GEORGE BROWN PRESS
JOURNALS, BOOKS & PROCEEDINGS

International Journal of Engineering Advances

Article

Research on Real-Time User Feedback Acceleration Mechanism Based on Genai Chatbot

Xiao Liu [1,*]

[1] Meta Monetization, Meta, Bellevue, Washington, 98004, USA
* Correspondence: Xiao Liu, Meta Monetization, Meta, Bellevue, Washington, 98004, USA

Abstract: As GenAI adoption continues to surge, AI chatbots have become one of the most prominent and widely deployed applications of artificial intelligence, especially across sectors such as customer service, education, healthcare assistance, and content generation. In these domains, the real-time feedback loop generated through user interactions-and the system's ability to adapt based on this feedback-plays a critical role in determining the quality, reliability, and intelligence level of chatbot responses. As user expectations for accuracy, immediacy, and personalization continue to rise, the demand for a more robust and efficient feedback mechanism has become increasingly urgent. However, existing GenAI chatbot systems often struggle to maintain sufficient speed and stability in their real-time feedback loops. Problems such as delayed feedback processing, insufficiently refined semantic interpretation, and occasional inaccuracies in generated responses weaken the system's overall performance and may negatively affect user experience. These challenges highlight the necessity of integrating real-time feedback more directly and effectively into the core model processing pipeline. This article aims to address the above limitations by investigating how real-time user feedback can be seamlessly incorporated into AI model interactions to enhance response efficiency and accuracy. The research proposes an integrated optimization strategy that includes streamlining data transmission paths, improving semantic parsing algorithms, refining intent recognition processes, and applying targeted model fine-tuning methods. These enhancements work together to accelerate the system's feedback processing capabilities and strengthen its adaptive response mechanism. Through extensive experimental testing and comparative analysis, the study demonstrates that the optimized system significantly improves the efficiency of the feedback loop, enhances the coherence and precision of generated responses, and ultimately boosts user satisfaction. The findings suggest that the integration of real-time feedback into GenAI model processing offers a viable pathway for building more intelligent, adaptive, and user-centered chatbot systems suitable for broader practical deployment.

Keywords: chatbot; user feedback; real time mechanism; model optimization

Received: 06 November 2025
Revised: 11 November 2025
Accepted: 02 December 2025
Published: 07 December 2025

1. Introduction

As artificial intelligence continues to evolve rapidly, it has been extensively integrated into a wide range of medium- to large-scale applications, with AI chatbot systems powered by advanced large language models standing out as one of the most visible and mature implementations. These systems rely heavily on continuous user interaction, and the feedback generated throughout the dialogue plays a crucial role in guiding model optimization, improving response strategies, and enhancing overall interaction quality. Real-time feedback has therefore become a fundamental component for enabling adaptive learning and maintaining high-quality conversational performance.

Despite this importance, the current mechanisms for processing and utilizing user feedback still face multiple technical challenges [1]. First, feedback collection channels are often inefficient or fragmented, making it difficult to capture user behavior signals in a timely and complete manner. Such delays hinder the system's ability to update dialogue

states dynamically and limit its capability to adjust responses based on the latest user intent. Second, user feedback is frequently expressed in non-standard language forms, such as colloquial phrases, implicit meanings, emotional expressions, or incomplete linguistic structures. These characteristics significantly increase the cognitive burden on the system, complicating semantic interpretation, intention extraction, and sentiment understanding [2].

Furthermore, the relatively slow pace of model updates introduces another bottleneck. In many existing systems, user feedback remains primarily at the data aggregation layer, waiting to be processed during periodic retraining cycles. The insights from this feedback are not rapidly incorporated into the real-time generation logic, resulting in response patterns that lag behind evolving user preferences. When interactions occur frequently or involve complex tasks, this gap becomes even more pronounced, weakening the system's ability to adapt promptly and reducing its perceived intelligence and utility in demanding environments [3].

Together, these limitations underscore the need for more efficient mechanisms that integrate real-time feedback into model processing pipelines. Addressing this challenge is essential for improving the responsiveness, adaptability, and stability of GenAI chatbot systems, particularly in scenarios that require long-duration interactions and context-sensitive decision-making.

2. Overview of Genai Chatbots

GenAI chatbot is a type of natural language processing system built on generative artificial intelligence technology. GenAI mainly achieves system goals through a large number of trained language models. The system expresses the user's intentions based on models in a large corpus, and automatically generates contextually coherent, semantically reasonable, and contextually appropriate natural language text. Compared to traditional conversational robots, GenAI has strong language comprehension ability, achieving more free language generation effects and deeper contextual grasp ability. Its application scope covers intelligent customer service, chatbots, question and answer services, writing generation, and educational assistants, becoming an important force driving the iterative evolution of artificial intelligence interaction systems [4].

Usually, the system includes a decoding input unit, a syntax generation module, an environment adjustment module, and a response strategy. During the conversation, the GenAI model outputs smoothly and accurately, while continuously integrating user behavior feedback to enhance the relevance and personalization of the content. With the increasing demand for real-time communication, intelligent systems are expected to have faster problem response times and stronger environmental adaptability, thus placing higher demands on generating rules and providing timely feedback. In order to meet these requirements, a more efficient feedback path should be constructed to enhance the learnability and adjustment ability of the model, so that GenAI has stronger flexibility to adapt to complex bidirectional interaction and intelligent decision-making collaborative tasks, and has more application scenarios [5].

3. Genai Chatbot Application

3.1. Practical Scenarios for Online Customer Service

In online customer service scenarios, GenAI chatbots, with their powerful language generation capabilities, have gradually replaced traditional online customer service and become an important tool for improving user satisfaction and service efficiency. The system can perform semantic understanding based on the questions raised by users, dynamically generate targeted answers in combination with context, and thus achieve a 24/7, high concurrency automated response mechanism. Compared to traditional manual customer service, it has a faster response speed, wider information coverage, and the ability to continuously learn and expand content, effectively alleviating the pressure of

tight human resources and rising service costs. In practical applications, GenAI systems are usually integrated into enterprise customer service platforms, responsible for tasks such as preliminary inquiries, answering common questions, and providing process guidance. By collecting and processing real-time user feedback, the system can also perform clustering analysis and knowledge base updates on high-frequency issues, improving response accuracy and semantic matching effectiveness in subsequent interactions. When evaluating the service effectiveness of such systems, the following comprehensive indicator model can be introduced:

$$E = \alpha R + \beta A + \Gamma s \quad (1)$$

Among them, E refers to the overall service efficiency, R to evaluate the response speed, A For the accuracy of the answer, SFor user satisfaction,α, β, γThe empirical weight coefficient. This model helps to systematically evaluate the performance of GenAI in online customer service scenarios and provides data basis for subsequent optimization.

3.2. Educational Q&A Assistance Function

In educational application scenarios, GenAI chatbots have been used for tasks such as knowledge explanation, homework guidance, and concept restatement, making learning more timely and personalized. It can answer questions raised by students, rather than relying solely on mechanical or formulaic answers. For problems with strong repetition and clear knowledge boundaries, such as formula derivation, definition analysis, and general exercise exercises, GenAI can provide accurate answers in an instant, helping teachers reduce their burden and allowing more time for student learning. These tools may become important tools in online learning platforms, virtual classrooms, and personalized learning systems. The effectiveness of such teaching question and answer systems can be determined by comprehensively evaluating three dimensions: knowledge matching, language expression, and student satisfaction. The following model can be constructed for representation:

$$S = \theta_1 M + \theta_2 L + \theta_3 R \quad (2)$$

Among them, STo represent the overall teaching assistance score, MFor the matching degree of knowledge points, LFor clarity in language expression, RScoring student satisfaction, $\theta_1, \theta_2, \theta_3$ For weight coefficients. This formula helps evaluate the comprehensive service capability of GenAI in educational Q&A tasks and provides a basis for subsequent strategy adjustments.

3.3. Content Creation and Generation Tools

In the content creation process, ChatGenAI is widely used in copywriting, article expansion, conversation simulation, and story generation. Based on its powerful writing ability, it can automatically conceive and generate structured and logically coherent articles based on user input keywords, themes, or genres. To meet the innovative writing needs in various environments, multiple forms and styles of articles can be generated in a short period of time, greatly improving the efficiency and diversity of content output, such as in industries such as social content management, advertising copy creation, and teaching material production. To quantify the generation quality of the system in content creation, a language generation quality rating model can be introduced:

$$G = \mu_1 C + \mu_2 D + \mu_3 F \quad (3)$$

Among them, G referts to generate an overall score for the content, C as a measure of content integrity, D for the level of semantic diversity, FTo score the conformity of the format,μ_1, μ_2, μ_3To adjust the weight coefficients. This model can effectively measure the performance of generated results under different tasks, providing quantitative support for system optimization and style control.

4. Problems with Real-Time User Feedback Acceleration Mechanism

4.1. Delay in Feedback Collection Process

The effectiveness of collecting user feedback during GenAI chatbot interaction directly affects the system's response speed and real-time performance. The existing system architecture has obvious shortcomings in front-end interface design and data channel construction, and user feedback cannot be immediately obtained and forwarded to the processing module. Some platforms submit feedback in the form of polling mechanisms or time triggers, resulting in delayed feedback data due to the inability to continuously monitor. Due to factors such as network conditions and service unavailability, delays can also be exacerbated, resulting in the inability to maintain the timeliness of feedback content during conversation rounds. For implicit actions such as quick change of operation, multiple clicks, pause of operation, etc., and the lack of a unified method system and structured collection mechanism, a large amount of implicit feedback has not been collected, which also affects the generation and improvement of backend models. The incomplete collection capability not only affects the overall perception of user behavior by the system, but also may lead to the omission of feedback data during the generation process, further weakening the model's ability to learn high-frequency interaction patterns and limiting the flexible response performance of intelligent systems in complex scenarios.

4.2. Feedback Content Semantic Ambiguity

In dialogue system interaction, user feedback often presents vague and unstructured expression features, lacking clear semantic direction. Many feedbacks are expressed in extremely simple or vague language, such as "no", "problematic", "say it again", making it difficult for the system to determine the specific content of their negation or questioning. The feedback from multiple conversations may lose consistency with past information, making it difficult for the model to construct semantic connections. Some feedback contains emotional statements, such as language or subjective opinions, which are difficult to analyze semantically. In addition, different users have significant differences in language styles and expression habits, and the same meaning can take on different forms. The above issues not only interfere with the consistency and adaptability of feedback recognition, but also easily lead to misjudgments in semantic understanding. Interfering with the correct reflection of user intent by the interference system reduces the reference value of feedback in subsequent strategy modifications.

4.3. Slow Feedback Update Cycle

For GenAI chatbots, responding to user feedback often relies on periodic training to adjust model parameters, making it difficult to achieve timely response. At present, the vast majority of systems adopt centralized offline fine-tuning, from collecting, filtering, and cleaning user feedback to incorporating feedback data into the training pipeline. The delay generated by this process reduces the system's self adaptability. In the actual interactive application process, users expect their feedback to immediately respond to the content. However, due to the current system structure being unable to meet this requirement, there is a phenomenon of model behavior being disconnected from user intent. Although online learning has made some progress, model stability issues may still arise during the process of model optimization. Therefore, there is a lack of opportunities for frequent updates, and due to the difficulty in effectively converting feedback data into regulatory signals in the generation logic, it is easy for the system to have a dead loop, which reduces the quality of dialogue and user satisfaction. Therefore, slow updates have become a major obstacle to the development of models and the improvement of feedback loop efficiency (As shown in Figure 1).

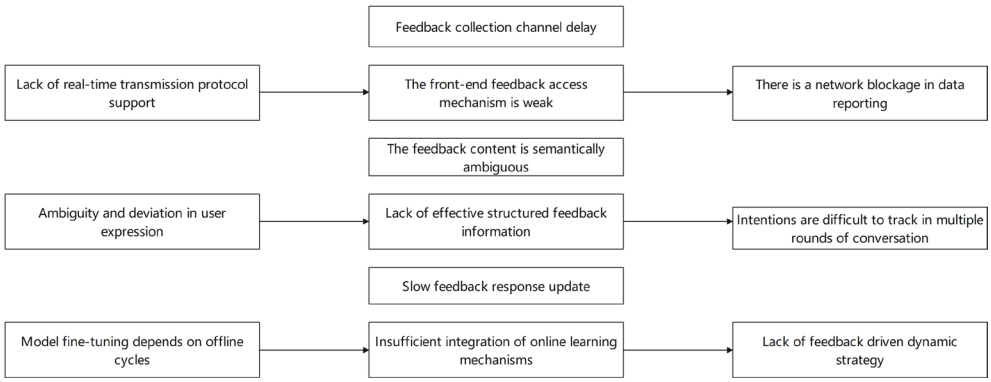

Figure 1. Issues with the Real time User Feedback Acceleration Mechanism.

5. Optimization Strategy for Real-Time User Feedback Acceleration Mechanism

5.1. Optimizing the Data Transmission Link

In order to improve the feedback response efficiency of GenAI chatbots during user interaction, enhance the system hierarchy, and make the data transmission path more efficient, otherwise the problems of a large amount of feedback data accumulation and feedback information processing delay cannot meet real-time performance requirements. Building a low latency, high bandwidth feedback communication channel is one of the key elements to achieve efficient interactive closed-loop. Use event driven architecture and long connection technologies (such as WebSocket, gRPC, etc.) to reduce the latency of front-end submission of feedback data to back-end processing. By configuring feedback event detectors on the user side, utilizing small batch buffering, and edge device forwarding mechanisms, feedback data can be quickly collected, processed, and organized, improving data flow in the system. The difference in transmission modes directly affects the time delay of feedback processing, and different methods have their own characteristics in delay control, data integrity, and system load. To compare the feedback delay performance of different transmission modes, refer to Table 1:

Table 1. Performance Comparison of Different Feedback Transmission Methods.

transmission mode	average delay (ms)	Packet loss rate (%)	Real time evaluation
HTTPPo lling	420	2.3	poor
WebSocketDirect connection	75	0.8	excellent
gRPCasynchronous flow	92	1.1	good
Kafkamessage channel	140	1.7	general

Through collaborative optimization of network communication and device layout, enhance feedback response efficiency and the system's ability to support high line loads. During the process of establishing a feedback path, the server needs to deploy a distributed receiving module to separately store high traffic and efficient feedback information data. Introducing an event priority queue to queue and allocate strategies for the importance of feedback information, prioritizing responses to feedback information that is clearly directed or highly relevant to the conversation content, and reducing the consumption of system resources by useless data; Establishing feedback status tracking and visualization modules is beneficial for transparent control of the data lines throughout the entire reaction process, improving overall response consistency, and providing basic support for real-time intelligent optimization in complex dialogue tasks.

5.2. Strengthen Semantic Understanding Ability

In real-time feedback mechanisms, improving semantic understanding ability is a prerequisite for accurate response. The dialogue system should have the ability to identify and extract the core essence of feedback, including simple, general, and scattered information. By utilizing context aware mechanisms and semantic enhancement techniques, it is possible to effectively locate the opposing, questioning, and supplementary information present in feedback. By establishing a semantic encoding structure based on BERT and Roberta's deep language model, it is possible to further optimize discourse understanding and combine dialogue state tracking strategies to restore the content targeted by feedback. At the same time, in situations where the context is unclear and the expression is inappropriate, situational construction and matching techniques can further optimize the system's effectiveness in reproducing feedback intentions. To evaluate the adaptation effect of different semantic understanding methods in user feedback processing, the following Table 2 can be established.

Table 2. Performance Comparison of Different Semantic Recognition Methods.

Method Type	Semantic recognition accuracy (%)	Context correlation ability	Real time processing performance
Template keyword extraction method	65	weak	strong
LSTM + Intent Classification Model	78	centre	centre
BERT embedding matching mechanism	90	strong	centre
RoBERTa Context Fusion Structure	93	strong	stronger

In multi round dialogue scenarios, using a dialogue level attention framework can help the model better extract conversational intentions and vague expressions, such as "can still do this", "is a bit wrong", and "put it another way". This can be determined uniformly through language similarity and emotion assisted annotation, and combined with semantic hierarchical classification label generation, to complete the transcoding of reply structures, achieving accuracy improvement while obtaining more reliable data sources for subsequent model training. Building a cross task universal semantic feature library can help improve the system's adaptability to new responses.

5.3. Building a Rapid Update Model

In the real-time feedback processing system, constructing an efficient model update process is the core link in transforming user feedback into system improvement behavior. The current system updates centered around large-scale models rely heavily on offline training methods. This model has a slow update frequency and slow response, which is not cost-effective for timely response to user needs. On more frequent and low latency model updates, an incremental learning strategy combining parameter freezing and local fine-tuning can be adopted. By introducing lightweight fine-tuning structures such as LoRA (Low Rank Adaptation), Adapter, etc., only a small number of intermediate layer parameters are updated, which can ensure the stability of the model backbone and quickly adjust some response behaviors. Suitable for application scenarios with limited device resources and capable of receiving real-time generated data streams. To compare the performance of different update strategies in terms of timeliness and resource utilization, the following comparison Table 3 can be established.

Table 3. Performance Comparison of Different Model Update Strategies.

Update method	Proportion of parameter quantity (%)	Update duration (seconds), accuracy change (Δ%), application flexibility	Update duration (seconds), accuracy change (Δ%), application flexibility	Update duration (seconds), accuracy change (Δ%), application flexibility
Total fine-tuning	100	1800	+6.5	low
Adapter fine-tuning	8	420	+5.2	tall
LoRA fine-tuning	6	310	+5.4	tall
Prompt Tuning	2	250	+3.8	centre

By selecting the appropriate fine-tuning strategy and adjusting the processing mode of local updates based on task types and feedback frequency. After dividing the model parts according to the types of tasks, the main model is maintained to run stably in low-frequency tasks, while the main model is kept running normally for infrequent tasks. Thus achieving an effective balance between cost control and performance improvement. In addition, a feedback selection mode has been introduced to further optimize the update speed and improve the relevance of the generated results.

6. Conclusion

The widespread application of AI chatbots in intelligent interaction scenarios is driven by user feedback behavior, which is the main driving force for improving system performance. This article proposes a series of optimization measures to improve feedback efficiency, including semantic enhancement and lightweight online updates, to address issues such as feedback information acquisition timeliness, accuracy in language meaning judgment, and time-consuming responses. This forms a fully functional and closed-loop feedback acceleration framework, which is proven through practical comparison and system reconstruction. It improves the overall feedback efficiency and model response flexibility, and is helpful for the system in complex interaction requirements and coverage. Not only does it improve the overall operational efficiency, but it also provides a powerful driving force for the further development of model intelligence. For future development, this closed-loop feedback may further promote human-machine collaboration, intelligent decision-making, and customization, and may become one of the key foundational technologies supporting the continuous advancement of generative AI.

References

1. Z. Jin, J. Wang, D. Li, M. Li, and B. Li, "MSCA: a multi-scale context-aware recommender system leveraging reviews and user interactions," *International Journal of Web Information Systems*, vol. 21, no. 3, pp. 205-229, 2025. doi: 10.1108/ijwis-10-2024-0311
2. C. E. Schilstra, C. E. Wakefield, J. McLoone, L. Wiener, M. W. Donoghoe, R. I. Hoffman, and J. Cayrol, "Development of a World Health Organization international survey assessing the lived experience of people affected by cancer: outcomes from pilot testing, user feedback, and survey revision," *Supportive Care in Cancer*, vol. 33, no. 5, p. 445, 2025. doi: 10.1007/s00520-025-09372-2
3. P. K. Das, and T. Kumar, "Ecommerce sellers' ratings: Is user feedback adequate?," *International Journal of Consumer Studies*, vol. 47, no. 4, pp. 1561-1578, 2023. doi: 10.1111/ijcs.12938
4. A. Almutairi, M. Al-Amri, and K. Button, "A preliminary evaluation of a virtual reality-based physiotherapy prototype toolkit: a usability study, content analysis of user feedback," *Osteoarthritis and Cartilage*, vol. 32, pp. S565-S566, 2024. doi: 10.1016/j.joca.2024.02.836
5. C. Wall, and R. Laing, "Achieving Net Zero: How Could User Feedback Be Leveraged to Promote Domestic Heat Pump Adoption in Scotland?," *Sustainability*, vol. 16, no. 17, p. 7833, 2024. doi: 10.3390/su16177833