

Multimodal Emotion-Aware Preference Modeling for Personalized Dialogue Generation

Ravi Melhatre

Department of Computer Science, University of New Hampshire, Durham, NH, USA.
ravi.work@unh.edu

Lucas Crawford

Department of Electrical Engineering and Computer Science, University of Missouri,
Columbia, MO, USA.
crawfordlucas@missouri.edu

Arthur Fex

Department of Computer Science, George Mason University, Fairfax, VA, USA.
afox@gmu.edu

Abstract

The growing integration of conversational agents into healthcare, education, customer service, and daily productivity has exposed the limits of purely text-based personalization. Users evaluate dialogue systems not only on factual accuracy but also on emotional attunement, interaction style, and continuity of preferences across sessions. This paper presents a system-level examination of multimodal emotion-aware preference modeling for personalized dialogue generation. We discuss how acoustic, visual, and linguistic affect signals can be combined with explicit and implicit preference signals to shape response generation without sacrificing architectural modularity or operational control. The analysis first establishes the foundations of affective computing and the requirements of user-state awareness in dialogue systems. It then develops a reference architecture that includes modality-specific encoders, cross-modal emotion recognition, preference adapters, and generation policies supported by parameter-efficient adaptation. Structural trade-offs among latency, interpretability, memory footprint, personalization granularity, and multimodal fusion depth are explored. The paper further addresses robustness, fairness, privacy, governance, and sustainability. Rather than advocating a single monolithic model, the discussion emphasizes modular integration, auditability, and staged deployment. The work contributes an interdisciplinary systems perspective for designing personalized dialogue pipelines that remain responsive to emotional context while adhering to responsible AI principles and practical infrastructure constraints.

Keywords

affective computing, multimodal emotion recognition, personalized dialogue generation, preference modeling, conversational AI, responsible AI, system architecture.

1. Introduction

Human conversation is multimodal, context-dependent, and emotionally regulated. Early dialogue systems optimized for task completion or response likelihood, often ignoring the affective dimension. The introduction of self-attentive sequence modeling [1] and large-scale pretraining [2] altered this landscape by enabling contextual response generation at scale. In parallel, parameter-efficient adaptation methods [3][4] have made it feasible to specialize

large models for individual users, domains, and interaction styles without retraining them from scratch. Dialogue-oriented pretraining [5] further demonstrated that conversational corpora can encode social and stylistic regularities. These advances create conditions for personalized systems, but they also expose a key gap: user preferences are not static textual properties. They shift with mood, conversational goals, and social context. Addressing this gap requires treating preference as a multimodal and emotion-sensitive construct.

Research on empathetic dialogue [6] and emotional chatting machines [7] showed that explicit emotional signals improve perceived response appropriateness. Multimodal emotion resources [8][9] and attention-based fusion models [10] have enabled systems to infer affect from speech, facial expression, and words. However, many existing systems treat emotion recognition and preference modeling as separate stages. A more robust design integrates them into a unified personalization layer that conditions generation on both the user’s current affective state and a learned preference profile. This paper examines the system-level consequences of such integration, with attention to architectural structure, resource constraints, governance, and long-term sustainability.

2. Foundations and System Requirements

Preference modeling in dialogue systems can be approached through supervised signals, comparative judgments, or interaction-level reward learning. Recent work on aligning language models with human feedback [11] demonstrates that pairwise and ranking-based preference data can be converted into policies that shape generation. Yet this approach brings systemic risks when models absorb biases from interaction data and become difficult to audit [12]. Affective computing provides theoretical grounding for emotion-aware systems [13]. Basic emotion theories [14] support categorical emotion labels, while dimensional and appraisal frameworks capture emotional intensity and valence. Multimodal affect detection research [15] shows that combining modalities improves reliability under noise, but also increases complexity. Nonverbal channels contribute substantially to perceived meaning [16], although voice and face data are both sensitive and identifiable. Ethical risk frameworks for language models [17] highlight the need for context-specific harm assessment before deployment.

From a systems perspective, emotion-aware preference modeling must satisfy several requirements. First, the system must estimate user state continuously without requiring explicit questionnaires. Second, it must represent preferences in a form that can influence generation without overriding factual correctness or user agency. Third, the system must operate under latency and privacy constraints that vary across deployment contexts. Fourth, the personalization layer must be isolable so that a faulty preference inference cannot contaminate all downstream responses. These requirements imply a modular architecture in which emotion recognition, preference estimation, and generation are separately observable yet tightly coordinated. They also imply that preference models should be updated through interaction rather than fixed at initialization.

3. Multimodal Emotion-Aware Preference Modeling Architecture

A reference architecture consists of four subsystems: multimodal emotion extraction, preference state modeling, policy adaptation, and controllable generation. Preference state modeling can be approached as a comparison task in which the system learns to rank candidate responses under a given emotional context. Recent dual-stage transformer work with gated fusion [18] illustrates how modality-specific preference signals can be integrated

through gated mechanisms to improve pairwise preference prediction in dialogue systems. Foundation model ecosystems [19] further allow such modular components to be composed across pretrained encoders, adapters, and reward models. For speech, self-supervised speech representations [20] provide robust acoustic embeddings without costly handcrafted features.

The architecture must decide where fusion occurs. Early fusion combines raw or lightly processed features from text, audio, and vision before preference estimation. Late fusion processes each modality independently and combines representations at the decision level. Hybrid fusion uses intermediate representations, cross-modal attention, or gated aggregation to preserve modality-specific structure while enabling interaction. Early fusion offers simplicity but can be brittle when a modality is missing or noisy. Late fusion supports modularity and easier debugging but may miss cross-modal dependencies such as sarcasm, where verbal content conflicts with prosody. Hybrid approaches are more expressive but introduce additional synchronization and alignment challenges. The choice depends on deployment constraints and the reliability of each modality.

Personalization can be implemented through parameter-efficient adapters attached to a frozen pretrained language model. These adapters may encode user-level style, topic preferences, and emotion regulation tendencies. A key design decision is whether preference representations are stored as explicit vectors, as learned tokens, or as low-rank updates. Explicit vectors are easier to audit and transfer across sessions. Learned tokens are flexible but less interpretable. Low-rank updates reduce memory use and make it practical to serve many users from one base model. However, they can create conflicts when multiple preferences interact. The system should therefore distinguish stable preferences, such as formality or preferred explanation depth, from dynamic emotional states that should not be permanently stored.

Generation control is another architectural concern. The dialogue generator must respond to current emotion signals without overfitting to a single affective label. For example, a user expressing frustration may benefit more from a calm, structured response than from a mirroring response that amplifies negative affect. Preference models can rank candidate generations, adjust decoding behavior, or provide control signals to the generator. The ranking approach is attractive because it keeps the generator unchanged and allows preference policies to be updated independently. The control-signal approach permits finer-grained modulation but can reduce stability if not carefully regularized. A staged deployment strategy, where ranking models are introduced before direct generation control, reduces risk and improves observability.

4. Robustness and Fairness in Affective Dialogue Systems

Multimodal systems are susceptible to missing modalities, sensor noise, cultural variation in emotional expression, and demographic bias. Robustness evaluation should therefore include corrupted inputs, partial occlusions, background noise, and cross-dataset transfer. Existing robustness benchmarks [21] provide a useful starting point for stress testing models under distributional shifts, although affective dialogue systems require additional evaluation focused on social appropriateness and emotional calibration. A model that remains accurate on clean validation data may still respond inappropriately when emotional signals are ambiguous. Therefore, uncertainty estimation and abstention mechanisms should be designed into the preference layer.

Fairness is equally critical because emotion recognition and preference modeling can encode demographic and cultural assumptions. Expression of anger, sadness, or enthusiasm varies across individuals and communities. A system that interprets subdued affect as low engagement may systematically misread users from cultures with different display rules. Similarly, preference models trained on interaction logs may inherit historical biases in response selection. Architectural remedies include separate group-specific calibration layers, adversarial debiasing, and audit interfaces that allow system operators to inspect preference inferences. However, technical interventions alone are insufficient. Fairness must be monitored throughout the lifecycle, with ongoing review of how multimodal signals are collected, annotated, and transformed into model behavior.

5. Governance, Privacy, and Policy Implications

The integration of voice, face, and language data in personalized dialogue systems raises substantial governance challenges. AI governance frameworks [22] emphasize accountability, transparency, and societal benefit as central principles. These principles translate into concrete requirements for data minimization, purpose limitation, user consent, and the right to delete or correct personal preference data. Voice and facial representations are particularly sensitive because they are biometric and can be used for identification. Systems should avoid storing raw multimodal signals when summary affect features are sufficient. Preference profiles should be treated as personal data and protected by the same controls as user communications.

Policy development must also account for context-specific risks. A healthcare dialogue system may require stricter oversight than a general-purpose assistant because emotional misrecognition can influence clinical decisions. Educational systems must protect minors and avoid overpersonalization that limits learning opportunities. Consumer-facing systems should provide clear notice when emotion recognition is active and allow users to disable it without losing core functionality. Such controls are not only compliance measures; they also improve user trust and long-term adoption. Governance mechanisms should include periodic audits, incident reporting channels, and independent evaluation of affective performance across demographic groups.

6. Deployment, Sustainability, and Infrastructure Considerations

Deployment of multimodal emotion-aware dialogue systems involves trade-offs between cloud and edge processing. Cloud deployment offers higher model capacity and easier updates but introduces latency and privacy risks. Edge processing reduces data movement and improves responsiveness but constrains model size and energy budget. A practical architecture may therefore use lightweight speech and facial feature extraction at the edge, while preference fusion and large-scale generation run in the cloud. This split can be adjusted according to application sensitivity and available hardware. Such staged infrastructure also supports graceful degradation when a modality becomes unavailable.

Sustainability is another central concern. Emotion recognition models, preference adapters, and large language generators each contribute to energy consumption. The cumulative cost increases when systems are replicated for many users. Parameter sharing, distillation, quantization, and selective activation can reduce inference cost. Caching of common preference states and session-level reuse of speech embeddings can further lower compute load. From an organizational perspective, sustainability metrics should be part of system evaluation rather than treated as an afterthought. A personalized dialogue system is only

viable at scale if its infrastructure remains affordable, energy-efficient, and compatible with long-term environmental goals.

The interdisciplinary nature of this design problem requires collaboration among machine learning engineers, human-computer interaction researchers, privacy lawyers, and domain experts. Conversational systems are not merely technical artifacts; they are socio-technical infrastructures that shape how people express emotion, seek information, and form expectations about technology. A system that performs well in offline benchmarks may fail in deployment if users find its emotional interpretations intrusive or unreliable. Therefore, iterative evaluation with diverse user populations should accompany architectural development. This includes qualitative assessments of perceived empathy, trust, and autonomy, which are not captured by standard accuracy metrics.

7. Conclusion

Multimodal emotion-aware preference modeling offers a pathway toward dialogue systems that are more responsive to the full communicative context of human interaction. However, realizing this potential requires more than improving emotion recognition accuracy. It requires a deliberate system-level design that aligns multimodal fusion, preference adaptation, generation control, privacy safeguards, and deployment infrastructure. This paper has examined these dimensions through an interdisciplinary systems lens. A reference architecture was presented in which emotion recognition and preference modeling are modular yet coordinated, enabling auditability and incremental deployment. Structural trade-offs were analyzed across early, late, and hybrid fusion strategies, as well as between ranking-based and control-based personalization approaches. Robustness and fairness considerations were discussed in relation to noisy modalities, cultural variability, and demographic bias. Governance and policy implications were framed around privacy, consent, and accountability for biometric data. Deployment analysis emphasized the value of edge-cloud partitioning and energy-aware design. The central argument is that personalized dialogue generation should not be built as a monolithic model focused only on response quality. It should be built as a governed, adaptive, and auditable system that respects user agency while responding sensitively to emotional context. Future work should develop standardized evaluation protocols that combine technical metrics with measures of perceived empathy, fairness, and long-term user well-being.

References

1. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems* 30 (pp. 5998–6008). Curran Associates.
2. Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 4171–4186). Association for Computational Linguistics.
3. Houlisby, N., Giurui, A., Jastrzebski, S., Morrone, B., de Laroussilhe, Q., Gesmundo, A., Attariyan, M., & Gelly, S. (2019). Parameter-efficient transfer learning for NLP. In *Proceedings of the 36th International Conference on Machine Learning* (pp. 2790–2799). PMLR.

4. Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., & Chen, W. (2022). LoRA: Low-rank adaptation of large language models. In International Conference on Learning Representations.
5. Zhang, Y., Sun, S., Galley, M., Chen, Y.-C., Brockett, C., Gao, X., Gao, J., Liu, J., & Dolan, B. (2020). DIALOGPT: Large-scale generative pre-training for conversational response generation. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations (pp. 270–278). Association for Computational Linguistics.
6. Rashkin, H., Smith, E. M., Li, M., & Boureau, Y.-L. (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 (pp. 5370–5381). Association for Computational Linguistics.
7. Zhou, H., Huang, M., Zhang, T., Zhu, X., & Liu, B. (2018). Emotional chatting machine: Emotional conversation generation with internal and external memory. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 32, No. 1, pp. 730–738). AAAI Press.
8. Busso, C., Bulut, M., Lee, C.-C., Kazemzadeh, A., Mower, E., Kim, S., Chang, J. N., Lee, S., & Narayanan, S. S. (2008). IEMOCAP: Interactive emotional dyadic motion capture database. *Language Resources and Evaluation*, 42(4), 335–359.
9. Poria, S., Hazarika, D., Majumder, N., Naik, G., Cambria, E., & Mihalcea, R. (2019). MELD: A multimodal multi-party dataset for emotion recognition in conversations. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (pp. 527–536). Association for Computational Linguistics.
10. Zadeh, A., Liang, P. P., Poria, S., Vij, P., Cambria, E., & Morency, L.-P. (2018). Multi-attention recurrent network for human communication comprehension. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 32, No. 1, pp. 5642–5649). AAAI Press.
11. Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., ... Lowe, R. (2022). Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems* 35 (pp. 27730–27744). Curran Associates.
12. Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (pp. 610–623). ACM.
13. Picard, R. W. (1997). *Affective computing*. MIT Press.
14. Ekman, P. (1992). An argument for basic emotions. *Cognition and Emotion*, 6(3-4), 169–200.
15. D'Mello, S. K., & Kory, J. (2015). A review and meta-analysis of multimodal affect detection systems. *ACM Computing Surveys*, 47(3), Article 43, 1–36.
16. Mehrabian, A. (1968). Communication without words. *Psychology Today*, 2(4), 53–56.

17. Weidinger, L., Mellor, J., Rauh, M., Griffin, C., Uesato, J., Huang, P.-S., ... Gabriel, I. (2021). Ethical and social risks of harm from language models. arXiv preprint arXiv:2112.04359.
18. Shih, K., Deng, Z., Chen, X., Zhang, Y., & Zhang, L. (2025, May). DST-GFN: A Dual-Stage Transformer Network with Gated Fusion for Pairwise User Preference Prediction in Dialogue Systems. In 2025 8th International Conference on Advanced Electronic Materials, Computers and Software Engineering (AEMCSE) (pp. 715-719). IEEE.
19. Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., ... Liang, P. (2021). On the opportunities and risks of foundation models. arXiv preprint arXiv:2108.07258.
20. Baevski, A., Zhou, H., Mohamed, A., & Auli, M. (2020). wav2vec 2.0: A framework for self-supervised learning of speech representations. In Advances in Neural Information Processing Systems 33 (pp. 12449–12460). Curran Associates.
21. Hendrycks, D., & Dietterich, T. (2019). Benchmarking neural network robustness to common corruptions and perturbations. In International Conference on Learning Representations.
22. Floridi, L., & Cowls, J. (2019). A unified framework of five principles for AI in society. Harvard Data Science Review, 1(1). <https://doi.org/10.1162/99608f92.8cd550d1>