A Modular Prototype of Emotion-Aware Proactive Voice Agent with Live2D Embodiment

Jae Young Suh[*1] **and Mingyu Jeon**[†1]

[1]*DELAB, MODULABS, Seoul, South Korea*

Abstract

We present a voice-based conversational agent that advances beyond reactive dialogue by integrating speech-to-text transcription with Whisper, emotion recognition, simple policy mechanisms, and Live2D embodiment. The system delivers supportive guidance either as inline prompts or card-style recommendations, while empathetic dialogue and expressive avatar cues enhance both transparency and user engagement. A log-based evaluation across ten sessions showed consistent stability, with an average latency of 7.1 seconds. This prototype illustrates a practical foundation for developing emotion-aware, proactive companions aligned with the vision of human-centered dialogue systems.

Keywords: Proactive Conversational Agent, Emotion-Aware Interaction, Live2D Embodied Dialogue

1 Introduction

Large Language Models (LLMs) have achieved remarkable progress in recent years, enabling conversational agents to generate fluent, coherent, and contextually relevant dialogue across diverse domains [20, 2]. However, most deployed systems remain essentially reactive and respond to user queries without taking initiative. Recent surveys highlight the importance of proactive conversational agents that anticipate user needs, recommend relevant content, and foster more human-centered interaction [6, 7]. Benchmarks such as ProCIS formalize proactive retrieval tasks and measure both timing and contextual appropriateness [19]. In emotionally sensitive contexts, poorly timed or irrelevant suggestions can undermine trust, which underscores the role of affect as a critical cue for initiative [11].

Research has approached proactivity in dialogue from multiple directions. Some studies develop retrieval-augmented methods for anticipatory interventions [19], while others draw on mixed-initiative principles to balance helpfulness with user autonomy [11]. Human-centered perspectives emphasize that initiative should remain non-intrusive, particularly in affect-rich scenarios where excessive intervention risks harming trust and long-term engagement [7]. Emotion-aware dialogue systems provide a promising foundation because affect recognition enables empathetic and context-sensitive responses

[*]Corresponding author. Email: tjwodud04@gmail.com

[†]Email: jkmcoma7@gmail.com

[15]. Nevertheless, most approaches remain reactive and focus on expressing empathy after emotion is observed rather than leveraging affective cues as triggers for proactive support in real time.

This paper introduces a prototype of a voice-based, emotion-aware conversational agent embodied through a Live2D animated character. The system integrates Whisper for speech-to-text transcription [18], affect inference under a compact taxonomy, and proactive suggestion mechanisms governed by simple policy rules with cooldowns. Embodiment through a visual avatar is intended to enhance transparency and user engagement while mitigating the perceived intrusiveness of unsolicited interventions.

This work makes two key contributions. First, we present an implemented modular prototype that links emotion recognition with proactive conversational strategies in a voice-based, embodied setting. Second, we analyze how avatar embodiment mediates unsolicited initiative and identify design levers that preserve user autonomy while enriching affective human-AI interaction. By addressing the gap between reactive empathy and affect-driven proactive assistance, this study advances the development of more human-centered dialogue systems.

2 Related Work

2.1 Proactive Conversational Agents

Conversational AI has progressed from purely reactive systems toward proactive designs that anticipate needs and decide when to intervene. This shift recognizes that effective interaction requires both answering explicit questions and offering timely, context-aware suggestions. Prior syntheses organize core challenges and methods in retrieval, dialogue planning, and evaluation [6], while benchmarks formalize proactive retrieval and quantify timing and contextual appropriateness [19]. Building on this foundation, recent perspectives broaden taxonomies and identify open evaluation gaps [8], and modeling of agents' latent deliberation offers mechanisms for calibrating initiative in multi-party settings [13]. Together, these directions underscore proactivity as a coupling of computational techniques with human-centered control over when, why, and how an agent should act.

2.2 Emotion-Aware Dialogue Systems

Emotion is central to social interaction, and affect recognition has been shown to support empathetic, context-sensitive responses that can improve user experience [15, 17]. Beyond text-only sentiment, multimodal approaches incorporate speech and visual signals, and avatar-mediated interaction further situates dialogue in affective context [9]. Extending this trajectory, recent work demonstrates gains when prosodic and visual cues are explicitly fused, highlighting the value of speech-based affect for empathy and naturalness [22]. These advances motivate designs where emotion is not only acknowledged post hoc but also informs proactive support.

2.3 Embodied Conversational Interfaces

Embodiment augments dialogue by increasing social presence and clarifying intent, complementing the technical and affective dimensions of interaction. Early studies reported richer protocols and stronger presence than text-only systems [4]; subsequent work explored expressive avatars, social robots, and VTuber-style characters that integrate visual performance with conversational intelligence [14], and showed how representation and agency shape copresence and user experience [16]. More recent syntheses in extended-reality contexts examine design choices that influence presence and trust [21], and broader views of embodied agents argue that virtual and physical forms enable tighter perception–action loops than disembodied chat alone [10]. In concert, this literature frames embodiment as a means to render initiative more legible while mitigating perceived intrusiveness.

3 System Overview

3.1 System Workflow

The architecture integrates voice input, emotion recognition, policy evaluation, and embodied output within a modular pipeline. Figure 1 illustrates the end-to-end workflow, where speech is captured, transcribed by a Whisper-based ASR, analyzed for affect, and processed by the policy engine to determine whether proactive intervention should occur.

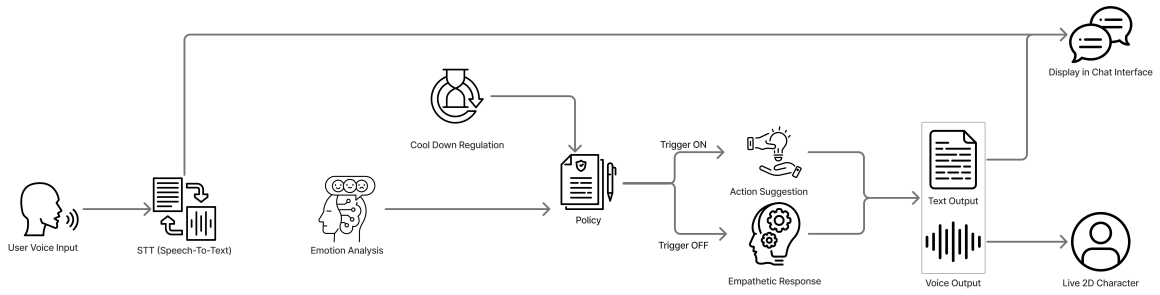

Figure 1: End-to-end workflow of the proposed system, showing the pipeline from speech input through transcription, emotion inference, policy evaluation, and proactive suggestion with embodiment.

Emotion inference follows a compact taxonomy inspired by the East Asian concept of the "Seven Emotions" (chiljeong, 七情) described in classical Confucian texts such as the *Book of Rites* and the *Doctrine of the Mean* [5]. This taxonomy was chosen for its cultural grounding and interpretability, offering a concise yet expressive set of categories that reflect everyday human experience. The module assigns confidence scores to each category and maintains a dynamic session state that continuously tracks both immediate cues and longer-term emotional trends throughout the interaction.

The policy engine evaluates this evolving state to decide when to intervene. A cooldown mechanism controls the frequency of proactive actions, ensuring that suggestions remain supportive rather than intrusive. When trigger conditions are met, the suggestion generator creates lightweight cards combining affect-sensitive dialogue with actionable resources, such as calming music or short videos. Each card includes plain text, brief rationale, and one or two external links, delivered as a structured payload.

The system output is returned to the client, where it can be displayed in the chat interface or converted into synthesized speech through a Live2D avatar. Embodiment reinforces initiative by aligning visible expressions with dialogue, allowing proactive actions to appear natural and socially grounded. The modular design emphasizes clarity, reproducibility, and extensibility, demonstrating how minimal yet well-defined components can operationalize proactive behavior in emotionally sensitive contexts.

Proactive Suggestion Policy

Beyond a simple cooldown heuristic, the policy combines hard guardrails (e.g., minimum intervals between suggestions, user-defined quiet hours) with a lightweight soft-scoring model that integrates affect intensity and conversational context. Only when the composite score exceeds a conservative threshold is a suggestion surfaced; otherwise, guidance remains inline. This layered design helps the system remain non-intrusive while maintaining interpretability and consistency.

3.2 User Interface

On the client side, the system is embodied through a Live2D character integrated into a web-based chat interface, as shown in Figure 2. Suggestion cards appear seamlessly within the dialogue bubble and are accompanied by affect-sensitive responses. Each card contains a concise title, a rationale, and one or two links for follow-up actions. By embedding these cards directly into the conversation flow, proactive interventions are perceived as integrated contributions rather than external interruptions.

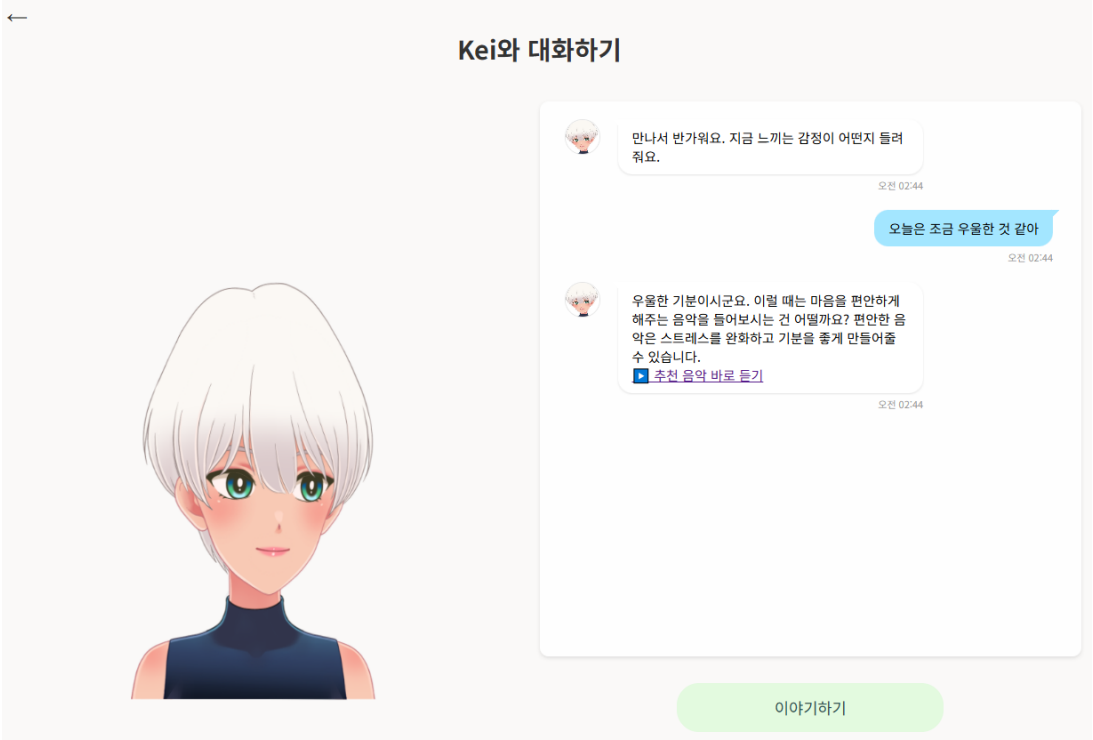

Figure 2: Client interface integrating Live2D embodiment with structured suggestion cards, illustrating proactive support within the conversational flow.

The Live2D avatar synchronizes lip movements and facial expressions with system responses, reinforcing the emotional intent of each suggestion. For example, calm or sympathetic expressions are paired with supportive dialogue. The avatar also enhances transparency by visually embodying the agent's initiative, reducing the impression of mechanical behavior. Together, visual embodiment and structured suggestion cards strengthen user trust, minimize intrusiveness, and encourage sustained engagement.

4 Preliminary System Evaluation

A lightweight log-based evaluation was conducted to assess the system's technical feasibility under demonstration conditions. In accordance with the workshop constraints, no user studies were performed. Instead, three objective aspects were analyzed: latency, proactive behavior, and stability. These indicators offer a transparent view of the system's operational performance without introducing subjective bias. Table 1 summarizes the overall results.

Under the current conservative thresholds, explicit suggestion cards were not triggered; instead, proactive support appeared as inline guidance with contextual links. This outcome indicates that the policy prioritized low-visibility initiative during short demo sessions, underscoring the need for threshold calibration in future iterations.

Table 1: Summary of log-based evaluation metrics across ten test sessions

Metric	Result	Notes
End-to-end latency	Avg. 7.1s (median 6.1s, 90th percentile 9.9s)	10 runs
Proactive outputs	Contextual responses with links, no separate card display	Observed in all runs
System stability	100%	All runs completed successfully

System latency, measured from the end of user speech to the display of the system response, averaged 7.1 seconds across ten sessions, with a median of 6.1 seconds and a 90th percentile of 9.9 seconds. Although longer than typical interactive systems, this delay reflects the cumulative processing of transcription, affect inference, policy evaluation, and rendering within a prototype environment.

Regarding proactive behavior, no standalone suggestion cards were issued. Under the existing threshold and cooldown configurations, the policy favored inline recommendations rather than separate card presentation. Nevertheless, the system consistently embedded context-appropriate suggestions (e.g., music links, breathing guidance, or short routines) within responses, demonstrating stable proactive behavior.

System stability was maintained throughout all ten test sessions without any critical failures, confirming that the modular and decoupled architecture reliably handled repeated interactions even under relatively high latency.

To illustrate how these outputs manifested in practice, Figure 3 presents an excerpt from the raw system log. The example captures a representative session in which the user expressed stress, and the system responded with empathetic dialogue accompanied by a supportive music link. Such instances connect the quantitative results above with concrete interaction traces.

Original Version


```
{
  "i": 1,
  "user_id": "u_mood",
  "session_id": "s_mood",
  "text": "조금 불안하고 답답해. 숨 고르는 법 알려줄래?",
  "ok": true,
  "elapsed_ms": 7609.66784399352,
  "error": "",
  "response": {
    "ai_text": "불안하고 답답한 마음, 정말 힘드시겠어요. 이럴 때는 편안한 음악을 들으며 마음을
    가라앉혀보는 건 어떨까요? 편안한 음악은 긴장을 완화하고 마음을 진정시키는 데 도움이 됩니다.
    <br><a href=\"https://www.youtube.com/watch?v=DWcJFNfaw9c\" target=\"_blank\">▶ 추천 음악
    바로 듣기</a>",
    "audio": "AAA=",
    "emotion_percent": {
      "낙": 0,
      "노": 0,
      "애": 20,
      "애(사랑)": 0,
      "오": 80,
      "욕": 0,
      "희": 0
    },
    "link": "https://www.youtube.com/watch?v=DWcJFNfaw9c",
    "proactive_card": null,
    "top_emotion": "오",
    "user_text": "조금 불안하고 답답해. 숨 고르는 법 알려줄래?"
  }
}
```


Translated Version


```
{
  "i": 1,
  "user_id": "u_mood",
  "session_id": "s_mood",
  "text": "I feel a bit anxious and suffocated. Could you teach me how to catch my breath?",
  "ok": true,
  "elapsed_ms": 7609.66784399352,
  "error": "",
  "response": {
    "ai_text": "Feeling anxious and suffocated must be really difficult. In times like this,
    how about calming yourself by listening to some relaxing music? Soothing music can help
    ease tension and bring peace of mind. <br><a href=\"https://www.youtube.com/watch?
    v=DWcJFNfaw9c\" target=\"_blank\">▶ Listen to the recommended music</a>)",
    "audio": "AAA=",
    "emotion_percent": {
      "Joy": 0,
      "Anger": 0,
      "Sorrow": 20,
      "Love": 0,
      "Dislike": 80,
      "Desire": 0,
      "Delight": 0
    },
    "link": "https://www.youtube.com/watch?v=DWcJFNfaw9c",
    "proactive_card": null,
    "top_emotion": "오",
    "user_text": "I feel a bit anxious and suffocated. Could you teach me how to catch my
    breath?"
  }
}
```


Figure 3: Excerpt from a system log showing user input, inferred emotion distribution, and generated response including supportive content.

While these findings do not directly capture user perception, they demonstrate that the prototype operated robustly and reliably under repeated trials, providing a clear foundation for refining how proactive behaviors can be surfaced and evaluated in future work.

5 Discussion

This demonstration shows that lightweight mechanisms, when combined with emotion recognition and embodiment, can effectively support proactive conversational behavior in practice. As outlined in the workflow, a modular integration of Whisper-based speech recognition, affect inference under a compact taxonomy, session state tracking, and policy evaluation enables the system to deliver supportive guidance both within responses and, when appropriate, through structured suggestion cards. The log-based evaluation further confirms feasibility, with sessions completing at 100% stability and an average end-to-end latency of 7.1 seconds. Taken together, these results indicate that a clear, decoupled architecture can provide contextually appropriate assistance under demonstration settings.

A primary strength of the system lies in its modular and interpretable design. Each component—from transcription to policy control and Live2D embodiment—can be independently replaced or extended. This flexibility allows for richer emotion models, adaptive policies, and improved retrieval strategies without major restructuring. The approach also aligns with human-centered and mixed-initiative design principles, which emphasize legibility of system intent, explicit control over initiation, and reproducibility of behavior [7, 11]. Furthermore, proactive retrieval formulations such as ProCIS offer a principled framework for measuring timing and contextuality, suggesting a standardized direction for future evaluation [19].

Embodiment contributes to reducing perceived intrusiveness while maintaining initiative. Prior research shows that avatars enhance social presence and enable richer interaction protocols beyond text-only communication [4]. Expressive interfaces, including VTuber-style characters and avatar-based collaboration, have been found to strengthen engagement and make unsolicited guidance feel more natural when paired with affective cues [14, 16]. In our prototype, synchronized facial expressions and voice delivery render supportive intent more understandable, aligning with findings that clarity and affective signaling are essential for the acceptance of proactive systems [7].

In interpreting the evaluation, explicit suggestion cards did not appear in the ten test runs. Instead, the system consistently embedded contextual recommendations directly into responses, such as music links, breathing guidance, or short routines. This suggests that current policy thresholds favor inline guidance over separate visual cards. Looking ahead, a mixed-initiative design could flexibly vary the form of proactive output: clarifying questions for low-risk entry points, inline suggestions for subtle affective cues, and structured cards for more salient needs [1, 11]. Such variations could be systematically examined using ProCIS-inspired protocols to quantify timing and contextual appropriateness [19].

Latency remains an important engineering concern. The observed delay reflects the cumulative cost of transcription, affect inference, policy evaluation, and rendering in a prototype environment. Future deployments could reduce this overhead through batching and token streaming, partial decoding with incremental TTS, optimized client-side VAD cutoff, lightweight affect models, and tighter client–server loops. In addition, retrieval-augmented methods could cache frequently used supportive resources to minimize external API calls while maintaining contextual relevance [12]. These optimizations complement the modular structure without sacrificing interpretability.

Policy design illustrates the balance between transparency and sophistication. Simple heuristics are easy to audit and explain but may overlook nuances in affective interaction. Human-centered design emphasizes maintaining user autonomy and providing clear rationales for unsolicited initiative [7, 11]. Learning-based strategies—from adaptive thresholds to long-horizon planning—can improve sensitivity and timing but require interpretability to sustain trust. A practical approach is to combine transparent rules with adaptive mechanisms: employing clarifying questions under uncertainty and gradually escalating support when affective and contextual cues align [1].

We also distinguish absolute latency (end-to-end) from perceived latency. While the measured delay averaged around seven seconds, token streaming and early delivery can reduce perceived waiting and preserve conversational flow. For affect inference, text-based analysis was chosen for simplicity;

future iterations will incorporate prosodic features such as pitch, tone, and rhythm to better capture emotional nuance. This trade-off represents a pragmatic balance between implementation feasibility and multimodal extensibility.

Finally, ethical and deployment considerations remain central. Proactive interventions should operate on an opt-in basis, respect interruptibility, and adopt privacy-preserving defaults. Prior analyses caution that proactive behavior can be perceived as intrusive when rationales or refusal mechanisms are lacking [3]. Our approach—combining embodiment with inline guidance—aims to balance helpfulness with autonomy. Future work will extend this approach through configurable disclosure, user-specific pacing, and local-first content options. As the system evolves toward real-world deployment, safeguards, standardized evaluation, and transparent reporting will be essential to ensure responsible development of proactive, emotion-aware agents [6].

6 Conclusion

This work presented a prototype of a proactive, emotion-aware conversational agent embodied with a Live2D avatar. The system integrates lightweight speech recognition, affect inference, and policy-based decision logic to demonstrate how relatively simple mechanisms, when carefully coordinated, can yield transparent and reproducible proactive behaviors. The modular architecture enables each component to be independently extended or replaced, allowing future incorporation of richer emotion models, adaptive policies, and advanced retrieval strategies.

Log-based evaluation verified the system's feasibility, showing consistent stability across all sessions. Although the end-to-end latency averaged slightly above seven seconds and explicit suggestion cards were not triggered, the agent reliably provided context-sensitive recommendations within its responses. These findings highlight both the practicality of the current design and key directions for refinement—namely, latency reduction, policy calibration, and adaptive adjustment for real-world use.

The discussion emphasized modularity, clarity, and embodiment as guiding principles, while underscoring ethical safeguards as essential for responsible scaling. Proactive interventions should remain interpretable, interruptible, and privacy-conscious. From this perspective, conversational assistants can be viewed not merely as reactive responders but as collaborative partners that anticipate user needs and support human goals in a socially grounded manner.

Although this demonstration focused on Korean, employing a culturally grounded affect taxonomy, the architecture itself is language-agnostic and can be localized with minimal modification. This prototype thus serves as a testbed for non-intrusive initiative, where embodiment and affective cues can be adapted across languages and cultures to enable human-centered and socially appropriate proactive AI.

In summary, the prototype demonstrates that proactive conversational agents can be developed in a practical and extensible way without reliance on overly complex pipelines. By combining emotion awareness with visual embodiment, it establishes a foundation for advancing dialogue systems that are both technically robust and ethically aligned with human-centered interaction.

Acknowledgments

This research was supported by Brian Impact Foundation, a non-profit organization dedicated to the advancement of science and technology for all.

Declaration on Generative AI

During the preparation of this work, the authors used large language models for grammar and style polishing and image tools for figure prototyping. After using these tools, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] Mohammad Aliannejadi, Hamed Zamani, Fabio Crestani, and W. Bruce Croft. Asking clarifying questions in open-domain information-seeking conversations. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR'19, pages 475–484, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450361729.

[2] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA, 2020. Curran Associates Inc. ISBN 9781713829546.

[3] Caterina Bérubé, Marcia Nißen, Rasita Vinay, Alexa Geiger, Tobias Budig, Aashish Bhandari, Catherine Rachel Pe Benito, Nathan Ibarcena, Olivia Pistolese, Pan Li, Abdullah Bin Sawad, Elgar Fleisch, Christoph Stettler, Bronwyn Hemsley, Shlomo Berkovsky, Tobias Kowatsch, and A. Baki Kocaballi. Proactive behavior in voice assistants: A systematic review and conceptual model. *Computers in Human Behavior Reports*, 14:100411, 2024. ISSN 2451-9588.

[4] J Cassell, T Bickmore, L Campbell, H Vilhjálmsson, and H Yan. More than just a pretty face: conversational protocols and the affordances of embodiment. *Knowledge-Based Systems*, 14(1): 55–64, 2001. ISSN 0950-7051.

[5] Edward Y. J. Chung and Jea Sophia Oh. *Introduction: Emotions (Jeong/Qing 情) in Korean Philosophy and Religion*, pages 1–93. Springer International Publishing, 2022.

[6] Yang Deng, Wenqiang Lei, Wai Lam, and Tat-Seng Chua. A survey on proactive dialogue systems: problems, methods, and prospects. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, IJCAI '23, 2023. ISBN 978-1-956792-03-4.

[7] Yang Deng, Lizi Liao, Zhonghua Zheng, Grace Hui Yang, and Tat-Seng Chua. Towards human-centered proactive conversational agents. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '24, pages 807–818, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400704314.

[8] Yang Deng, Lizi Liao, Wenqiang Lei, Grace Hui Yang, Wai Lam, and Tat-Seng Chua. Proactive conversational ai: A comprehensive survey of advancements and opportunities. *ACM Trans. Inf. Syst.*, 43(3), March 2025. ISSN 1046-8188.

[9] Hao Fei, Han Zhang, Bin Wang, Lizi Liao, Qian Liu, and Erik Cambria. EmpathyEar: An open-source avatar multimodal empathetic chatbot. In Yixin Cao, Yang Feng, and Deyi Xiong, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 61–71, Bangkok, Thailand, August 2024. Association for Computational Linguistics.

[10] Pascale Fung, Yoram Bachrach, Asli Celikyilmaz, Kamalika Chaudhuri, Delong Chen, Willy Chung, Emmanuel Dupoux, Hongyu Gong, Hervé Jégou, Alessandro Lazaric, Arjun Majumdar, Andrea Madotto, Franziska Meier, Florian Metze, Louis-Philippe Morency, Théo Moutakanni, Juan Pino, Basile Terver, Joseph Tighe, Paden Tomasello, and Jitendra Malik. Embodied ai agents: Modeling the world, 2025.

[11] Eric Horvitz. Principles of mixed-initiative user interfaces. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '99, pages 159–166, New York, NY, USA, 1999. Association for Computing Machinery. ISBN 0201485591.

[12] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA, 2020. Curran Associates Inc. ISBN 9781713829546.

[13] Xingyu Bruce Liu, Shitao Fang, Weiyan Shi, Chien-Sheng Wu, Takeo Igarashi, and Xiang 'Anthony' Chen. Proactive conversational agents with inner thoughts. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, CHI '25, New York, NY, USA, 2025. Association for Computing Machinery. ISBN 9798400713941.

[14] Zhicong Lu, Chenxinran Shen, Jiannan Li, Hong Shen, and Daniel Wigdor. More kawaii than a real-person live streamer: Understanding how the otaku community engages with and perceives virtual youtubers. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI '21, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450380966.

[15] Yukun Ma, Khanh Linh Nguyen, Frank Z. Xing, and Erik Cambria. A survey on empathetic dialogue systems. *Information Fusion*, 64:50–70, 2020. ISSN 1566-2535.

[16] Payod Panda, Molly Jane Nicholas, Mar Gonzalez-Franco, Kori Inkpen, Eyal Ofek, Ross Cutler, Ken Hinckley, and Jaron Lanier. Alltogether: Effect of avatars in mixed-modality conferencing environments. In *Proceedings of the 1st Annual Meeting of the Symposium on Human-Computer Interaction for Work*, CHIWORK '22, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450396554.

[17] Aravind Sesagiri Raamkumar and Yinping Yang. Empathetic conversational systems: A review of current advances, gaps, and opportunities. *IEEE Trans. Affect. Comput.*, 14(4):2722–2739, October 2023. ISSN 1949-3045.

[18] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org, 2023.

[19] Chris Samarinas and Hamed Zamani. Procis: A benchmark for proactive retrieval in conversations. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR 2024, pages 830–840. ACM, July 2024.

[20] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, pages 6000–6010, Red Hook, NY, USA, 2017. Curran Associates Inc. ISBN 9781510860964.

[21] Fu-Chia Yang, Pedro Acevedo, Siqi Guo, Minsoo Choi, and Christos Mousas. Embodied conversational agents in extended reality: A systematic review. *IEEE Access*, 13:79805–79824, 2025.

[22] Han Zhang, Zixiang Meng, Meng Luo, Hong Han, Lizi Liao, Erik Cambria, and Hao Fei. Towards multimodal empathetic response generation: A rich text-speech-vision avatar-based benchmark. In *Proceedings of the ACM on Web Conference 2025*, WWW '25, page 2872–2881, New York, NY, USA, 2025. Association for Computing Machinery. ISBN 9798400712746.