Text Overlap: An LLM with Human-like Conversational Behaviors

JiWoo Kim[1], **Minsuk Chang**[2*], **JinYeong Bak**[1*]
[1]Sungkyunkwan University, Suwon, South Korea
[2]Google Deepmind, Seattle, USA
wldn9705@skku.edu, minsukchang@google.com, jy.bak@skku.edu

Abstract

Traditional text-based human-AI interactions typically follow a strict turn-taking approach. This rigid structure limits conversational flow, unlike natural human conversations, which can freely incorporate overlapping speech. However, our pilot study suggests that even in text-based interfaces, overlapping behaviors such as backchanneling and proactive responses lead to more natural and functional exchanges. Motivated by these findings, we introduce text-based overlapping interactions as a new challenge in human-AI communication, characterized by real-time typing, diverse response types, and interruptions. To enable AI systems to handle such interactions, we define three core tasks: deciding when to overlap, selecting the response type, and generating utterances. We construct a synthetic dataset for these tasks and train OverlapBot, an LLM-driven chatbot designed to engage in text-based overlapping interactions. Quantitative and qualitative evaluations show that OverlapBot increases turn exchanges compared to traditional turn-taking systems, with users making 72% more turns and the chatbot 130% more turns, which is perceived as efficient by end-users. This finding supports overlapping interactions and enhances communicative efficiency and engagement.

1 Introduction

Human-to-human conversations differ from chess, where turns are strictly alternated. In human-to-human conversation, overlaps and interruptions are common, requiring participants to coordinate who speaks, when to stop, and when to continue (Duncan, 1972; Sacks et al., 1974). On the other hand, current text-based chat interactions follow strict turn-taking, similar to playing chess. This applies not only to human-human interactions but also to interactions with Large Language Models (LLM), where users must wait for the chatbot to respond before the conversation can be continued (Zhou et al., 2023).

User: I watched a movie recently –

OverlapBot: Uh-huh.

User: – and loved how the director handled the big twist. But I can't remember who –

OverlapBot: I think you are mentioning Bong Joon-ho.

User: Can you suggest more of his –

OverlapBot: Here are some of his most notable films that you should check –

User: Oh, only horror movies.

OverlapBot: If you're looking for horror movies –

Table 1: Examples of the types of overlap made by OverlapBot. While the user is typing (–), OverlapBot can provide listener cues indicating attention (*Uh-huh*) or generate a response even if the user's typing is not finished.

Refining strict turn-taking remains relatively underexplored in NLP, despite efforts in speech and robotics to improve turn-taking dynamics (Aylett and Romeo, 2023; Aylett et al., 2023; Ehret et al., 2023; Janowski and André, 2018; Skantze, 2021a). Speech-based systems have primarily focused on reducing awkward silences (Phukon et al., 2022; Ma et al., 2024; Chang et al., 2022), while robotic systems have shown that improved turn-taking enhances conversational naturalness (Paetzel-Prüsmann and Kennedy, 2023; Lala et al., 2019; Moujahid et al., 2022). Although a recent text-based study (Zhang et al., 2024) introduced duplex response generation, existing studies have yet to refine turn-taking based on specific conversational behaviors observed in human dialogue.

Strict turn-taking may overlook important conversational features, both in terms of naturalness and functionality. To investigate this, we conducted a pilot study where seven pairs of participants engaged in text-based conversations using a real-time chat interface that allowed simultaneous typing and message visibility. Our observations revealed

[*]Corresponding authors

Proceedings of the Third Workshop on Social Influence in Conversations (SICon 2025), pages 127–139
July 31, 2025 ©2025 Association for Computational Linguistics

that overlapping interactions enabled functional conversational behaviors, particularly backchanneling, where participants provided brief acknowledgments (e.g., *yeah* or *got it*) while reading the other person's message, and preemptive answering, where they anticipated and responded before a question or statement was fully articulated. These findings suggest that allowing overlap in text-based interactions fosters more natural and functional exchanges.

Motivated by this finding, we introduce text overlapping interactions as a new challenge in human-AI communication, where interactions involve real-time typing, diverse response types, and interruptions. Unlike strict turn-taking systems, AI capable of handling overlap must dynamically manage when to interject, provide backchannels, or generate preemptive responses. To address this, we define three core evaluation tasks: (1) Timing classification, deciding whether to overlap or wait; (2) Action classification, determining the appropriate response type, such as backchanneling or producing a full response; and (3) Utterance generation, producing natural overlapping responses that maintain conversational flow. By tackling these challenges, we aim to develop AI systems that better align with human conversational behaviors.

Thus, we develop OverlapBot, an LLM-driven chatbot, using a synthetic dataset constructed from a conversation dataset (Godfrey et al., 1992) and an instruction-tuning dataset (Taori et al., 2023). OverlapBot is finetuned on Llama3-8B with parameter-efficient tuning, optimizing it for the Timing classification (when to overlap), Action classification (response type selection), and Utterance generation (producing overlapping responses). We develop a dedicated chat interface that supports overlapping functionalities such as real-time typing and interruptions.

Our evaluation shows that OverlapBot improves both system performance and the end-user experience in overlapping interactions. It demonstrated better performance than the baselines in timing accuracy, act classification, and utterance generation, while a user study with 18 participants found it more communicative and immersive than a conventional turn-taking chatbot. OverlapBot generates more concise messages and increases and enables faster turn-taking, highlighting the benefits of overlapping interactions for efficiency and engagement.

In summary, our contributions include:

- We define text overlapping interactions in human-LLM conversations based on observed human behaviors in our pilot study.
- We establish key evaluation tasks for assessing timing, response type selection, and conversational coherence.
- We develop OverlapBot, an LLM-driven chatbot that manages overlaps through backchanneling and preemptive responses.
- We show that overlapping in human-AI interactions facilitates faster turn-taking and make conversations feel more natural and engaging.

2 Text Overlap Interactions

We characterize text-based overlapping interactions based on key findings from our pilot study (Appendix B). In this study, 14 participants engaged in 10-minute real-time chat conversations on decision-making tasks. The interface allowed them to see their partner's typing as it happened, creating a conversational flow similar to spoken dialogue. From our observations, we identify three key elements that characterize text-overlapping interactions:

Real-time Typing The interface displays participants' typing activities in real-time, allowing both parties to see input as it is being written. This shared visibility creates opportunities for overlap by enabling responses before message completion. For instance, if a user types *I want to be*, and their conversation partner simultaneously responds with *Yeah*, the overlap occurs at the word *be* in the user's utterance.

Types of Response Text-based overlap manifests in two primary forms: backchanneling and preemptive answering (Table 1). Backchanneling involves brief, real-time acknowledgments (e.g., *uh-huh*, *I see*) that signal active listening without disrupting the conversation. In preemptive answering, a speaker anticipates and responds to an incomplete utterance before the other party finishes typing.

Interruptions and Deletions Speakers often adjust their responses when an overlap occurs, either deleting unfinished text or rephrasing to maintain conversational coherence. For example, if a user begins typing *I was thinking we could try –* , but the other person interrupts with *Let's go to the Italian place!*, the user deletes their unfinished sentence

Dialogues	Timing	Action	Utterance
User: Have **Ground Truth:** [Await] **Hypothesis:** [Await] **Hypothesis:** [Overlap] [Answer] *I tried, but I couldn't.*	✓ ✗	– ✗	– ✗
User: Have you painted **Ground Truth:** [Overlap] [Understanding] *Mm-hmm.* **Hypothesis:** [Overlap] [Answer] *I painted yesterday.*	✓	✗	✗
User: Have you painted anything recently? **Ground Truth:** [Overlap] [Answer] *Yes, I painted a small landscape last weekend.* **Hypothesis:** [Overlap] [Answer] *Mm-hmm.*	✓	✓	✗

Table 2: Examples of Timing, Action, and Utterance tasks with correct (✓) and incorrect (✗) predictions. '–' indicates exclusion from score calculation.

and instead replies with *Yeah, that works!*, adjusting their response to fit the new conversational direction.

3 Approach

3.1 Training Strategy

To enable LLMs to handle text overlap interactions, we establish three core evaluation tasks as shown in Table 3 and Table 2.

We created a synthetic dataset by modifying existing datasets to align with the three core tasks. The final dataset consists of 15,377 training samples, 6,482 validation samples, and 6,978 test samples. An example of the modified format is shown below.

Instruct Evaluate whether the interlocutor would overlap this utterance or wait his turn to come. If your evaluation is to overlap, return your evaluation as [Overlap] _dialogue_act_ _answer_. You have to choose a _dialogue_act_: either [Understanding] or [Answer]. You have to fill _answer_ with your own answer to this utterance. Otherwise, if your evaluation is to wait, return your evaluation as only [Await].

User Have you painted

Assistant [Overlap] [Understanding] *Mm-hmm.*

We created the synthetic dataset from two existing datasets: a conversation dataset and an instruction-tuning dataset. The first dataset, the Switchboard Dialogue Act Corpus (SWDA), consists of 1,155 five-minute telephone conversations between 440 participants discussing various topics such as child care, recycling, and news media (Godfrey et al., 1992). We selected SWDA for its detailed dialogue annotations, which include overlapping behaviors such as backchanneling and sentence completion. The second dataset was an instruction-tuning dataset (Taori et al., 2023). Since SWDA is primarily a conversational dataset, we recognized that a model trained solely on SWDA might struggle with task-oriented dialogues. For the instruction-tuning dataset, we randomly segmented and reformulated responses to synthesize assistant replies that align with overlapping interactions.

Task	Description
Timing	Decide whether to overlap or wait.
Example	User typing *"Have you painted,"* then model predicts [Overlap] or [Await].
Action	Choose the type of response when overlapping.
Example	If [Understanding], model selects backchanneling. If [Answer], model selects full answer.
Utt.	Generate a natural response based on the Action selection.
Example	If [Understanding], model generates *Um-hmm.* If [Answer], model generates *I painted something (...)*

Table 3: Evaluation tasks for overlapping interactions. Details are on Appendix C.

We finetuned the Llama3 8B instruct model using parameter-efficient techniques (AI@Meta, 2024; Hu et al., 2022). Training details are provided in Appendix D. We evaluated the chatbots' automatic performance on classification accuracy (F1 score) and reference-based generation accuracy (Bleu (Papineni et al., 2002), Rouge-L (Lin, 2004)). For our baseline models, we used Llama3

Model	Timing	Action	Utterance
Llama3 8B	0.46 (±0.01)	0.37 (±0.03)	0.16 (±0.08) / 0.11 (±0.01)
GPT4o	0.47 (±0.00)	0.73 (±0.02)	0.18 (±0.06) / 0.16 (±0.02)
GPT4 turbo	0.46 (±0.02)	0.73 (±0.07)	0.22 (±0.04) / 0.15 (±0.02)
OverlapBot	**0.65** (±0.04)	**0.80** (±0.07)	**0.55** (±0.02) / **0.30** (±0.02)

Table 4: Automatic evaluation results. Timing and Action values represent F1 scores. Utterance values represent BLEU and Rouge-L F1 scores, respectively. Standard deviations obtained 3-fold cross-validation are shown in parentheses.

Metric	Role	Conventional	OverlapBot
Message Length	User	62.36 (±22.49)	43.18 (±12.74)
	Chatbot	177.64 (±34.65)	133.40 (±42.19)
Total Turns	User	7.56 (±2.59)	13.00 (±3.93)
	Chatbot	7.33 (±2.40)	16.89 (±7.19)
# Turns / Minute	User	1.28 (±0.45)	1.93 (±0.82)
	Chatbot	1.25 (±0.45)	2.48 (±1.33)
Overlap Ratio		-	6.0% (±3.0%)
# Deletes / Minute	User	-	11.10 (±6.62)
	Chatbot	-	2.98 (±1.70)

Table 5: Quantitative comparison of conventional chatbot and OverlapBot in our study. Overlap Ratio represents the percentage of total conversation time where simultaneous keystrokes occurred between the User and OverlapBot.

8B instructed tuned model, GPT-4o (gpt-4o-2024-08-06), GPT-4 Turbo (gpt-4-turbo-2024-04-09) through the OpenAI API.

3.2 Evaluation Results

Automatic evaluation results indicated that OverlapBot exhibited better performance across all assessed dimensions, including timing, action execution, and utterance generation (Table 4).

In addition, we conducted a user study where 18 participants engaged in free topic conversations with both the conventional turn-taking chatbot and OverlapBot. For comparison with the conventional turn-taking system, we implemented a chat system where neither users nor the chatbot could see each other's typing. In this system, we employed the vanilla Llama3-8B instruct-tuned model. We analyzed participants' conversation logs and interview transcripts. The overall procedure of our study was conducted after obtaining IRB approval from the university. Details on user study are in Appendix E.

Table 5 presents the quantitative results of the user study, showing that OverlapBot facilitated shorter message lengths and a higher number of turns exchanged compared to the conventional chatbot. Here, turns are calculated based on Send actions, not typing status. Notably, the OverlapBot sent messages more frequently than the conventional chatbot, indicating its ability to provide more information within the same timeframe. Interestingly, the ratio of turns exchanged between the user and the chatbot, which was nearly a balanced exchange of turns in the conventional interface, shifted in the OverlapBot interaction. This shift could be attributed to OverlapBot's backchanneling behavior, which might not have elicited responses from users. Additionally, users deleted messages more frequently than OverlapBot, possibly due to revising their written content before resending it to the LLM, or intentionally removing their input to avoid leaving their words in the conversation logs.

Additionally, an analysis of interview transcripts revealed three general impressions of OverlapBot compared to the conventional chatbot. First, interactions felt similar to conversing with a real person. Participants specifically noted that OverlapBot felt more communicative and immersive compared to the conventional chatbot. Second, OverlapBot enabled more efficient interactions. Since it could provide preemptive responses while users were typing and users could interrupt it, conversations became more fast-paced and efficient. Third, while the increased speed was generally perceived positively, some participants noted that OverlapBot's responses tended to be shorter and less structured.

4 Conclusion and Future Work

We introduce text-based overlapping features into human-AI interactions. We show the key characteristics of text overlapping and develop specific tasks for LLMs to handle such interactions. Our implementation with a finetuned LLM shows improvements in interaction efficiency and naturalness compared to traditional turn-taking systems.

Our results highlight key directions for extending this work. While our implementation shows the potential of text-based overlapping, further research is needed to assess its effectiveness across different interaction scenarios. Additionally, developing metrics to balance interaction speed and response quality is meaningful for real-world applications. Furthermore, extending this work to multimodal interactions that integrate text and speech can be a meaningful direction (Cho et al., 2022).

Understanding how LLMs process these overlaps could lead to more responsive AI systems across modalities.

Limitations

We implemented deletions systematically rather than relying on the LLM to delete messages on its own, as language models inherently predict the next token rather than modify past outputs. Due to this limitation, deletion was not included as one of the evaluation tasks.

Further, the more natural interaction with OverlapBot does not mitigate common limitations of LLMs, such as hallucinations, limited knowledge, and lack of long-term memory (Laskar et al., 2024).

Ethical Considerations

We used publicly available data to create a synthetic dataset for training our model. During the user study, we provided participants with appropriate guidelines, ensuring that they were aware of their tasks and how their data will be utilized. After the study, all personal information was deleted.

Acknowledgments

We would like to thank the anonymous reviewers for their helpful questions and comments. We further would like to express our gratitude to Jonas Belouadi for discussions, proofreading, and comments on our work. This work was partly supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (RS-2019-II190421, AI Graduate School Support Program (Sungkyunkwan University) and RS-2025-02263169, Detection and Prediction of Emerging and Undiscovered Voice Phishing) and National Research Foundation of Korea (NRF-RS-2025-00523385).

References

AI@Meta. 2024. Llama 3 model card.

Matthew Peter Aylett and Marta Romeo. 2023. You don't need to speak, you need to listen: Robot interaction and human-like turn-taking. In *Proceedings of the 5th International Conference on Conversational User Interfaces, CUI 2023, Eindhoven, The Netherlands, July 19-21, 2023*, pages 11:1–11:5. ACM.

Matthew Peter Aylett, Éva Székely, Donald McMillan, Gabriel Skantze, Marta Romeo, Joel E. Fischer, and Gisela Reyes-Cruz. 2023. Why is my agent so slow? deploying human-like conversational turn-taking. In *International Conference on Human-Agent Interaction, HAI 2023, Gothenburg, Sweden, December 4-7, 2023*, pages 490–492. ACM.

Adrian Bennett. 1978. Interruptions and the interpretation of conversation. In *Annual Meeting of the Berkeley Linguistics Society*, pages 557–575.

Richard E Boyatzis. 1998. *Transforming qualitative information: Thematic analysis and code development*. Sage.

Shuo-Yiin Chang, Bo Li, Tara N. Sainath, Chao Zhang, Trevor Strohman, Qiao Liang, and Yanzhang He. 2022. Turn-taking prediction for natural conversational speech. In *23rd Annual Conference of the International Speech Communication Association, Interspeech 2022, Incheon, Korea, September 18-22, 2022*, pages 1821–1825. ISCA.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, Wei Ye, Yue Zhang, Yi Chang, Philip S. Yu, Qiang Yang, and Xing Xie. 2024. A survey on evaluation of large language models. *ACM Trans. Intell. Syst. Technol.*, 15(3):39:1–39:45.

Eugene Cho, Nasim Motalebi, S. Shyam Sundar, and Saeed Abdullah. 2022. Alexa as an active listener: How backchanneling can elicit self-disclosure and promote user experience. *Proc. ACM Hum.-Comput. Interact.*, 6(CSCW2).

Patricia M Clancy, Sandra A Thompson, Ryoko Suzuki, and Hongyin Tao. 1996. The conversational use of reactive tokens in english, japanese, and mandarin. *Journal of pragmatics*, 26(3):355–387.

Elizabeth Clark, Anne Spencer Ross, Chenhao Tan, Yangfeng Ji, and Noah A. Smith. 2018. Creative writing with a machine in the loop: Case studies on slogans and stories. In *Proceedings of the 23rd International Conference on Intelligent User Interfaces, IUI 2018, Tokyo, Japan, March 07-11, 2018*, pages 329–340. ACM.

Jennifer Coates. 1994. No gap, lots of overlap: Turn-taking patterns in the talk of women friends. *Researching language and literacy in social context*, pages 177–192.

Hai Dang, Lukas Mecke, and Daniel Buschek. 2022a. Ganslider: How users control generative models for images using multiple sliders with and without feed-forward information. In *CHI '22: CHI Conference on Human Factors in Computing Systems, New Orleans, LA, USA, 29 April 2022 - 5 May 2022*, pages 569:1–569:15. ACM.

Hai Dang, Lukas Mecke, Florian Lehmann, Sven Goller, and Daniel Buschek. 2022b. How to prompt? opportunities and challenges of zero- and few-shot learning for human-ai interaction in creative applications of generative models. *CoRR*, abs/2209.01390.

Laurie P Dringus. 1991. *A study of delayed-time and real-time text-based computer-mediated communication systems on group decision-making performance*. Nova University.

Starkey Duncan. 1972. Some signals and rules for taking speaking turns in conversations. *Journal of personality and social psychology*, 23(2):283.

Olga Egorow and Andreas Wendemuth. 2022. On emotions as features for speech overlaps classification. *IEEE Trans. Affect. Comput.*, 13(1):175–186.

Jonathan Ehret, Andrea Bönsch, Patrick Nossol, Cosima A. Ermert, Chinthusa Mohanathasan, Sabine J. Schlittmeier, Janina Fels, and Torsten W. Kuhlen. 2023. Who's next?: Integrating non-verbal turn-taking cues for embodied conversational agents. In *Proceedings of the 23rd ACM International Conference on Intelligent Virtual Agents, IVA 2023, Würzburg, Germany, September 19-22, 2023*, pages 27:1–27:8. ACM.

John J Godfrey, Edward C Holliman, and Jane McDaniel. 1992. Switchboard: Telephone speech corpus for research and development. In *Acoustics, speech, and signal processing, ieee international conference on*, volume 1, pages 517–520. IEEE Computer Society.

Julia A Goldberg. 1990. Interrupting the discourse on interruptions: An analysis in terms of relationally neutral, power-and rapport-oriented acts. *Journal of pragmatics*, 14(6):883–903.

Bettina Heinz. 2003. Backchannel responses as strategic responses in bilingual speakers' conversations. *Journal of pragmatics*, 35(7):1113–1142.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. Lora: Low-rank adaptation of large language models. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Jie Huang and Kevin Chen-Chuan Chang. 2023. Towards reasoning in large language models: A survey. In *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 1049–1065. Association for Computational Linguistics.

Zainab Iftikhar, Yumeng Ma, and Jeff Huang. 2023. "together but not together": Evaluating typing indicators for interaction-rich communication. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems, CHI 2023, Hamburg, Germany, April 23-28, 2023*, pages 724:1–724:12. ACM.

Kathrin Janowski and Elisabeth André. 2018. Decision-theoretic personality-based reasoning about turn-taking conflicts. In *Proceedings of the 18th International Conference on Intelligent Virtual Agents, IVA 2018, Sydney, NSW, Australia, November 05-08, 2018*, pages 349–350. ACM.

Peiling Jiang, Jude Rayan, Steven P. Dow, and Haijun Xia. 2023. Graphologue: Exploring large language model responses with interactive diagrams. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology, UIST 2023, San Francisco, CA, USA, 29 October 2023- 1 November 2023*, pages 3:1–3:20. ACM.

Chang-Min Kim, Hyeon-Beom Yi, Ji-Won Nam, and Geehyuk Lee. 2017. Applying real-time text on instant messaging for a rapid and enriched conversation experience. In *Proceedings of the 2017 Conference on Designing Interactive Systems, DIS '17, Edinburgh, United Kingdom, June 10-14, 2017*, pages 625–629. ACM.

Divesh Lala, Koji Inoue, and Tatsuya Kawahara. 2019. Smooth turn-taking by a robot using an online continuous model to generate turn-taking cues. In *International Conference on Multimodal Interaction, ICMI 2019, Suzhou, China, October 14-18, 2019*, pages 226–234. ACM.

Md Tahmid Rahman Laskar, Sawsan Alqahtani, M Saiful Bari, Mizanur Rahman, Mohammad Abdullah Matin Khan, Haidar Khan, Israt Jahan, Amran Bhuiyan, Chee Wei Tan, Md Rizwan Parvez, Enamul Hoque, Shafiq Joty, and Jimmy Huang. 2024. A systematic survey and critical review on evaluating large language models: Challenges, limitations, and recommendations. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 13785–13816, Miami, Florida, USA. Association for Computational Linguistics.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Zhaoyang Liu, Yinan He, Wenhai Wang, Weiyun Wang, Yi Wang, Shoufa Chen, Qinglong Zhang, Zeqiang Lai, Yang Yang, Qingyun Li, Jiashuo Yu, Kunchang Li, Zhe Chen, Xue Yang, Xizhou Zhu, Yali Wang, Limin Wang, Ping Luo, Jifeng Dai, and Yu Qiao. 2023. Interngpt: Solving vision-centric tasks by interacting with chatbots beyond language. *CoRR*, abs/2305.05662.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Ryan Louie, Andy Coenen, Cheng Zhi Huang, Michael Terry, and Carrie J. Cai. 2020. Novice-ai music co-creation via ai-steering tools for deep generative models. In *CHI '20: CHI Conference on Human Factors* *in Computing Systems, Honolulu, HI, USA, April 25-30, 2020*, pages 1–13. ACM.

Ziyang Ma, Yakun Song, Chenpeng Du, Jian Cong, Zhuo Chen, Yuping Wang, Yuxuan Wang, and Xie Chen. 2024. Language model can listen while speaking. *Preprint*, arXiv:2408.02622.

Damien Masson, Sylvain Malacria, Géry Casiez, and Daniel Vogel. 2024. Directgpt: A direct manipulation interface to interact with large language models. In *Proceedings of the CHI Conference on Human Factors in Computing Systems, CHI 2024, Honolulu, HI, USA, May 11-16, 2024*, pages 975:1–975:16. ACM.

Bonan Min, Hayley Ross, Elior Sulem, Amir Pouran Ben Veyseh, Thien Huu Nguyen, Oscar Sainz, Eneko Agirre, Ilana Heintz, and Dan Roth. 2024. Recent advances in natural language processing via large pre-trained language models: A survey. *ACM Comput. Surv.*, 56(2):30:1–30:40.

Meriam Moujahid, Helen F. Hastie, and Oliver Lemon. 2022. Multi-party interaction with a robot receptionist. In *ACM/IEEE International Conference on Human-Robot Interaction, HRI 2022, Sapporo, Hokkaido, Japan, March 7 - 10, 2022*, pages 927–931. IEEE / ACM.

Kumiko Murata. 1994. Intrusive or co-operative? a cross-cultural study of interruption. *Journal of pragmatics*, 21(4):385–400.

Humza Naveed, Asad Ullah Khan, Shi Qiu, Muhammad Saqib, Saeed Anwar, Muhammad Usman, Nick Barnes, and Ajmal Mian. 2023. A comprehensive overview of large language models. *CoRR*, abs/2307.06435.

Maike Paetzel-Prüsmann and James Kennedy. 2023. Improving a robot's turn-taking behavior in dynamic multiparty interactions. In *Companion of the 2023 ACM/IEEE International Conference on Human-Robot Interaction, HRI 2023, Stockholm, Sweden, March 13-16, 2023*, pages 411–415. ACM.

Xingang Pan, Ayush Tewari, Thomas Leimkühler, Lingjie Liu, Abhimitra Meka, and Christian Theobalt. 2023. Drag your GAN: interactive point-based manipulation on the generative image manifold. In *ACM SIGGRAPH 2023 Conference Proceedings, SIGGRAPH 2023, Los Angeles, CA, USA, August 6-10, 2023*, pages 78:1–78:11. ACM.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Joon Sung Park, Joseph C. O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology, UIST 2023, San Francisco, CA, USA, 29 October 2023- 1 November 2023*, pages 2:1–2:22. ACM.

Mridumoni Phukon, Abhishek Shrivastava, and Bruce Balentine. 2022. Can VUI turn-taking entrain user behaviours?: Voice user interfaces that disallow overlapping speech present turn-taking challenges. In *Proceedings of the 13th Indian Conference on Human Computer Interaction, IndiaHCI 2022, Hyderabad, India, November 9-11, 2022*, pages 42–56. ACM.

Martin Podlubny, John Rooksby, Mattias Rost, and Matthew Chalmers. 2017. Synchronous text messaging: A field trial of curtains messenger. *Proc. ACM Hum. Comput. Interact.*, 1(CSCW):86:1–86:20.

Mark Rejhon, Christian Vogler, Norman Williams, and Gunnar Hellström. 2013. Standardization of real-time text in instant messaging. In *The 15th International ACM SIGACCESS Conference on Computers and Accessibility, ASSETS '13, Bellevue, WA, USA, October 21-23, 2013*, pages 66:1–66:2. ACM.

Harvey Sacks, Emanuel A Schegloff, and Gail Jefferson. 1974. A simplest systematics for the organization of turn-taking for conversation. *language*, 50(4):696–735.

Emanuel A Schegloff. 2000. Overlapping talk and the organization of turn-taking for conversation. *Language in society*, 29(1):1–63.

Sakib Shahriar and Kadhim Hayawi. 2023. Let's have a chat! A conversation with chatgpt: Technology, applications, and limitations. *CoRR*, abs/2302.13817.

Gabriel Skantze. 2021a. Turn-taking in conversational systems and human-robot interaction: A review. *Comput. Speech Lang.*, 67:101178.

Gabriel Skantze. 2021b. Turn-taking in conversational systems and human-robot interaction: A review. *Comput. Speech Lang.*, 67:101178.

Jacob Solomon, Mark W. Newman, and Stephanie D. Teasley. 2010. Speaking through text: the influence of real-time text on discourse and usability in IM. In *Proceedings of the 2010 International ACM SIGGROUP Conference on Supporting Group Work, GROUP 2010, Sanibel Island, Florida, USA, November 6-10, 2010*, pages 197–200. ACM.

Tanya Stivers, Nicholas J Enfield, Penelope Brown, Christina Englert, Makoto Hayashi, Trine Heinemann, Gertie Hoymann, Federico Rossano, Jan Peter De Ruiter, Kyung-Eun Yoon, and 1 others. 2009. Universals and cultural variation in turn-taking in conversation. *Proceedings of the National Academy of Sciences*, 106(26):10587–10592.

Sangho Suh, Bryan Min, Srishti Palani, and Haijun Xia. 2023. Sensecape: Enabling multilevel exploration and sensemaking with large language models. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology, UIST 2023, San Francisco, CA, USA, 29 October 2023- 1 November 2023*, pages 1:1–1:18. ACM.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford alpaca: An instruction-following llama model. https://github.com/tatsu-lab/stanford_alpaca.

Maria Dolores C Tongco. 2007. Purposive sampling as a tool for informant selection.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Jules White, Quchen Fu, Sam Hays, Michael Sandborn, Carlos Olea, Henry Gilbert, Ashraf Elnashar, Jesse Spencer-Smith, and Douglas C. Schmidt. 2023. A prompt pattern catalog to enhance prompt engineering with chatgpt. *CoRR*, abs/2302.11382.

Sarita Yardi. 2006. The role of the backchannel in collaborative learning environments. In *Making a Difference...: Proceedings of the 7th International Conference for the Learning Sciences, ICLS 2006, Bloomington, IN, USA, June 27 - July 1, 2006*. International Society of the Learning Sciences.

Ann Yuan, Andy Coenen, Emily Reif, and Daphne Ippolito. 2022. Wordcraft: Story writing with large language models. In *IUI 2022: 27th International Conference on Intelligent User Interfaces, Helsinki, Finland, March 22 - 25, 2022*, pages 841–852. ACM.

Xinrong Zhang, Yingfa Chen, Shengding Hu, Xu Han, Zihang Xu, Yuanwei Xu, Weilin Zhao, Maosong Sun, and Zhiyuan Liu. 2024. Beyond the turn-based game: Enabling real-time conversations with duplex models. *CoRR*, abs/2406.15718.

Qi Zhou, Bin Li, Lei Han, and Min Jou. 2023. Talking to a bot or a wall? how chatbots vs. human agents affect anticipated communication quality. *Comput. Hum. Behav.*, 143:107674.

Don H Zimmermann and Candace West. 1996. Sex roles, interruptions and silences in conversation. In *Amsterdam Studies in the Theory and History of Linguistic Science Series 4*, pages 211–236. John Benjamins BV.

A Related Work

A.1 Large Language Models, Text-based Conversational Agent, Interactive Designs

Recent advancements have led to the widespread development of Large Language Models, or text-based conversational agents (LLMs). LLMs are increasingly being applied across various domains due to their interactivity (Min et al., 2024; Shahriar and Hayawi, 2023; Dang et al., 2022b; White et al., 2023; Park et al., 2023). These interactions typically rely on verbose textual prompting, sometimes complemented by graphical manipulations such as buttons or mouse pointer movements.

A.1.1 Verbose Textual Prompting

The primary mode of interaction with LLMs is through a prompting interface (Chang et al., 2024). Users craft specific prompts to guide LLMs in performing tasks such as email generation, text summarization, or question-answering. Additionally, users can engage in dialogue-like interactions, allowing for natural language conversations with the models. Several widely adopted techniques enhance textual prompting. For instance, the Chain-of-Thought method enables LLMs to provide step-by-step reasoning (Huang and Chang, 2023; Wei et al., 2022), while Multi-Turn instructions allow for iterative problem-solving by incorporating user feedback into subsequent prompts (Naveed et al., 2023). These approaches align with a strict turn-taking conversational paradigm, where users input a prompt, wait for the model's response, and repeat the process. However, few studies have explored interaction paradigms that move beyond traditional turn-taking in text-based human-AI exchanges. Our work introduces overlapping capabilities to LLMs, broadening the interaction design space by enabling overlapping functionality. This enables forms of interaction that expands the possibilities for "how" users and LLM can interact with.

A.1.2 Graphical Manipulations Combined with Textual Prompting

Many integrations of LLMs incorporate graphical elements (Jiang et al., 2023; Suh et al., 2023), including widgets like buttons and sliders to trigger predefined textual or system prompts. For example, buttons are often used as shortcuts for tasks such as editing text or generating code (Yuan et al., 2022; Clark et al., 2018). OpenAI's ChatGPT API, for instance, includes a "stop generating" button, which requires users to use their mouse to pause the model's response. In comparison, our proposed interface enables users to stop the chatbot by simply sending a textual command that overlaps with the ongoing interaction. In addition, sliders are commonly utilized to adjust model parameters, allowing users to modify continuous variables that affect the generation of outputs like images or music

(Dang et al., 2022a; Louie et al., 2020). In addition, gestures and physical metaphors are sometimes employed to refine LLM outputs. For example, pointing to a specific area can highlight elements of an image or guide the model to regenerate only a selected part (Liu et al., 2023). Similarly, dragging gestures can be used to adjust spatial attributes of an image, such as pose, facial expressions, or layout (Masson et al., 2024; Pan et al., 2023). Our proposed interface eliminates the need for buttons, sliders, or gestures. Instead, it relies exclusively on text-based interactions, such as stopping the LLM's response by overlapping functionality.

A.2 Overlap in Human Communication, Cooperative Overlap and Competitive Overlap

Human-to-human conversations generally follow a pattern where one person speaks at a time, yet overlap in speech is a frequent occurrence (Skantze, 2021b; Zimmermann and West, 1996). It is important to recognize that these overlaps should not merely be viewed as failures in turn-taking, as they often fulfill important functions and contribute to the smooth flow of interaction (Coates, 1994). Overlapping speech is not always a sign of dominance or unfriendliness (Goldberg, 1990). Previous studies have identified two distinct types of overlap: cooperative and competitive (Schegloff, 2000; Murata, 1994; Egorow and Wendemuth, 2022). Cooperative overlap involves both speakers contributing to the conversation collaboratively, without competing for control. A common example of this is back-channeling (Yardi, 2006; Heinz, 2003), where the listener provides brief, often subtle vocalizations such as "mm hmm," "uh huh," or "yeah." These responses, although frequent, are not typically considered full "turns" in conversation. Another form of cooperative overlap is terminal overlap, where the listener anticipates the speaker's turn ending and begins to speak before the turn is fully completed. Conversely, competitive overlap occurs when speakers vie for control of the conversation, with one eventually needing to relinquish their turn. Unlike cooperative overlap, competitive overlap requires a resolution mechanism to determine which speaker should continue (Goldberg, 1990; Skantze, 2021b). Previous research highlights that while overlaps can be objectively identified in a corpus, interruptions require interpretation, as one speaker is seen as violating the other's right to speak (Bennett, 1978).

A.3 Real-time Text Messaging

Research on real-time messaging in text-based interaction has uncovered various effects on collaboration and communication (Rejhon et al., 2013; Iftikhar et al., 2023). Some studies have shown that when messages are visible to interlocutors as they are being typed, user coordination improves and message editing decreases (Solomon et al., 2010; Dringus, 1991). Field trials have indicated that synchronous communication can foster greater cooperation and engagement, particularly in close relationships (Podlubny et al., 2017). Further studies have suggested that real-time messaging enhances conversational experiences by minimizing silence and incorporating nonverbal cues, such as pauses and typing speed, into the communication process (Kim et al., 2017). These findings illustrate the positive impact of real-time messaging, highlighting its potential to facilitate smoother interactions. Our study differs by enabling real-time text-based messaging between a human and an LLM-powered chatbot, where the chatbot is inherently capable of managing overlap.

B Pilot Study

In this study, we conducted a pilot study where seven pairs of participants engaged in text-based conversations using a real-time chat interface that allowed simultaneous typing and message visibility. We focused on a task that could induce users to naturally overlap with each other in text-based interaction. As chat conversations can vary depending on the relationship between the partners, we gathered participants by purposive sampling (Tongco, 2007). A total of 14 participants took part in discussions. Their average age was 26 (SD = 2.09), and 8 of them were female and 6 male. 12 participants were native South Korean speakers, 1 participant was a native German speaker, and 1 participant was a native Chinese speaker. These participants formed seven pairs, with six pairs conversing in Korean and one pair (German and Chinese) using English. The pairs were intentionally made up of individuals with different levels of familiarity, including close friends, colleagues, and strangers.

To encourage conversation, we instructed the pairs to decide on things about a group retreat workshop. They had to decide on three songs to listen to, three dinner menus, and three movies to watch. They were given a 10-minute time limit for these decisions. After the discussion, participants were asked to complete open-ended questions about their overall experience and their intention to use it in the future. We interviewed them when more detailed explanations were needed in open-ended responses. All conversations were recorded, and the participants' typing logs were saved as files, with their consent.

We collected three types of data: open-ended survey responses, interview transcripts and recorded videos. By observing the recorded videos, we were able to determine the types of overlapping behaviors that occurred. By having the first author and an independent researcher thematically analyze the open-ended responses and interview transcripts (Boyatzis, 1998), we were able to understand the intentions behind the overlapping behaviors.

B.1 Findings

First, we observed that participants frequently engaged in overlapping behaviors. Specifically, participants overlapped with their interlocutor's typing by starting to type even before the other person finished typing. All participants showed and acknowledged this behavior. Participants reported their intentions as follows, which were related to cooperative overlap (Section A.2).

1. **Preemptive response:** Predicting the end of the turn and starting to reply before it is completed. For example, participants preemptively gave answers to the interlocutor's questions as in "A: Do you remember who the movie direc–" "B: You mean Bong Jun Ho?"

2. **Backchanneling:** Showing one is paying attention or giving instant agreement on others' perspective. For example, participants gave backchanneling to the interlocutors as in "A: Today I went to–" "B: yeah."

Second, we observed that participants frequently engaged in **deletion** behaviors. Specifically, to resolve **interruption** from their interlocutor, participants deleted their typed messages. This happened when participants encountered simultaneous typing by interlocutors. As mentioned in Section A.2 about competitive overlap, the concept of interruptions necessitates some level of interpretation, where one participant is perceived as violating the other's right to speak. We interpreted this deletion behavior as the resolution mechanism for interruption, to determine which speaker should continue. All participants demonstrated and acknowledged this counteracting behavior to interruptions. Participants reported their reasons as follows, which are related to competitive overlap (Section A.2).

1. Adjusting responses based on the interlocutor's actions, such as transitioning topics when there is a mismatch or addressing questions and refutations during simultaneous typing.

2. Removing brief real-time feedback, including backchanneling cues, typos, or profanity.

In addition, participants perceived conversations as authentically similar to a real conversation. They noted that the flow of conversation with overlapping was uninterrupted, enhancing the presence of the interlocutor and fostering greater engagement. *"It made me focus more on the chat because I could see what the other person was typing (and they might even delete it)." (P3)*; *"When the content I was about to type matched what the other person was typing, it felt like a boost in closeness." (P5)* The prevailing sentiment was that overlapping effectively promoted the exchange of opinions: *"It felt like the limitations of online discussion were reduced." (P1)*

However, certain participants experienced a psychological burden due to the transparency of their thought processes while typing (Podlubny et al., 2017). *"Since everything I typed was visible to the other person in real time, even what I typed unconsciously, I became more cautious." (P8)*; *"If I had to chat for an extended period with this interface, I think I would feel fatigued, as if my initial thoughts were being monitored." (P4)* Some participants expressed a preference for using this interface exclusively in intimate relationships: *"I would use it with close friends, but probably not with people I am not as familiar with." (P12)*

In conclusion, these findings reveal that people instinctively engage in overlapping during text-based interactions – something traditional chatbot systems don't allow. We have grown so accustomed to strict turn-taking with chatbots that we may not have realized what has been missing. When given the chance to overlap, people naturally embrace it, opening up possibilities in chatbot design. This naturally occurred conversational behavior presented a new technical challenge where text-based chatbot cannot naturally overlap people, which we solved by finetuning LLM with publicly available datasets customized for overlapping.

C Details of Tasks

To enable overlapping interactions in LLMs based on human conversational behaviors, we introduce a three-stage prediction framework consisting of the following tasks.

Timing Prediction (When to Overlap?) The model first determines whether to overlap with the user's ongoing utterance or wait until they complete their turn. Given the user's typed tokens, the model selects between two options: `[Await]`, where the model does not interrupt and waits for the user to finish, or `[Overlap]`, where the model initiates an overlapping response.

Action Selection (What to Do When Overlapping?) If the model selects `[Await]`, no output is generated. If `[Overlap]` is chosen, the model must further decide on the appropriate dialogue action: `[Understanding]`, which signals active listening without disrupting the user's speech (e.g., *Uh-huh, Yeah*), or `[Answer]`, which provides a preemptive response before the user's utterance is fully completed.

Utterance Generation (What to Say?) Based on the selected action, the model generates the corresponding response. If `[Understanding]` was chosen, the model produces a brief backchanneling utterance (e.g., *Mm-hmm.*). If `[Answer]` was chosen, the model generates a relevant response to the user's unfinished input. While the second task (Action Selection) determines only the action token, the third task (Utterance Generation) ensures that the generated response aligns with the selected action.

D Training Details

We finetuned the Llama 3 8B instruct model (AI@Meta, 2024) on 1 NVIDIA RTX A6000 GPU. We employed QLoRA with 4-bit quantization, setting the LoRA rank (Hu et al., 2022) and alpha value to 16, and targeted all attention and feed-forward layers. The model was trained with a maximum sequence length of 2048 tokens, using a batch size of 16 with gradient accumulation steps of 4. We used the AdamW 8-bit optimizer (Loshchilov and Hutter, 2019) and implemented a learning schedule with 30 warmup steps over 300 total training steps. Training was conducted using 3-fold cross-validation, with each fold taking approximately 4 hours. We applied early stopping with a patience of 5 steps based on validation loss and saved model checkpoints every 100 steps. Mixed precision training was used with bfloat16 where supported, falling back to float16 otherwise.

E User Evaluation Details

A total of 18 participants were recruited by voluntarily responding to the experiment participation post on the university's website. 10 of them were South Koreans, 6 of them Indonesians, 1 of them Nepali, and 1 of them is Vietnamese. Their average age was 23 (SD=2.42), and 9 of them were female and 9 were male. They all self-reported frequent usage of the OpenAI chatGPT website. As compensation for their participation, all participants were paid 50K KRW. Each experiment lasted approximately 60 min on average. All sessions were conducted remotely using Google Meets with audio and video recordings and were conducted in Korean or English, based on the nationality. The overall procedure of our study was conducted after obtaining IRB approval from the university.

Before the experiments began, participants received a detailed explanation of how to use OverlapBot and the conventional chat system. The tutorial introduced key functionalities of OverlapBot, such as its ability to display real-time typing and provide understanding reactions (e.g., "yeah") or answers before the participant's utterance was complete. Participants were also instructed on how to interrupt the chatbot's response. For the conventional chat system, they were informed that neither they nor the chatbot could see each other's typing in real-time. During the tutorial, participants were given examples of potential conversation topics, such as discussing hypothetical scenarios like "Would you rather speak every language or communicate with animals?" or "Would you rather die in 20 years with no regrets or live to 100 with a lot of regrets?" The explanation and tutorial session was conducted for approximately 10 minutes.

Following the tutorial, each participant engaged in a 10-minute conversation with the conventional chat system and then with the OverlapBot, with the order randomized. Participants were free to choose any topic including the hypothetical scenarios for their interactions in English, ensuring that the conversations were natural and varied. After completing the interactions, participants were asked to fill out an open-ended survey and participate in a semi-structured interview to gather qualitative feedback on their experience. The survey and interview included questions designed to explore participants' perceptions and preferences regarding the two chatbots. Key questions addressed the main differences participants noticed between the OverlapBot and conventional chatbots, their overall impressions of each chatbot, and specific aspects of the OverlapBot that they found most useful or convenient. Participants were also asked to indicate which interface they preferred and to explain their reasons. Additionally, the survey inquired about any difficulties or discomfort experienced while using the Overlapbot.

We collected four types of data: open-ended survey responses, interview responses, recorded videos, and conversation logs. We analyzed participants' conversation logs to conduct a quantitative comparison between OverlapBot and a conventional chatbot. We utilized open-ended survey responses and transcribed interview responses to analyze general impressions of OverlapBot compared to the conventional chat system. For the analysis, thematic analysis (Boyatzis, 1998) was conducted by three authors. We repeatedly observed recorded videos to learn new interaction patterns users showed using OverlapBot. Three authors conducted a thematic analysis (Boyatzis, 1998) of the transcribed interview and open-ended survey responses to gather insights on participants' impressions of OverlapBot.

F Discussion

F.1 Relationships

The conversational relationship between humans and AI also requires further exploration. Participants in our formative study observed that the transparency of typing might feel more appropriate in casual relationships, such as with close friends, but less suitable in hierarchical or unfamiliar relationships: *"I would use it with close friends, but probably not with people I am not as familiar with." (P12)* This suggests that the relational context of human-AI interactions — whether focused on companionship, practical assistance, or other roles — may influence how overlapping features are perceived and received. For instance, socially isolated individuals, such as the elderly or those living alone, may appreciate OverlapBot's overlapping features as part of its role as a conversational partner. On the other hand, users engaging with AI in professional or hierarchical settings may favor stricter turn-taking norms. These nuanced preferences highlight the need to design overlapping interactions that are sensitive to the role and context of the relationship.

F.2 Rethinking the Necessity of Prompting Design

Numerous studies have shown that LLMs produce varying outputs based on the prompts they receive, prompting users to carefully craft precise prompts. Our findings suggest that overlap may reduce the need for highly detailed prompts. By observing the user's input in real time as they type, the LLM can infer intent without relying on a fully developed prompt. As the LLM anticipates the user's intended response, users can provide immediate confirmation or correction. However, while some participants appreciated this as a convenient and effective feature, others found it uncomfortable, viewing the typing process as a critical step for clarifying and organizing their thoughts. This feedback indicates that overlapping interface should offer users control, enabling them to adjust the visibility of their typing to match their interaction preferences.

F.3 User-Customizable Overlap

When designing overlapping chatbots, it is essential to consider user preferences and provide adjustable settings that accommodate diverse interaction styles. Some users, particularly those accustomed to signaling the end of their turn with the Enter key, may find the chatbot's proactive behavior intrusive or disruptive. To address this, the chatbot must carefully determine the right moment to offer a preemptive response, ensuring users feel they have communicated enough before being interrupted. As one participant shared: *"I wish it would let me finish what I have to say. (...) I feel like I have to finish speaking quickly or say something just to keep up, and that made me feel uncomfortable and uneasy." (P16)*

The absence of non-verbal cues in text-based interactions complicates this further. As another participant noted: *"In human conversations, you can usually guess from my facial expressions or tone, but here, it only relies on the text, so I thought there might be more room for error." (P12)* One possible solution is to adjust overlap frequency based on the user's typing speed. For example, slower typists may benefit from more frequent overlaps to maintain flow, whereas faster typists might find them disruptive.

F.4 Culturally Adaptive Overlap

When designing overlapping chatbots, it is essential to account for cultural differences, as these significantly influence how overlapping is perceived (Stivers et al., 2009; Clancy et al., 1996). In some cultures, conversational overlap is considered a sign of active engagement and is viewed positively. Users from these backgrounds may appreciate chatbot's overlap as a natural part of the interaction. Conversely, in cultures that prioritize clear turn-taking, such interruptions could be seen as rude or disruptive. This cultural variability underscores the need for configuration to be adaptable. By learning and adjusting to the conversational norms of individual users over time, the AI chatbot can better align its behavior with the user's cultural background.