Towards Human-centered Proactive Conversational AI

Yang Deng

Singapore Management University
ydeng@smu.edu.sg

Conversational AI agents are envisioned to provide social support or functional service to human users via natural language interactions. However, typical conversational systems are built to follow instructions, which means that the conversation is led by the user, and the system simply follows the user's instructions or intents. My research endows the conversational AI with the capabilities of creating or controlling the conversation to achieve the conversational goals by taking initiative and anticipating impacts on themselves or human users (**INTELLIGENCE**), namely *Proactive Conversational AI*. I will also highlight the importance of moving towards building human-centered proactive conversational AI (Deng et al. 2024a) that emphasize human needs and expectations (**ADAPTIVITY**), and that considers ethical and social implications of these agents (**CIVILITY**), rather than solely focusing on technological capabilities.

PART I: INTELLIGENCE Proactive features in dialogue systems (Deng et al. 2023a,b, 2025) can significantly enhance user engagement and service efficiency. I will introduce a comprehensive evaluation on the proactivity of LLMs across multiple proactive dialogue tasks (Deng et al. 2023c) and a new dialogue policy planning paradigm (Deng et al. 2024b) to strategize large language models (LLMs) for proactive dialogue problems with a tunable language model plug-in as a plug-and-play dialogue policy planner, which can be supervisedly fine-tuned over available human-annotated data as well as conduct reinforcement learning from goal-oriented AI feedback with dynamic interaction data. This framework is further applied into various applications, such as target-guided conversational recommendation (Dao et al. 2024), asking clarification questions in conversational information seeking (Chen et al. 2024), and noncollaborative dialogues (Zhang et al. 2024a).

PART II: ADAPTIVITY A human-centered proactive conversational AI must recognize and adapt to diverse user needs, preferences, and values. To enable strategic planning tailored to heterogeneous users, we developed a population-based training framework (Zhang et al. 2024a) that employs diversified user simulators with varied personas and decision-making behaviors. This approach allows agents to learn robust strategies that generalize across a broad spectrum of human profiles. Under this new training paradigm, there are two questions that are worth further studying: 1) How reliable is the persona-driven user simulation? (Wu et al. 2025), and 2) How well can the LLM understand user mental state? (Li, Shi, and Deng 2026).

PART III: CIVILITY As LLMs serve as foundation of the conversational AI, the trust and reliability of LLMs becomes utmost important. We need to identify and understand the potential causes and mechanisms of unintended behaviors in the LLM-powered conversational agents and develop techniques to reduce the likelihood of such behaviors occurring and the potential harm that may be caused by them. LLMs often struggle with different trust and reliability issues, including generating factually incorrect content (Qin et al. 2024) and producing toxic or disruptive content (Zhao et al. 2025a,b). Specifically, I investigate the knowledge boundary of LLMs (Li et al. 2025; Zhou et al. 2026), such as mitigating unknown questions (Deng et al. 2024c) and improving faithful integrity (Zhao et al. 2025c).

I will conclude by outlining future directions for proactive interactions beyond dialogue (Zhang et al. 2024b), including multi-agent environments and AI mentoring systems, and will discuss the emerging risks and opportunities that accompany these next-generation proactive AI agents.

References

Chen, Y.; Huang, C.; Deng, Y.; Lei, W.; Jin, D.; Liu, J.; and Chua, T. 2024. STYLE: Improving Domain Transferability of Asking Clarification Questions in Large Language Model Powered Conversational Agents. In *Findings of ACL, ACL 2024*.

Dao, H.; Deng, Y.; Bui, K.; Le, D. D.; and Liao, L. 2024. Experience as Source for Anticipation and Planning: Experiential Policy Learning for Target-driven Recommendation Dialogues. In *Findings of ACL: EMNLP 2024*.

Deng, Y.; Lei, W.; Huang, M.; and Chua, T. 2023a. Goal Awareness for Conversational AI: Proactivity, Non-collaborativity, and Beyond. In *ACL 2023*.

Deng, Y.; Lei, W.; Lam, W.; and Chua, T. 2023b. A Survey on Proactive Dialogue Systems: Problems, Methods, and Prospects. In *IJCAI 2023*.

Deng, Y.; Liao, L.; Chen, L.; Wang, H.; Lei, W.; and Chua, T. 2023c. Prompting and Evaluating Large Language Models for Proactive Dialogues: Clarification, Target-guided, and Non-collaboration. In *Findings of the Association for Computational Linguistics: EMNLP 2023*.

Deng, Y.; Liao, L.; Lei, W.; Yang, G. H.; Lam, W.; and Chua, T. 2025. Proactive Conversational AI: A Comprehensive Survey of Advancements and Opportunities. *ACM Trans. Inf. Syst.*, 43(3): 67:1–67:45.

Deng, Y.; Liao, L.; Zheng, Z.; Yang, G. H.; and Chua, T. 2024a. Towards Human-centered Proactive Conversational Agents. In *SIGIR 2024*.

Deng, Y.; Zhang, W.; Lam, W.; Ng, S.; and Chua, T. 2024b. Plug-and-Play Policy Planner for Large Language Model Powered Dialogue Agents. In *ICLR 2024*.

Deng, Y.; Zhao, Y.; Li, M.; Ng, S.; and Chua, T. 2024c. Don't Just Say "I don't know"! Self-aligning Large Language Models for Responding to Unknown Questions with Explanations. In *EMNLP 2024*.

Li, M.; Shi, X.; and Deng, Y. 2026. RecToM: A Benchmark for Evaluating Machine Theory of Mind in LLM-based Conversational Recommender Systems. In *AAAI 2026*.

Li, M.; Zhao, Y.; Zhang, W.; Li, S.; Xie, W.; Ng, S.; Chua, T.; and Deng, Y. 2025. Knowledge Boundary of Large Language Models: A Survey. In *ACL 2025*.

Qin, P.; Huang, C.; Deng, Y.; Lei, W.; and Chua, T. 2024. Beyond Persuasion: Towards Conversational Recommender System with Credible Explanations. In *Findings of ACL: EMNLP 2024*.

Wu, S.; Zhu, Y.; Hsu, W.; Lee, M. L.; and Deng, Y. 2025. From Personas to Talks: Revisiting the Impact of Personas on LLM-Synthesized Emotional Support Conversations. In *EMNLP 2025*.

Zhang, T.; Huang, C.; Deng, Y.; Liang, H.; Liu, J.; Wen, Z.; Lei, W.; and Chua, T. 2024a. Strength Lies in Differences! Improving Strategy Planning for Non-collaborative Dialogues via Diversified User Simulation. In *EMNLP*.

Zhang, X.; Deng, Y.; Ren, Z.; Ng, S.; and Chua, T. 2024b. Ask-before-Plan: Proactive Language Agents for Real-World Planning. In *Findings of ACL: EMNLP 2024*.

Zhao, W.; Hu, Y.; Deng, Y.; Guo, J.; Sui, X.; Han, X.; Zhang, A.; Zhao, Y.; Qin, B.; Chua, T.; and Liu, T. 2025a. Beware of Your Po! Measuring and Mitigating AI Safety Risks in Role-Play Fine-Tuning of LLMs. In *ACL 2025*.

Zhao, W.; Hu, Y.; Deng, Y.; Wu, T.; Zhang, W.; Guo, J.; Zhang, A.; Zhao, Y.; Qin, B.; Chua, T.; and Liu, T. 2025b. MPO: Multilingual Safety Alignment via Reward Gap Optimization. In *ACL 2025*.

Zhao, Y.; Deng, Y.; Ng, S.; and Chua, T. 2025c. Aligning Large Language Models for Faithful Integrity Against Opposing Argument. In *AAAI 2025*.

Zhou, Y.; Huang, H.; Liu, Y.; Dai, R.; Wang, X.; Zhang, X.; Shi, S.; and Deng, Y. 2026. Do Retrieval Augmented Language Models Know When They Don't Know? In *AAAI 2026*.