Real-Time AI Mediation for Human-Human Digital Communications

Elisa Salatti*, Susanna Pelagatti*, Antonio Caruso†, Stefano Chessa*
*Dipartimento di Informatica, Università di Pisa, Pisa, Italy
†Dipartimento di Matematica e Fisica, Università del Salento, Salento, Italy

***Abstract*—Artificial Intelligence Mediated Communication (AIMC) is emerging as a promising paradigm for supporting human communication through AI systems. In this work, we present an architecture in which an agentic Large Language Model (LLM) operates as a mediation layer between users in a chat environment. Rather than acting as an additional participant, the AI analyzes messages before delivery to avoid harmful interactions and maintain discussions aligned with a shared objective. We describe the proposed architecture and outline a preliminary evaluation study to investigate users acceptance and perceived usability of AI mediation.**

***Index Terms*—Large Language Model, Agents, AI-mediated communication**

I. Introduction

Digital communication platforms play an increasingly important role in collaborative activities such as online learning, where participants must coordinate, exchange ideas, and work toward shared objectives. In these environments, the quality of interactions can significantly influence collaboration outcomes and learning effectiveness [1], [2].

Large Language Models (LLMs) are increasingly integrated into digital communication environments. Their growing adoption is influencing how individuals create, revise, and exchange messages, contributing to the emergence of AI-Mediated Communication (AIMC) [3].

One of the most widespread configurations involves the use of LLM-based tools, such as ChatGPT or Copilot, to support the creation and revision of content before it is shared. These systems are commonly used to draft or refine emails, documents, social media posts, and chat messages. In this configuration, AI plays an assistive role, while the human user remains responsible for reviewing, approving, and ultimately sending the content.

More recently, LLM-based conversational agents have been employed as interlocutors within digital conversations and, in some cases, as moderators aimed at regulating interactions, structuring discussions, and promoting participant engagement [4], [5]. To ensure safe and appropriate interactions, these systems typically incorporate safety and guardrail mechanisms designed to constrain inappropriate behaviors and reduce the risk of harmful outputs [6].

To the best of our knowledge, research on AI acting as an intermediary between human communicators remains limited. A recent example is the work by Aoyama et al. [7] which introduced an LLM-based adaptive agent pair for mediating human-to-human digital communication.

Building on this emerging research direction, we investigate a configuration in which AI operates as a real-time mediation layer between human interlocutors. The proposed architecture mediates messages before delivery with the objective of supporting constructive and goal-oriented communication. This architecture may be particularly relevant in collaborative settings, such as online learning activities, where individuals are required to work toward a shared objective through digital communication.

II. System design

In the proposed architecture, messages are not exchanged directly between participants. As shown in Fig. 1, each user is associated with a dedicated LLM-based agent.

Whenever a user sends a message, the corresponding agent intercepts and analyzes it before it is delivered in the shared chat environment. The agent may then decide to forward the original message or adapt its content according to the communication objectives defined for the application context. Incoming messages are delivered to users without further modification. The agent does not participate in the conversation as an additional interlocutor and does not introduce new information. It can filter inappropriate content, reduce offensive or aggressive language, and prevent the conversation from deviating from the intended topic or task.

All participants are aware that the conversation is mediated by AI. However, users only interact with the mediated version of messages and do not have access to the original content produced by the other participant.

III. Agentic system

Considering the scenario described above, the challenges at the system architecture design level are manifold. These challenges arise both in the interaction between a single LLM and a user, and in the interaction between the LLM and the broader conversational context (or more generally, the chosen interaction mechanism). In the following points we discuss some architectural issues ranging from *system prompt engineering* to the design of fully agentic environments, in which LLM-based agents operate within a surrounding orchestration and control framework.

- **Prompt-Guided Interaction.** The *system prompt*, hidden from the user, defines the behavior of the LLM as a

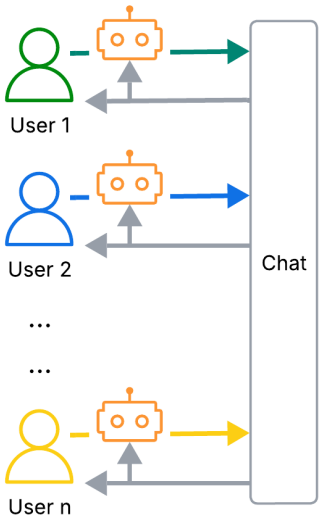

Fig. 1. Overview of the proposed AI-mediated communication architecture. Each participant is associated with a dedicated LLM agent that intercepts and processes outgoing messages before they are delivered in the shared chat environment.

conversational proxy. It specifies the interaction policy, ranging from simple message forwarding to more sophisticated mediation and adaptation strategies.

- **User-Specific Prompt Customization.** System prompts can be dynamically tailored to individual user requirements, such as preferred language, communication style, accessibility needs, or context-dependent message reformulation.
- **Adaptive Agentic Prompting.** The LLM can be embedded within a *harnessing* environment that manages a repository of prompts and interaction policies. In this setting, prompts are not static but evolve according to interaction history and feedback signals, enabling reinforcement-learning-based optimization of system behavior.

This depicts a large space of possibilities in system design, that must be guided by a proper definition of suitable quality metrics; the latter, are discussed in the next section.

IV. Human Factors and Evaluation

While AI-mediated mechanisms for human-human communication may contribute to goal-oriented conversations, their impact on user experience cannot be assumed. The presence of an AI mediator could influence how users perceive communication, autonomy, transparency, and trust. Awareness of AI intervention may affect how individuals perceive their interlocutors, as it may become more difficult to determine to what extent a message reflects the sender's own views, intentions, and communicative abilities rather than the contribution of the AI mediator [8].

Despite these concerns, AI-mediated communication may offer benefits in contexts characterized by well-defined collaborative objectives. In educational group activities, for example, it could help maintain discussions aligned with the assigned task and facilitate more goal-oriented collaboration.

Consequently, a central aspect of this research concerns the evaluation of users' perceptions, acceptance, and trust toward AI-mediated communication. To this end, we plan to evaluate the proposed system through a controlled study comparing a standard chat with an AI-mediated chat. Participants will complete a collaborative decision-making task, allowing us to analyze both communication patterns and task-related outcomes. In addition, questionnaires will be used to investigate users' perceptions of the system.

V. Conclusions

This paper presented an AI-mediated communication architecture in which an LLM acts as an intermediary between human users during digital interactions. In the proposed approach, the LLM can adapt messages before delivery to avoid harmful exchanges while helping conversations remain focused on a shared task or objective. Current and future work focuses on implementing and evaluating the system in collaborative scenarios, investigating both communication outcomes and users' perceptions of AI-mediated interactions.

References

[1] A. Bach and F. Thiel, "Collaborative online learning in higher education—quality of digital interaction and associations with individual and group-related factors," *Frontiers in Education*, vol. Volume 9 - 2024, 2024.

[2] A. A. M. Kasim, N. A. Darus, N. A. A. Lee, A. L. B. Subramaniam, and J. Januin, "Online collaborative performance in group-based tasks among learners of higher education," *Studies in English Language and Education*, vol. 9, no. 3, pp. 948–966, 2022. [Online]. Available: https://doi.org/10.24815/siele.v9i3.24861

[3] J. T. Hancock, M. Naaman, and K. Levy, "AI-Mediated Communication: Definition, Research Agenda, and Ethical Considerations," *Journal of Computer-Mediated Communication*, vol. 25, no. 1, pp. 89–100, Mar. 2020.

[4] Y. Mao, Y. Li, H. Liang, Z. Shen, X. Xu, and D. Wang, "Multi-user chat assistant (muca): A framework using llms to facilitate group conversations," 2024. [Online]. Available: https://arxiv.org/abs/2401.04883

[5] S. Yoo, "Read the room: Exploring when llm agents should participate in everyday mixed-context group chats on mobile messaging platforms," in *Proceedings of the Extended Abstracts of the 2026 CHI Conference on Human Factors in Computing Systems*, ser. CHI EA '26. New York, NY, USA: Association for Computing Machinery, 2026. [Online]. Available: https://doi.org/10.1145/3772363.3798392

[6] H. Abdelkader, M. Abdelrazek, S. Singh, I. Logothetis, P. Rani, R. Vasa, and J.-G. Schneider, "Safeguarding LLM-Applications: Specify or Train?" in *2025 IEEE/ACM 4th International Conference on AI Engineering – Software Engineering for AI (CAIN)*. Ottawa, ON, Canada: IEEE, Apr. 2025, pp. 275–276.

[7] S. Aoyama, R. Chujo, A. Hautasaari, and T. Naemura, "Intersubjective Model of AI-mediated Communication: Augmenting Human-Human Text Chat through LLM-based Adaptive Agent Pair," Feb. 2025.

[8] S. Sahebi and P. Formosa, "The AI-mediated communication dilemma: Epistemic trust, social media, and the challenge of generative artificial intelligence," *Synthese*, vol. 205, no. 3, p. 128, Mar. 2025.