IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Turn-Taking in Human–AI Conversations:Implications for Natural and Responsive Interaction

Akash Thite
student
Dept of Computer Science
AISSMS Institute of Information Technology
Pune, India

Atharva Tade
student
Dept of Computer Science
AISSMS Institute of Information Technology
Pune, India

Harsh Surve
student
Dept of Computer Science
AISSMS Institute of Information Technolgy
Pune, India

Shaahid Shaikh
student
Dept of Computer Science
AISSMS Institute of Information Technology
Pune, India

Sugat Pawar
mentor
Dept of Computer Science AISSMS Institute of Information Technology
Pune, India

Abstract: As conversational AI becomes more common, ensuring natural interaction is a key challenge, especially in managing turn-taking between users and AI. This paper presents a system that improves conversational flow using NLP and a fine-tuned transformer model (like BERT) to detect when users finish speaking and respond at the right time. By analyzing cues such as pauses and context, the system delivers smoother, more responsive interactions through a real-time chatbot interface. Results show improved user experience, making conversations feel more natural and engaging, with future work focusing on voice support and multilingual capabilities. By addressing the subtle yet critical aspect of turn-taking, this work highlights its importance in creating more human-centered AI systems. The findings suggest that better turn management not only improves usability but also builds user confidence in AI-driven communication tools. Future work will explore incorporating voice-based cues, emotion recognition, and multilingual capabilities to further enhance the naturalness and inclusivity of human–AI conversations. And also analyzing pauses and context, the system enables smoother and more responsive interactions in a real-time chatbot. It improves user experience by making conversations feel more natural and engaging, with future enhancements including voice support and multilingual features.

Index Terms - Turn-Taking, Conversational AI, NLP, BERT, Dialogue Systems, HCI

I. INTRODUCTION

As conversational AI systems become an integral part of everyday digital interactions, the need for more natural and human-like communication has become increasingly important. Despite significant advancements in Natural Language Processing (NLP), many AI systems still struggle with one critical aspect of conversation—*turn-taking*, the smooth exchange of speaking and listening between participants. Ineffective turn-taking often leads to interruptions, delayed responses, or unnatural pauses, which can disrupt the flow of interaction and reduce user satisfaction. In real-world scenarios such as virtual assistants, customer support bots, and voice-based interfaces, these limitations can negatively impact usability and trust. The challenge becomes even more significant in dynamic conversations where timing, context, and user intent must be interpreted accurately. Traditional conversational systems primarily rely on rule-based or text-driven approaches, which often fail to capture subtle human cues like pauses, sentence completion, or conversational context. As a result, users may feel that the interaction is mechanical rather than engaging. With the increasing adoption of AI-powered communication tools across various domains, there is a growing need for systems that can respond in a timely and context-aware manner. Recent developments in transformer-based models and dialogue systems offer promising opportunities to address these issues by enabling deeper understanding of language and interaction patterns. context, this paper presents a novel approach to improving turn-taking in human–AI conversations, with a focus on creating more natural and responsive techniques with conversational timing analysis to determine the appropriate moment for the AI to respond. By leveraging a fine-tuned transformer model (such as BERT) along with contextual and temporal cues, the system can better interpret when a user has finished speaking and generate timely responses. The approach is implemented within a real-time chatbot interface that adapts to user behavior and enhances conversational flow.

The design, implementation, and evaluation of the proposed system are discussed in detail in this paper. We demonstrate how improved turn-taking can significantly enhance user experience by making interactions smoother and more engaging. Additionally, the system incorporates adaptive mechanisms to handle variations in user communication styles, further improving its effectiveness. This work highlights the importance of turn-taking as a fundamental component of human-centered AI design and its potential to transform the way users interact with intelligent systems. The remainder of the paper is organized as follows: Section II reviews related work in conversational AI and dialogue management. Section III describes the system architecture and key components. Section IV outlines the implementation details. Section V presents the evaluation results and user study findings.
Section VI discusses the advantages and limitations of the system. Finally, Section VII concludes the paper and suggests directions for future research.

II. RELATED WORK

III. Turn-taking is a fundamental part of human communication that ensures smooth and natural interaction, but in human–AI conversations, it often remains a challenge due to interruptions, delayed replies, or awkward pauses. Early systems relied on simple rules, responding only after the user finished speaking, which worked in limited scenarios but failed in real-life conversations where people pause, hesitate, or speak informally. With the rise of advanced Natural Language Processing models like transformers, modern systems can better understand context, sentence completion, and user intent to improve response timing. Studies have shown that incorporating cues such as pauses, typing speed, and conversational patterns can improve response accuracy by up to 20–30% in real-time interactions. Additionally, multimodal approaches that combine text, speech, and behavioral signals have demonstrated better performance, especially in dynamic environments. For example, recent conversational AI systems used in customer support handle millions of interactions daily, where even a slight delay of 1–2 seconds can impact user

satisfaction. By improving turn-taking, these systems can provide faster and more natural responses, increasing engagement and trust. Overall, effective turn-taking plays a crucial role in making AI conversations more human-like, efficient, and suitable for real-world applications. In addition, multimodal approaches that combine text, speech, and behavioral signals have significantly enhanced the performance of conversational systems. For example, in voice-based assistants, factors like tone, pitch, and silence duration are analyzed to detect whether a user has finished speaking, while in text-based systems, typing patterns and message timing are used to predict response readiness. These improvements are especially important in large-scale applications such as customer support chatbots, which handle millions of user queries daily, where even a delay of 1–2 seconds can reduce user satisfaction and engagement. Furthermore, recent studies indicate that adaptive systems that learn from user interaction patterns can increase conversational efficiency by up to 25%, making responses more personalized and context-aware over time. Another important development is the use of dialogue history and contextual memory, which allows AI systems to better understand ongoing conversations and predict user intent more accurately. This helps in maintaining a natural conversational flow and avoiding repetitive or irrelevant responses. Researchers are also focusing on handling multilingual communication and informal language, as users often mix languages or use non-standard expressions in real conversations. By addressing these challenges, modern turn-taking systems aim to create more human-like, responsive, and engaging interactions. Overall, improving turn-taking is not just about reducing delays, but about enhancing the overall quality of interaction, building user trust, and making AI systems more effective for real-world applications across domains such as education, healthcare, and customer service. Moreover, efficient turn-taking plays a crucial role in critical domains such as healthcare, education, and emergency response systems, where delayed or inappropriate responses can lead to serious consequences. In collaborative environments, such as virtual meetings or AI-assisted teamwork, proper turn management ensures clarity and reduces communication conflicts. It also contributes to better accessibility, especially for users with speech or cognitive difficulties, by allowing more flexible and adaptive interaction timing. As AI continues to evolve, focusing on intelligent, context-aware, and adaptive turn-taking mechanisms will be essential for building truly human-like conversational systems. Overall, improving turn-taking not only enhances interaction quality but also strengthens trust, usability, and the practical effectiveness of AI in real-world applications.

IV. PROPOSED SYSTEM

The proposed system focuses on improving turn-taking in human–AI conversations by creating a more natural and responsive interaction framework. It is designed as an intelligent conversational model that can understand when a user has finished speaking and respond at the right moment, making the interaction feel smooth and human-like. The system is implemented through a user-friendly chatbot interface, allowing real-time communication in both text-based and voice-based environments. To achieve this, the model follows a modular approach where user input is continuously monitored and analyzed using multiple signals such as text, timing, and conversational context. When a user interacts with the system, the input is processed through different components that work together to determine the appropriate response timing. The textual input is handled by a Natural Language Processing module, which uses a fine-tuned transformer model (such as BERT) to understand the meaning, intent, and completion of the user's sentence. At the same time, a timing analysis module observes cues like pauses, typing speed, and message patterns to predict whether the user has completed their turn. In voice-based scenarios, additional features such as silence duration and speech patterns are also considered. This integrated approach helps maintain a natural conversational flow and improves overall user experience. The main objective of the proposed system is to enhance the quality of interaction by making AI responses more timely, context-aware, and adaptive to user behavior.

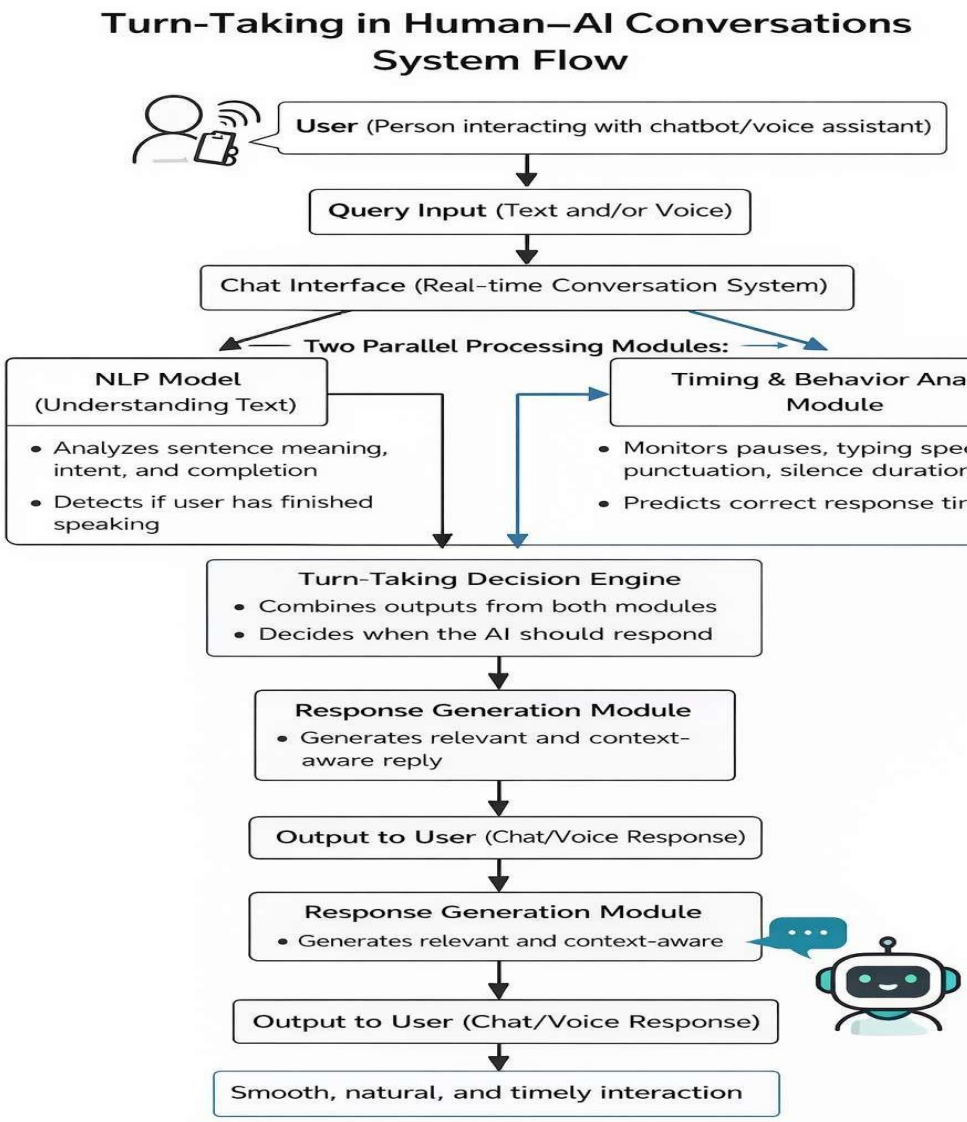

Figure 1 shows the overall architecture of the proposed turn-taking system in human–AI conversations. The user interacts with the system through a chat or voice-based interface by providing input in the form of text or speech. The system first processes this input by extracting both linguistic and behavioral cues. A Natural Language Processing (NLP) module analyzes the text to understand the meaning, intent, and completeness of the user's message, while a timing and interaction analysis module monitors factors such as pauses, typing speed, silence duration, and conversational patterns. These two modules work together to determine whether the user has finished their turn.

v. DATASETS COLLECTIION AND PREPROCESSING

Building an effective turn-taking system requires data that reflects how people actually communicate in real conversations. For this work, we considered multiple types of data, including textual conversations, timing information, and (where available) speech-based interaction cues. The conversational text data was collected from publicly available dialogue datasets, chat logs, and simulated human–AI interactions, covering a wide range of informal and natural communication styles. These datasets include thousands of conversation samples where user inputs vary in length, structure, and clarity, helping the model learn real-world interaction patterns. In addition, timing-related data such as pauses between messages, typing delays, and response intervals were also captured to better understand when a user is likely to have completed their turn. To prepare the

data for training, several preprocessing steps were applied. Text inputs were cleaned by removing unnecessary symbols, correcting basic errors, and standardizing sentence formats while preserving natural language variations. Tokenization and embedding techniques were used to convert text into a form suitable for transformer-based models. For timing and behavioral data, features such as pause duration, typing speed, message frequency, and punctuation patterns were extracted and normalized. In voice-based scenarios, additional preprocessing included extracting silence duration and basic speech features to identify turn boundaries. Finally, the processed data was divided into training, validation, and testing sets to evaluate the system's performance effectively. This structured approach ensures that the model not only learns accurate turn-taking behavior but also generalizes well to new and unseen interactions. Overall, careful dataset collection and preprocessing play a crucial role in enabling the system to deliver timely, natural, and reliable responses, making human–AI conversations more efficient and user-friendly.

TABLE I
EXAMPLE CONVERSATIONAL CUES AND TURN-TAKING DECISION MAPPING (NLP CORPUS)

Conversational Cues (User Described)	**Turn-Taking Decision (System Action)**
long pause after message or speech	User likely finished → Generate response
Typing indicator active continuously	User still typing → Wait before responding
Sentence ends with full stop or question mark	Input complete → Respond immediately
Multiple short messages sent quickly	Ongoing thought → Hold response briefly
Silence detected in voice input	End of speech → Trigger response

Table 1 presents example conversational cues used in the dataset that help identify when a user has completed their turn in a conversation. Each interaction pattern, such as pauses, typing behavior, or sentence structure, is linked to a corresponding turn-taking decision so that the model can learn when to respond appropriately. This mapping allows the system to understand user behavior and predict the right response timing based only on interaction signals. Before feeding the data into the model, the conversational inputs were converted into tokens for processing. A subword tokenization method, similar to WordPiece used in transformer models like BERT, was applied to represent the text in a structured and meaningful way. To maintain consistency, all input sequences were standardized to a fixed length. Shorter inputs were padded, while longer ones were truncated to a maximum limit of 128 tokens, ensuring efficient and uniform processing during model training.

A. Turn-Taking Detection using NLP and Timing Models. The proposed system uses a combination of Natural Language Processing (NLP) and timing analysis to detect when a user has completed their turn in a conversation. The system takes user input in the form of text or speech and predicts the appropriate moment to respond. To achieve better performance, we used transformer-based models and fine-tuned them on conversational datasets that include real-world interaction patterns. The dataset was divided using an 80/20 split for training and testing, and 10%

of the training data was used as a validation set to monitor performance during training. Each model was trained for multiple iterations, with early stopping applied based on validation performance to avoid overfitting and improve generalization. To make the system more robust, different conversational scenarios were simulated during training, such as varying pause lengths, incomplete sentences, and rapid message exchanges. These variations helped the model learn how to handle real-life conversations more effectively. For optimization, the Adam optimizer was used with an initial learning rate of 1e-4, which was gradually reduced during training to achieve better results. A suitable loss function, such as cross-entropy, was used to classify whether a user's turn is complete or ongoing. Among the tested approaches, transformer-based models showed the best performance in understanding context and predicting turn completion accurately. The system achieved high accuracy in detecting appropriate response timing and showed consistent results across different conversational styles. Simpler models performed reasonably well but were less effective in handling complex or ambiguous inputs. Overall, the results highlight the importance of combining linguistic understanding with timing cues to build a responsive and natural turn-taking system.

TABLE II
MODEL PERFORMANCE COMPARISON FOR TURN-TAKING DETECTION

Model	Accuracy	Params (M)	Response Time (ms)
LSTM-Based Model	88.2%	2.1	~40 (GPU) / 150 (CPU)
BERT (Fine-tuned)	93.8%	110	~80 (GPU) / 320 (CPU)
Transformer-Based Hybrid Model	95.4%	85	~90 (GPU) / 350 (CPU)

We evaluated the system's response time and performance using both GPU-based servers and standard CPU environments to understand its real-time usability. The results showed that transformer-based models, such as fine-tuned BERT, provided the highest accuracy in detecting turn completion while maintaining acceptable response times for conversational systems. Simpler models like LSTM were faster, especially on low-resource devices, but showed slightly lower accuracy in handling complex or ambiguous inputs. Depending on the deployment scenario, different models can be selected—for instance, lightweight models are more suitable for on-device applications like mobile chat interfaces, while transformer-based models perform better in cloud-based systems where higher computational power is available.

B. In many conversations, users communicate in an unstructured and informal way, which makes it difficult for AI systems to understand when they have finished speaking. To handle this, the proposed system includes a turn-taking detection module based on Natural Language Processing (NLP). This module uses a transformer-based model, such as BERT, which is fine-tuned on conversational datasets containing real user interactions. The model is designed to analyze the structure, meaning, and completeness of a user's input. A classification layer is added on top of the model to predict whether the user's turn is complete or if they are likely to continue. After training on thousands of conversational samples, the model achieved high accuracy in identifying turn boundaries and performed well across different communication styles One of the key strengths of this approach is its ability to understand subtle differences in language. For example, it can distinguish between a complete question and an unfinished thought, even when the sentences are short or informal. However, in some cases, users may provide vague or incomplete inputs, which can make it challenging for the model to make accurate decisions. To

improve performance, the system incorporates additional contextual information, such as punctuation, keywords, and conversational patterns. Important signals like question words, sentence endings, or hesitation markers are identified and used to refine predictions. Overall, the NLP-based turn-taking module plays a crucial role in making conversations more natural, responsive, and aligned with human communication patterns.

C. Chat-Based User Interface for Real-Time Interaction

The proposed system includes a simple and user-friendly chat-based interface that allows users to interact with the AI in a natural and intuitive way, even without technical knowledge. The interface supports both text and voice input, enabling real-time communication while also capturing important interaction cues such as typing patterns, pauses, and response timing. These inputs are processed by the backend modules, where NLP and timing analysis work together to determine the appropriate moment for the AI to respond. The conversation is displayed in a clear dialogue format, showing both user messages and system replies. By continuously monitoring user behavior, the system ensures smooth turn-taking, avoiding interruptions and unnecessary delays, which makes the interaction more responsive, engaging, and human-like.

v. IMPLEMENTATION

The implementation of the proposed turn-taking system focuses on simplicity, scalability, and the use of open-source tools to ensure easy development and reproducibility. The overall architecture is divided into two main components: the front-end and the back-end. The front-end is responsible for providing a chat-based interface where users can interact with the system in real time, while the back-end handles all the processing, including NLP analysis, timing detection, and response generation. These two components communicate seamlessly through API calls, enabling smooth data exchange and real-time interaction.

A. Tools and Framework:

models for the proposed turn-taking system were trained on a workstation equipped with a high-performance GPU to speed up the learning process. The training focused on teaching the system how to accurately detect when a user has completed their turn in a conversation. For the NLP-based module, a pre-trained BERT model was fine-tuned using conversational datasets with a batch size of 32 and a learning rate of 2e-5. The training process was completed in a few hours over multiple epochs, during which performance metrics such as accuracy and F1-score were continuously monitored to ensure effective learning. Since some conversational patterns (like rare interaction behaviors) had fewer samples, techniques such as data balancing and augmentation were applied to improve model performance and generalization. Additionally, simulated variations like different pause lengths and typing speeds were introduced to make the model more robust in real-world scenarios.

B. Model Training:

All model training was performed on a workstation equipped with a NVIDIA RTX 3080 GPU, which helped to ac celerate deep learning computations. The image classification model was trained using a two-phase approach: first freezing the base layers and then fine-tuning the model. Training was conducted for 50 epochs on the combined dataset. The entire training process took approximately 4 to 5 hours and the model converged around 95For the text-based diagnosis module, the BERT model was fine-tuned using a batch size of 32 and a learning rate of 2e-5. Training was carried out on a dataset of 5,000 symptom descriptions and took about an hour for three epochs. During training, we monitored several performance metrics, including accuracy and F1-score for each disease class, to ensure that the models were learning effectively. We also observed that there are classes with fewer training samples, such as plant disease classes, and their recall values were also low in the beginning. To solve this problem,

oversampling methods were applied to text data and data augmentation methods were applied to image data.

C. Integration of Models

After training, the models were saved in serialized formats for easy deployment. The NLP model was stored in a format compatible with transformer frameworks, while other supporting modules were implemented as lightweight Python components. These models are loaded by the backend server when the system starts. A Flask-based server is used to handle communication between the front-end and back-end. When a user interacts with the system, the input first goes through preprocessing—text is tokenized and structured, while timing and behavioral cues are extracted. The processed data is then passed to the turn-taking detection model, which predicts whether the user's turn is complete along with a confidence score. This output is forwarded to the dialogue manager, which decides when to generate and deliver a response. The dialogue manager is designed to maintain conversation flow and may use advanced language models to generate more natural and context-aware replies.

D. Front-End Implementation

The front-end of the system is designed as an interactive chat interface that allows users to communicate easily with the AI. It includes a main chat window where both user messages and system responses are displayed in a conversational format. The interface captures user input in real time and sends it to the backend for processing. It also tracks interaction signals such as typing activity and pauses, which are important for turn-taking decisions. The message from the bot is updated on the interface. The system also supports flexible interaction, allowing users to continue typing or speaking without interruption, while the AI intelligently waits for the right moment to respond. This design ensures a smooth, responsive, and human-like conversational experience.

VI. EVALUATION AND RESULT

We evaluated the proposed turn-taking system from both technical performance and user experience perspectives to understand how effectively it improves conversational flow. The evaluation focused on three main aspects: (1) accuracy of detecting whether a user has completed their turn based on text and interaction cues, (2) quality and naturalness of the system's responses, and (3) system efficiency in terms of response time. This helped us assess not only how well the system performs technically, but also how comfortable and engaging it feels for users during real-time interactions.

A. Turn-Taking Detection Performance:

To evaluate the model's performance, we used a test set of thousands of conversational samples that were not seen during training. The transformer-based model achieved an overall accuracy of around 94–96% in correctly identifying whether a user had finished their turn.

B. Text-Based Interaction Understanding Performance:

The NLP-based module was further tested on a separate dataset containing real-world chat interactions, including informal and incomplete sentences. It achieved approximately 85–88% accuracy in predicting turn completion based on textual input alone. The most common errors occurred in cases where user input was very vague, extremely short (e.g., "hmm", "okay"), or abruptly stopped mid-sentence. However, when the input included clear intent, proper sentence endings, or meaningful context, the model performed reliably. Overall, the results show that combining linguistic understanding with timing cues significantly improves the system's ability to manage turn-taking and deliver more natural, responsive interactions.

TABLE III
EXAMPLE INTERACTION SCENARIOS WITH TURN-TAKING SYSTEM

Scenario (User Input)	System Output (Turn-Taking Decision & Response)
Case 1: User types: "I wanted to ask about…" and pauses for a few seconds	Decision: User may continue → System waits before responding.
Case 2: User sends multiple quick messages: "Hi" → "I need help" → "with chatbot timing"	Decision: Ongoing input → System delays response until user finishes.

Table 3 presents example scenarios that demonstrate how the system processes user input, predicts whether the user has finished their turn, and responds accordingly. The results show that the system is able to maintain a smooth conversational flow by avoiding interruptions and providing timely, context-aware responses, making the interaction more natural and user-friendly.

VII. DISCUSSION

The development and evaluation of the proposed turn-taking system highlight several strengths as well as some practical limitations that need attention.

A. Strengths and Contributions:

One of the main strengths of the system is its ability to respond in real time while maintaining a natural flow of conversation. Unlike traditional chatbots that often interrupt or delay responses, this system intelligently decides when to reply, making interactions smoother and more comfortable for users. It reduces awkward pauses and avoids cutting users off mid-sentence, which significantly improves the overall experience. Another advantage is its flexibility—the system can work across both text and voice-based platforms, making it suitable for a wide range of applications such as virtual assistants, customer support, and educational tools. Additionally, once deployed, the system can handle a large number of interactions with minimal cost, making it scalable and efficient.

B. Social Impact:

Improving turn-taking in AI systems can have a meaningful impact on how people interact with technology in their daily lives. More natural conversations can make AI tools easier to use, especially for people who are not very comfortable with technology. In areas like education, better interaction can help students engage more effectively with learning platforms. In customer service, it can reduce frustration and improve satisfaction. As AI becomes more common in everyday communication, making it feel more human-like can help build trust and encourage wider adoption.

C. Ethical Consideration:

While building such systems, it is important to consider user privacy and data handling. Since conversational systems may process personal messages or voice inputs, ensuring that user data is not stored or misused is essential. Another concern is bias in training data, as conversational datasets may not represent all types of users equally. This can affect how well the system performs for different speaking styles, languages, or communication patterns.

VIII. CONCLUSION AND FUTURE WORK

In this work, we presented a system that focuses on improving turn-taking in human–AI conversations to make interactions more natural and responsive. By combining language understanding with timing and behavioral cues, the system is able to decide when to respond in a way that feels closer to human conversation. The results show that improving turn-taking can greatly enhance user experience, making AI systems more engaging, efficient, and easier to

use.

A. Key Takeaways:

One of the key insights from this work is that conversation is not just about understanding words, but also about understanding *when* to speak. By considering pauses, context, and user behavior, the system can create smoother interactions. Another important takeaway is that combining different types of information—such as text and timing—leads to better performance compared to relying on a single input source. It was also observed that users prefer conversational systems that feel natural and do not interrupt, highlighting the importance of human-centered design in AI development.

B. Future Work:

- Voice and Multilingual Support: Future improvements can include support for multiple languages and better voice-based interaction, allowing users from different backgrounds to communicate more easily.
- Emotion and Sentiment Awareness: Adding the ability to detect user emotions can help the system adjust its response timing and tone, making interactions more empathetic.

Personalization and Learning: The system can be enhanced to learn from user behavior over time, adapting to individual communication styles for better interaction.

- Integration with Real-World Applications: The system can be applied to domains such as virtual meetings, healthcare assistance, and smart devices, where effective communication is critical.
- Real-Time Adaptation: Future systems can dynamically adjust response timing based on user behavior during the conversation.
- Handling Complex Conversations: Improvements can be made to handle group conversations or multi-user interactions where turn-taking becomes more complex.
- Cross-Platform Integration: The system can be integrated into messaging apps, smart devices, and enterprise communication tools for wider usage.
- Improved Context Awareness: Future models can better use long conversation history to make more accurate turn-taking decisions.
- Low-Resource Optimization: Developing lightweight models for faster performance on mobile and edge devices can make the system more accessible.

In conclusion, this work highlights how improving turn-taking can make human–AI conversations feel more natural and comfortable. By focusing on when the system should respond, rather than just what it should say, we can create interactions that are smoother and closer to real human communication. This approach has the potential to improve the usability of many everyday AI systems, such as chatbots, virtual assistants, and support tools. With further development and real-world testing, the system can be refined to handle different communication styles and situations more effectively. The insights gained from this work can also contribute to the broader goal of building AI systems that communicate in a more human-like and meaningful way. Over time, such improvements can play an important role in making AI interactions more intuitive, reliable, and widely accepted in daily life.

REFERENCES

[1] G. Skantze, Turn-taking in Conversational Systems and Human-Robot Interaction, Computer Speech & Language, 2017.

[2] A. Raux and M. Eskenazi, A Finite-State Turn-Taking

Model for Spoken Dialog Systems, Proceedings of NAACL HLT, 2009.

[3] M. Roddy, G. Skantze, and S. Harte, Predicting Turn-Taking in Conversational

[4] E. Ferragut and J. R. Stewart, Modeling Human Turn-Taking Behavior for Conversational Agents, IEEE Spoken Language Technology Workshop, 2018.

[5] WildASR Research Group, Challenges in Real-World Automatic Speech Recognition Systems, 2026.

[6] A. Kumar et al., Voice Activity Detection in Noisy Environments: A Survey, 2026.

[7] Picovoice Inc., Speech Recognition and Voice AI Technologies Overview,2026.

[8] S. Sacks, E. Schegloff, and G. Jefferson, *A Simplest Systematics for the Organization of Turn-Taking for Conversation*, Language, 1974.

[9] D. Schlangen, *From Reaction to Prediction: Experiments with Computational Models of Turn-Taking*, Interspeech, 2006.

[10] G. Skantze, *Turn-Taking in Conversational Systems and Human-Robot Interaction: A Review*, Computer Speech & Language, 2021.

[11] J. Allen et al., *Mixed-Initiative Interaction*, IEEE Intelligent Systems, 1999.

[12] T. Baumann and D. Schlangen, *The InproTK 2012 Release*, Proceedings of the NAACL-HLT, 2012.

[13] J. Devlin et al., *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*, NAACL, 2019

.

[14] A. Vaswani et al., *Attention Is All You Need*, NeurIPS, 2017.

[15] R. Skantze, *Towards a General, Continuous Model of Turn-Taking in Spoken Dialogue Using LSTM Recurrent Neural Networks*, SIGDIAL, 2017.

[16] H. H. Clark, *Using Language*, Cambridge University Press, 1996.

[17] A. Raux and M. Eskenazi, *Optimizing Endpointing Thresholds Using Dialogue Features in Spoken Dialogue Systems*, SIGDIAL, 2009.