Emotion-Aware Conversational Agents for Mental Health Support: Leveraging NLP-Based Self-Reflection and Personalized Dialogue Generation

Roy M. Ferguson
Department of Electrical Engineering and Computer Science, University of Missouri, Columbia, MO, USA.
roy.m.ferguson@missouri.edu

Jereimy Bheambers
Department of Computer Science, University of Alabama at Birmingham, Birmingham, AL, USA.
jeremyc@uab.edu

Abstract

The escalating global burden of mental health disorders has precipitated an urgent need for scalable, accessible, and effective support mechanisms. Emotion-aware conversational agents have emerged as a promising digital intervention, offering continuous, non-stigmatizing, and personalized mental health assistance. This paper presents a system-level investigation into the design, deployment, and governance of such agents, focusing on the integration of natural language processing (NLP) techniques for self-reflection and adaptive dialogue generation. We examine the layered architecture that combines emotion recognition, user modeling, and context-aware response synthesis, highlighting structural trade-offs among latency, accuracy, privacy, and interpretability. Particular attention is given to self-reflection mechanisms, wherein agents guide users through structured, introspective conversations by analyzing linguistic patterns in real time. We discuss how transformer-based language models enable personalized dialogue generation that adapts to evolving affective states and user histories while maintaining therapeutic alignment. The analysis extends to infrastructure considerations, including cloud-edge orchestration, model compression, and continuous monitoring, as well as robustness against adversarial perturbations and distributional shifts. Fairness, accountability, and transparency are addressed through the lens of bias mitigation in training data, differential privacy, and explainable AI interfaces. Furthermore, we explore the policy implications of deploying mental health chatbots within fragmented regulatory landscapes and advocate for harmonized frameworks that balance innovation with clinical safety. The paper concludes by outlining forward-looking pathways for sustainable, equitable, and empathetic conversational systems in mental healthcare, underscoring the need for interdisciplinary collaboration across engineering, clinical psychology, ethics, and public policy.

Keywords

emotion-aware conversational agents, mental health, natural language processing, self-reflection, personalized dialogue generation, fairness, robustness, ethical AI, system architecture, governance.

1. Introduction

The global mental health crisis, exacerbated by socioeconomic stressors, the COVID-19 pandemic, and persistent gaps in care provision, has catalyzed a reexamination of how support services are delivered at scale. Traditional face-to-face therapy, while effective, remains inaccessible to large segments of the population due to cost, stigma, geographic disparities, and a chronic shortage of trained professionals. In this landscape, conversational agents designed for mental health support have attracted considerable attention from researchers, clinicians, and technology developers. These systems, ranging from simple scripted bots to sophisticated neural dialogue agents, promise always-available, low-barrier emotional support. Randomized controlled trials evaluating automated conversational agents such as Woebot have demonstrated preliminary efficacy in reducing symptoms of depression and anxiety among young adults [1]. Similarly, real-world data from the empathy-driven Wysa chatbot indicate that users find value in the nonjudgmental and immediate interactions these agents provide [2]. These findings anchor the rationale for deeper system-level inquiry into the architectures, algorithms, and governance mechanisms that underpin effective emotion-aware conversational systems.

The technological trajectory of mental health chatbots intersects with significant advances in natural language processing, affective computing, and personalized machine learning. Yet deploying such systems in high-stakes settings demands far more than assembling state-of-the-art language models. Developers must reconcile competing design goals: therapeutic fidelity versus conversational spontaneity, personalization versus privacy, and system scalability versus clinical accountability. Research on NLP for mental health has underscored the rich psychological signals embedded in non-clinical texts, including linguistic markers of depression, anxiety, and cognitive distortions [3], while integrative reviews of social media analysis have established the feasibility of inferring mental states from language at scale [4]. These capabilities provide foundational inputs for emotion-aware agents, but translating passive detection into ethically grounded, proactive dialogue generation remains a complex systems challenge.

This paper offers a comprehensive systems analysis of emotion-aware conversational agents for mental health support, with an emphasis on two pivotal components: NLP-based self-reflection and personalized dialogue generation. Self-reflection here denotes the agent-guided process of helping users examine their thoughts, emotions, and behaviors through natural language interaction, a mechanism rooted in cognitive behavioral therapy and mindfulness traditions. Personalized dialogue generation extends this by dynamically tailoring linguistic style, content depth, and conversational pacing to the individual's momentary affective state and longitudinal history. We argue that effective integration of these capabilities requires a carefully orchestrated architecture spanning emotion perception, user state tracking, reflective prompting, and safe response generation, all while operating within the constraints of real-time interaction and regulatory compliance. The paper further addresses infrastructure design, robustness testing, fairness auditing, and the evolving policy frameworks that will shape the next generation of mental health chatbots.

2. Foundational Technologies and Related Work

The present generation of conversational agents draws heavily on the transformer architecture introduced by Vaswani et al., which has revolutionized sequence modeling through self-attention mechanisms [7]. Large-scale pretrained language models such as BERT [5] and GPT-2 [6] have enabled unprecedented fluency and contextual coherence in text generation, forming the backbone of many modern dialogue systems. While general-domain chatbots like

XiaoIce illustrate the potential for sustained, emotionally resonant conversation with millions of users [8], their adaptation to mental health contexts necessitates specialized training, safety guardrails, and domain-specific evaluation rubrics. Empathetic dialogue research, specifically, has produced benchmarks and models that explicitly condition responses on emotional cues, moving beyond mere topic relevance toward affective alignment [9].

In the healthcare domain, conversational agents have been rapidly deployed for pandemic response and psychological triage, often leveraging pre-existing frameworks and content libraries [10]. A systematic review of conversational agents in healthcare highlights their growing prevalence but also identifies persistent gaps in evidence quality, user engagement sustainability, and safety reporting [11]. Mental health chatbots face unique challenges including the detection of crisis states, the management of therapeutic ruptures, and the prevention of over-reliance on automated support, challenges that demand rigorous architectural planning and continuous post-deployment monitoring [12]. While these systems leverage advances in affective computing and sentiment analysis, most existing implementations lack deep self-reflection capabilities, operating instead through scripted psychoeducation or superficial empathetic mirroring. The missing element is a theoretically grounded mechanism that not only recognizes what a user feels but guides them toward exploring why they feel that way, thereby facilitating genuine cognitive restructuring.

3. System Architecture and Design Considerations

Designing an emotion-aware conversational agent for mental health requires a layered, modular architecture that separates concerns while enabling tight feedback loops. At the lowest layer, speech or text input undergoes preprocessing, tokenization, and initial feature extraction, feeding into an emotion recognition module that estimates dimensional affect (valence, arousal) and discrete emotional categories from linguistic and, potentially, acoustic or facial features. This affective signal then informs a user state tracker that maintains a dynamic profile encompassing current emotional state, conversation history, therapeutic goals, and known user attributes such as demographic information and clinical history, where available. The state representation is passed to a dialogue manager that selects or generates the next conversational act, which may be an empathetic reflection, a Socratic question to promote self-reflection, or a personalized cognitive restructuring exercise.

A critical architectural trade-off lies in the choice between retrieval-based and generative response strategies. Retrieval-based systems select predefined responses from a curated library, offering reliability, safety, and low latency but limited conversational diversity. Generative models, particularly those fine-tuned on domain-specific corpora, can produce novel and contextually rich utterances but risk generating hallucinated or clinically inappropriate content. Hybrid architectures that combine both approaches are increasingly common, using retrieval for high-risk situations such as expressed suicidal ideation and generation for low-risk, exploratory dialogue. The system must also integrate a safety layer that monitors both user inputs and agent outputs for harmful content, employing classifiers and rule-based filters that can escalate to human supervisors when necessary. Real-time performance requirements, often demanding sub-second response latency for natural conversational flow, impose constraints on model size and server inference capacity, pushing deployment toward model distillation, quantization, and specialized hardware acceleration.

The architecture's self-reflection subsystem deserves particular attention. It must access the user state tracker to decide when to deploy reflective prompts, typically when the user expresses negative affect or contradictory statements indicative of cognitive distortions. This

subsystem relies on a dedicated NLP pipeline capable of identifying linguistic markers of absolutist thinking, overgeneralization, and emotional reasoning. Once triggered, it selects from a repertoire of reflective strategies, such as asking the user to describe alternative interpretations or to recall past successes that challenge the negative belief. The subsystem's effectiveness hinges on the quality of the underlying language understanding and the ability to personalize the reflection style to the user's communication preferences, a topic explored further in Section 4.

4. NLP-Based Self-Reflection Mechanisms

Self-reflection through conversational interaction demands that the agent not only comprehend surface-level user sentiment but also identify deeper cognitive patterns and gently challenge maladaptive thinking. NLP provides the technical means to operationalize therapeutic reflection at scale. Transformer-based encoder models fine-tuned on mental health corpora can detect linguistic markers of depression, including increased first-person singular pronoun use, absolutist words, and reduced lexical diversity, and these features serve as triggers for the reflective dialogue branch. The agent then generates prompts that are deliberately open-ended, avoiding prescriptive advice and instead inviting the user to articulate connections between thoughts, emotions, and behaviors. For instance, upon detecting a pattern of self-critical language, the agent might respond with a reflective question that encourages the user to consider what they would say to a friend in the same situation, a classic cognitive restructuring technique.

The design of such mechanisms must account for individual differences in readiness for self-reflection and in language style. Some users respond well to direct, structured questioning, while others benefit from a more narrative, storytelling approach. The work by Solanki et al. [13] underscores the importance of self-concept representations in shaping how people process information about themselves, suggesting that reflective interventions are most effective when they align with the user's existing self-schema. An NLP-enabled agent can infer elements of the user's self-concept from longitudinal dialogue data and adjust the framing of reflective prompts accordingly. For example, a user who consistently uses achievement-oriented language might be guided to reflect on self-worth outside of performance metrics. This alignment is technically achieved through a combination of topic modeling, sentiment history, and personal pronoun analysis, which collectively build a dynamic, vectorized representation of self-concept.

A further challenge is maintaining coherent and productive self-reflection dialogue over multiple turns without explicit therapist oversight. Dialogue state tracking must be extended to include a reflection depth metric that measures how deeply the user has engaged with a reflective thread, allowing the agent to decide when to deepen the exploration, when to consolidate insights, and when to transition back to supportive listening. This requires fine-grained natural language inference and commonsense reasoning to validate that the user's responses genuinely address the reflective prompt rather than deflecting or ruminating unproductively. Reinforcement learning techniques have been explored to optimize reflective dialogue policies, using rewards that combine user self-reported insight, linguistic indicators of cognitive change, and session completion rates. However, the specification of safe reward functions in mental health remains an open research problem, as maximizing engagement can inadvertently encourage emotional venting without resolution.

5. Personalized Dialogue Generation

Personalized dialogue generation elevates the conversation beyond generic empathy by adapting every aspect of the agent's linguistic output to the individual user. This personalization encompasses lexical choice, syntactic complexity, discourse style, emotional tone, and even metaphor preferences. Modern large language models, when conditioned on user-specific prompts and conversation history, can generate highly tailored responses. Personalization is informed by an evolving user model that integrates static attributes such as age and cultural background with dynamic features including recent emotional trajectory, stated preferences, and inferred personality traits. The model may learn that a particular user responds more positively to humorous deflections during mild stress but requires serious, structured support during acute distress, and adapt its tone accordingly.

Technically, personalized generation can be realized through a combination of prompt engineering, adapter modules, and retrieval-augmented approaches. Adapter layers inserted into pretrained transformer models allow for efficient per-user fine-tuning without catastrophic forgetting, enabling the model to capture idiosyncratic language usage and emotional expression patterns. Retrieval-augmented generation empowers the system to pull relevant therapeutic content from a curated knowledge base, such as a specific cognitive behavioral therapy worksheet or a mindfulness exercise, and weave it seamlessly into the conversation in a personal manner. The generation process is further constrained by therapeutic fidelity filters that ensure the personalized output does not deviate from evidence-based principles or introduce harmful suggestions. This filtering is particularly critical when the user's history includes self-harm, trauma, or psychosis, where even well-intentioned personalization may inadvertently trigger distress.

The system-level trade-off between personalization depth and privacy is acute. Rich personalization demands collecting and processing sensitive data, including mental health histories, emotional diaries, and intimate conversational logs. While on-device learning and federated architectures can mitigate centralization risks, they also complicate model evaluation and fairness auditing. Personalization algorithms themselves can introduce disparities if certain demographic groups are underrepresented in the training data, leading to lower-quality experiences for marginalized users. Therefore, personalization modules must be subjected to rigorous subgroup performance testing and fairness constraints, ensuring that the benefits of tailored support are equitably distributed. The ongoing debate over whether mental health chatbots should adapt to user personality versus gently steer users toward healthier communication styles embodies the deeper ethical tension between accommodation and intervention.

6. Infrastructure, Scalability, and Deployment

Deploying emotion-aware conversational agents at population scale requires a robust infrastructure that balances computational efficiency, data governance, and service reliability. Cloud-native architectures built on container orchestration platforms allow dynamic scaling of inference services to meet fluctuating demand, which in mental health contexts often exhibits diurnal and seasonal patterns. However, real-time latency requirements and concerns about data sovereignty increasingly motivate hybrid architectures where sensitive NLP tasks such as emotion recognition and reflective prompt generation are performed on-device or at the edge, while computationally intensive model updates and aggregate analytics occur in the cloud. Edge deployment reduces network dependency and enhances privacy but introduces device heterogeneity challenges, necessitating model compression techniques such as knowledge

distillation, quantization, and pruning without unacceptable degradation in therapeutic accuracy.

Continuous integration and continuous deployment (CI/CD) pipelines for mental health dialogue systems must incorporate extensive safety and quality gates. Unlike typical software, updates to language model weights, prompt templates, or safety classifiers can fundamentally alter the agent's therapeutic persona and risk profile. A rigorous A/B testing framework that includes clinical supervision, adversarial scenario simulation, and user feedback loops is essential to monitor drift in conversational quality and empathetic accuracy. Monitoring systems must detect not only technical failures such as latency spikes or out-of-memory errors but also clinical signals such as increased user distress, topic avoidance, or dangerous command misinterpretation. Logging and alerting mechanisms must be designed to comply with health data protection regulations, often requiring pseudonymization and strict access controls.

Sustainability is an emerging infrastructure concern. Training and serving large language models consume substantial energy, raising questions about the carbon footprint of widely deployed mental health chatbots. Techniques such as sparse attention, model recycling, and green data center utilization can mitigate environmental impact, yet the tension between ever-larger models for better emotional nuance and the imperative for sustainable computing remains unresolved. The mental health domain, with its inherently high social value, offers a compelling case for prioritizing efficient model architectures and evaluating performance per watt alongside traditional accuracy metrics. Lifecycle assessment of such systems should become standard practice, especially for publicly funded deployments.

7. Fairness, Robustness, and Ethical Governance

Fairness in emotion-aware conversational agents is multifaceted, encompassing demographic parity in emotion recognition accuracy, equitable quality of personalized dialogue across user populations, and unbiased handling of sensitive topics. Emotion classification models trained predominantly on data from specific linguistic, racial, or cultural groups may misinterpret affective expressions from underrepresented populations, leading to inappropriate or unhelpful responses [14]. Mitigation strategies include balanced data collection, adversarial debiasing during training, and post-hoc calibration of model outputs. Beyond technical fairness, procedural fairness demands transparency in how the agent makes reflective or therapeutic decisions, enabling users to understand and, where appropriate, contest the system's behavior. Explainable artificial intelligence (XAI) techniques, such as attention visualization and concept-based explanations, are being adapted to convey why the agent chose a particular reflective question or emotional framing, although the challenge of rendering such explanations accessible to non-experts persists.

Robustness against adversarial inputs and out-of-distribution scenarios is paramount. Users in distress may express themselves in atypical, fragmented, or metaphor-laden language that can confuse standard NLP models, while deliberate adversarial attempts to elicit harmful outputs must be defended against. Techniques such as adversarial training, randomized smoothing, and input perturbation monitoring have been applied to dialogue systems to improve resilience [20]. However, robustness in the mental health context extends beyond technical perturbations to include resilience against real-world ambiguities such as sarcasm, cultural idioms, and code-switching, failure to handle which could rupture the therapeutic alliance. Continuous collection of edge cases and integration of human-in-the-loop feedback into model retraining cycles are essential to maintain robustness over time.

Ethical governance frameworks must address the unique position of mental health chatbots as neither purely software products nor regulated medical devices. The absence of clear global standards has led to divergent approaches, with some jurisdictions classifying them as medical devices subject to stringent clinical evaluation, and others treating them as wellness apps with minimal oversight. A unified governance model would incorporate key pillars: clinical safety validation, informed consent that accurately conveys the agent's capabilities and limitations, crisis escalation protocols with seamless transfer to human care, data stewardship that exceeds minimum regulatory requirements, and independent auditing for bias and effectiveness [15]. Institutional review boards, ethics committees, and multidisciplinary advisory panels should be integral to the development lifecycle, not gatekeepers at the end. The deployment of mental health chatbots in low-resource settings further demands participatory design methods that center community values and avoid digital colonialism.

8. Policy Implications and Sustainability

The policy landscape for digital mental health interventions is fragmented and reactive, lagging behind the pace of technological change. Current regulatory instruments, such as the FDA's digital health precertification program and the EU's Medical Device Regulation, are beginning to adapt but provide limited guidance on the unique challenges posed by large language models that generate open-ended, non-deterministic therapeutic content [17]. Policymakers must grapple with liability allocation when an agent's personalized response contributes to a negative clinical outcome, a question that implicates developers, clinical supervisors, healthcare organizations, and the users themselves. We argue for a tiered regulatory framework that escalates requirements in proportion to clinical risk: low-risk wellness agents may operate with transparency disclosures and basic safety mechanisms, while high-risk therapeutic agents must undergo pre-market clinical trials, post-market surveillance, and algorithmic auditability akin to those for medical devices.

Data governance policies are equally critical. Mental health conversations contain extraordinarily sensitive information that, if breached, could lead to discrimination, stigma, or psychological harm. The principles of data minimization, purpose limitation, and user-controlled data portability must be enshrined in the system architecture from the outset, supported by technical mechanisms such as differential privacy and secure multi-party computation [15]. Synthetic data generation offers a promising avenue for training and evaluating mental health dialogue systems without exposing real user data, though careful validation is required to ensure that synthetic corpora faithfully represent the linguistic and emotional diversity of target populations [24]. International cross-border data flows further complicate compliance, as mental health data may be subject to multiple, conflicting legal regimes. Harmonization efforts through organizations such as the World Health Organization's digital health governance initiative are essential to enable global scalability while protecting vulnerable individuals.

Long-term sustainability of emotion-aware conversational agents depends on more than funding and infrastructure. It requires building trust with users, clinicians, and regulators through consistent, evidence-based performance and transparent reporting of limitations. Business models that rely on engagement metrics or advertising are fundamentally misaligned with therapeutic goals, and alternative models such as outcome-based reimbursement by health systems, public funding, or nonprofit service provision must be explored. Sustainability also encompasses workforce development: the clinicians, data scientists, and

ethicists who design and oversee these systems need interdisciplinary training that bridges computational and psychological disciplines. Academic-industry partnerships, open-source toolkits for safety evaluation, and publicly available benchmarks can accelerate progress while preventing the concentration of power in a few technology conglomerates. Ultimately, the integration of emotion-aware agents into mental healthcare should be viewed not as a replacement for human therapists but as a complementary layer in a stepped-care model, where automation handles routine monitoring and psychoeducation, freeing human clinicians to focus on complex cases and therapeutic relationships.

9. Conclusion

Emotion-aware conversational agents represent a transformative frontier in mental health support, combining advances in NLP, affective computing, and personalized machine learning to deliver scalable, empathetic, and self-reflective interactions. This paper has examined the system architecture, core mechanisms of self-reflection and personalized dialogue, infrastructure demands, and the intertwined ethical and policy challenges that define this rapidly maturing field. We have argued that the integration of NLP-based self-reflection moves these agents beyond superficial sentiment mirroring toward genuinely therapeutic engagement, while personalized dialogue generation enhances user relevance and long-term engagement. However, the deployment of such systems is fraught with structural trade-offs: the push for personalization must be balanced against privacy; the desire for conversational fluidity must not compromise clinical safety; and the ambition for scale must be accompanied by rigorous fairness and robustness guarantees. Future progress will depend on interdisciplinary collaboration that embeds clinical wisdom into technical design, on governance frameworks that evolve with technological capabilities, and on a steadfast commitment to centering the well-being of the individuals who entrust these systems with their most vulnerable moments. As the field moves forward, cautious optimism and systematic evaluation must walk hand in hand, ensuring that emotionally intelligent machines serve as a genuine force for psychological healing and social equity.

References

1. Fitzpatrick, K. K., Darcy, A., & Vierhile, M. (2017). Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (Woebot): A randomized controlled trial. JMIR Mental Health, 4(2), e19.

2. Inkster, B., Sarda, S., & Subramanian, V. (2018). An empathy-driven, conversational artificial intelligence agent (Wysa) for digital mental well-being: Real-world data evaluation mixed-methods study. JMIR mHealth and uHealth, 6(11), e12106.

3. Calvo, R. A., Milne, D. N., Hussain, M. S., & Christensen, H. (2017). Natural language processing in mental health applications using non-clinical texts. Natural Language Engineering, 23(5), 649–685.

4. Guntuku, S. C., Yaden, D. B., Kern, M. L., Ungar, L. H., & Eichstaedt, J. C. (2017). Detecting depression and mental illness on social media: An integrative review. Current Opinion in Behavioral Sciences, 18, 43–49.

5. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. Proceedings of the 2019

Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 4171–4186.

6. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. OpenAI Blog.

7. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. Advances in Neural Information Processing Systems, 30.

8. Zhou, L., Gao, J., Li, D., & Shum, H. Y. (2020). The design and implementation of XiaoIce, an empathetic social chatbot. Computational Linguistics, 46(1), 53–93.

9. Rashkin, H., Smith, E. M., Li, M., & Boureau, Y. L. (2019). Towards empathetic open-domain conversation models: A new benchmark and dataset. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 5370–5381.

10. Miner, A. S., Laranjo, L., & Kocaballi, A. B. (2020). Chatbots in the fight against the COVID-19 pandemic. npj Digital Medicine, 3(1), 65.

11. Laranjo, L., Dunn, A. G., Tong, H. L., Kocaballi, A. B., Chen, J., Bashir, R., Surian, D., Gallego, B., Magrabi, F., Lau, A. Y. S., & Coiera, E. (2018). Conversational agents in healthcare: A systematic review. Journal of the American Medical Informatics Association, 25(9), 1248–1258.

12. Denecke, K., Abd-Alrazaq, A., Househ, M., & Bewick, B. M. (2021). Artificial intelligence for chatbots in mental health: Opportunities and challenges. In Artificial Intelligence in Medicine (pp. 115–129). Springer.

13. Solanki, D., Hsu, H. M., Zhao, O., Zhang, R., Bi, W., & Kannan, R. (2020, July). The way we think about ourselves. In International Conference on Human-Computer Interaction (pp. 276-285). Cham: Springer International Publishing.

14. Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. ACM Computing Surveys, 54(6), 1–35.

15. Torous, J., & Roberts, L. W. (2017). Needed innovation in digital health and smartphone applications for mental health: Transparency and trust. JAMA Psychiatry, 74(5), 437–438.

16. Ram, A., Prasad, R., Khatri, C., Venkatesh, A., Gabriel, R., Liu, Q., Nunn, J., Hedayatnia, B., Cheng, M., Nagar, A., King, E., Bland, K., Wartick, A., Pan, Y., Song, H., Jayadevan, S., Hwang, G., & Pettigrue, A. (2018). Conversational AI: The science behind the Alexa Prize. arXiv preprint arXiv:1801.03604.

17. Topol, E. J. (2019). High-performance medicine: The convergence of human and artificial intelligence. Nature Medicine, 25(1), 44–56.

18. Deriu, J., Rodrigo, A., Otegi, A., Echegoyen, G., Rosset, S., Agirre, E., & Cieliebak, M. (2021). Survey on evaluation methods for dialogue systems. Artificial Intelligence Review, 54, 755–810.

19. Floridi, L., & Cowls, J. (2019). A unified framework of five principles for AI in society. Harvard Data Science Review, 1(1).

20. Jia, R., & Liang, P. (2017). Adversarial examples for evaluating reading comprehension systems. Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, 2021–2031.

21. Poria, S., Cambria, E., Bajpai, R., & Hussain, A. (2017). A review of affective computing: From unimodal analysis to multimodal fusion. Information Fusion, 37, 98–125.

22. Zhang, Y., Sun, S., Galley, M., Chen, Y. C., Brockett, C., Gao, X., Gao, J., Liu, J., & Dolan, B. (2020). Personalizing dialogue agents: I have a dog, do you have pets too? Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2204–2214.

23. Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (2020). Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Information Fusion, 58, 82–115.

24. El Emam, K., Mosquera, L., & Hoptroff, R. (2020). A method for managing re-identification risk from synthetic data derived from health-related population data. Journal of Medical Internet Research, 22(10), e19410.

25. Strubell, E., Ganesh, A., & McCallum, A. (2019). Energy and policy considerations for deep learning in NLP. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 3645–3650.

25. Zhou, D. (2026, May). A Code Visualization Graph-Based Method for Vulnerability Severity Assessment Using a Multi-Scale Feature Fusion Network. In 2026 3rd International Conference on Image Processing and Artificial Intelligence (ICIPAI) (pp. 306-309). IEEE.