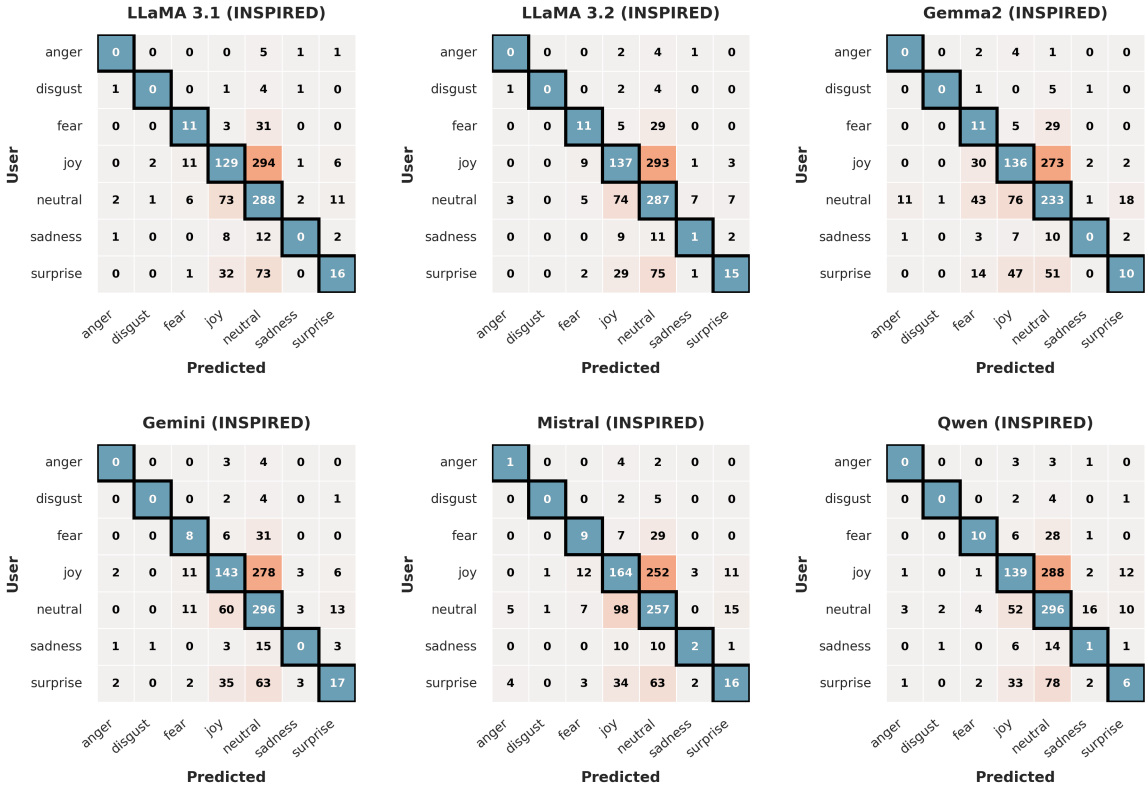


Figure 4: Emotion confusion matrices for **INSPIRED**. Diagonal entries represent correct predictions; off-diagonal cells reveal common misclassifications.