# Communicating for Trustworthiness

Opportunities for Chat Improvement in Fintech

Vivian Jiang
Lead Conversation Designer
Rocket Mortgage
Detroit, Michigan, United States
vivianjiang@rocketmortgage.com

Maria Karamihaylova
Conversation Designer
Rocket Mortgage
Detroit, Michigan, United States
mariakaramihaylova@rocketmortgage.com

## ABSTRACT

LLM-driven generative chatbots are gaining increased popularity across business applications. We outline four areas of opportunity to improve upon a fintech AI chat agent in the mortgage industry based on academic investigation around the application of Theory of Mind (ToM) to conversational agents, with a particular focus on trust. We propose methods to refine the chat experience rooted in research to increase the perception of trustworthiness, thereby driving user adoption, which includes: improving conversational breakdown recognition, limiting the possibility of conversational fatigue, augmenting intent-detection and task-resolution accuracy, and ameliorating the response time perception via visual- and text-based justifications.

## CCS CONCEPTS

• **Human-centered computing → Natural language interfaces**

## KEYWORDS

Theory of Mind, conversational agents, chatbot response delay, trust, transparency, conversational fatigue

**ACM Reference format:**

Vivian Jiang and Maria Karamihaylova. 2025. Communicating for Trustworthiness: Opportunities for Chat Improvement in Fintech. In *ACM Conversational User Interfaces 2025 (CUI '25). ACM, Waterloo, ON, Canada, 5 pages.*

## 1 INTRODUCTION

Financial technology (fintech) companies have increasingly been deploying generative AI-enabled solutions for tasks including fraud detection, customer service operations, automated data insights, and transactional data analysis [2]. AI chat assists in the banking and finance worlds are generally considered useful for customer service automation, particularly for simple tasks; frustration primarily arises when responses are inaccurate or unreliable [8]. As conversation designers at Rocket Mortgage, an American mortgage lender, we work to mitigate the potential for these stumbling blocks in communication by providing careful attention to the structure of our chat assistants' underlying dialogue flows.

Our team is responsible for scoping, designing, testing, and monitoring conversational flows and metrics for several different LLM-driven chat assistants that serve both our general customer base and Business-to-Business (B2B) partners. The creation and upkeep of our AI chat features are anchored in observations of how real users interact with our chat interface; designs are frequently driven by needs uncovered through professional internal research and work to address issues frequently encountered through industry experience. Establishing trust is key to providing AI-powered chat-based services to our users, particularly in the context of significant financial decisions (such as purchasing or refinancing a home) [12].

In this paper, we highlight opportunities for addressing common obstacles across two of our chat interfaces and provide examples and solutions for each. We seek to improve recognition of conversational breakdown and reduce conversational fatigue for our customer-facing AI, and to also heighten user trust in the system to resolve issues, as well as provide more visibility into response times for our B2B AI.

Consensus on potential solutions for new issues is often achieved through A/B variant testing based on hypotheses developed from the previously mentioned processes. We believe that interpreting the highlights of these pain points through a Theory-of-Mind (ToM) lens helps inform our forthcoming design solutions and UX research roadmaps while remaining focused on fostering user trust and staying rooted in a user-centric approach.

## 2 Consumer-facing AI chat case study

Rocket Mortgage's client-facing chat interface has two primary instances: one that appears after a lead form accessed through paid search ("Post-Form Chat") and another that appears on most public-facing pages that receive organic traffic ("General Chat"). Traffic for these chat instances is chiefly composed of users who are in different stages of the mortgage process, trending towards those who are early in exploring their financing options and evaluating potential lenders. For General Chat, engagement topics vary, ranging from general knowledge mortgage questions to those more specific to a user's individual finances. For Post-Form Chat, users are prompted to provide more details about their finances to better match them with a loan product.

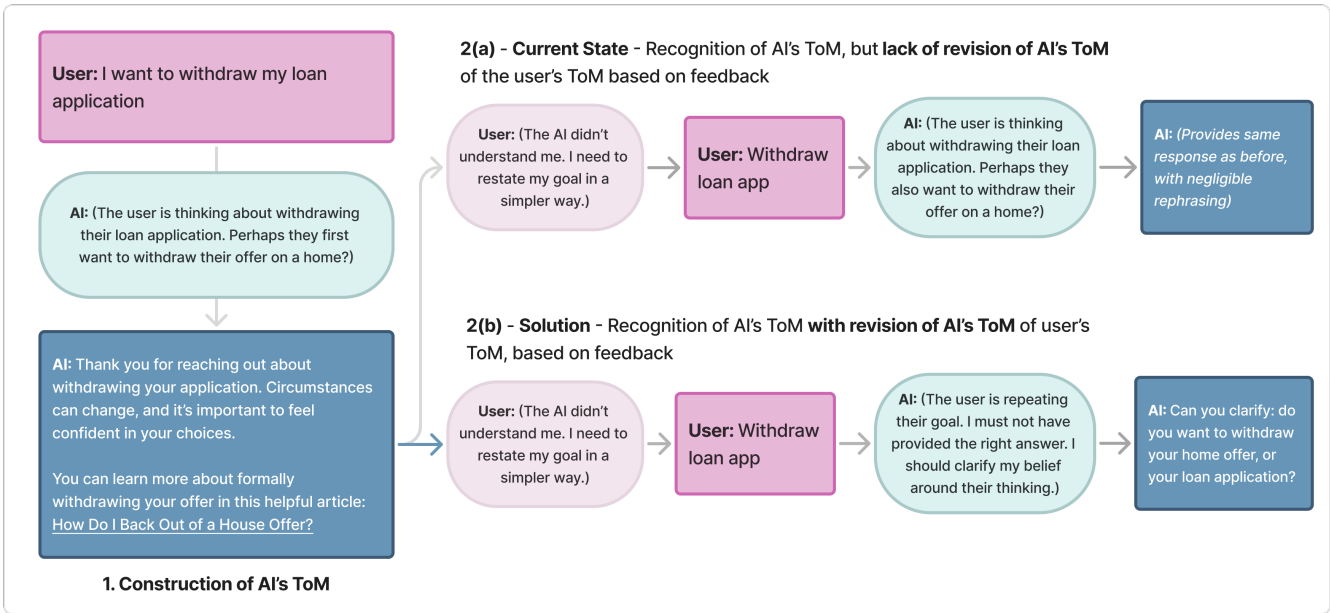


Figure 1: An abstraction of a real dialogue between a human user and Rocket Mortgage's "General Chat." This image highlights the process of MToM between user an AI as detailed in Wang and Goel (2022) [17]. It also offers the potential revised ToM of the AI in regards to the user, as well as the applicable system-repair solution.

In both these instances, the AI attempts to triage utterances and guide the user through different semi-structured pathways based on its evaluation of the topic. It is also able to route users to internal human agents ("experts") with respect to the specific business area of the conversation. Additionally, both instances utilize a combination of several LLMs to drive and manage the dialogue.

## 2.1 Improve recognition of conversational breakdown

In examining a dialogue (pictured above in Figure 1) between a user and General Chat, we see an opportunity for revision in the AI's ToM of the user to better address their goal based on the framework of Mutual Theory of Mind (MToM) by Wang and Goel (2022) [17].

In Figure 1, the user is trying to find assistance in withdrawing an in-process loan application. We can infer from their phrasing ("I want to withdraw my loan application") in combination with elements of Mutual Theory of Mind (MToM) that the user's initial constructed *perception* of the AI is that it can understand the user's specific goal and will help the user fulfill it. In response, the AI intakes this utterance and constructs its own ToM: it concludes the user is feeling concerned about their loan withdrawal and would like general support for their situation. Resultantly, the AI responds with unbiased advice and links to what it thinks is an informational article. Misunderstanding arises when the article shared is not directly relevant to the user's ask, revealing a gulf in understanding; in this moment of conversational breakdown, the user may question the AI's ability and therefore its trustworthiness [5].

The user's *recognition* of the AI's theory of the user's mind leads them to respond in a less embodied manner, switching to a shorter, more commanding statement. We infer the user thinks the AI's answer did not grasp their underlying request; consequently, they repeat their intent more simply, likely indicating a desire to continue the conversation but with a new level of reservation. This *feedback* from the AI leads the user to reshape their understanding of the AI's mindset to one that may perform better with a minor clarification. However, based on the similarities in the user's utterances, the AI does not find a need to revise its ToM and therefore provides the same messaging as earlier, causing a breakdown.

To improve this exchange, the AI should recognize repeated dialogue turns with little variance as a form of conversational breakdown requiring a repair strategy. Despite only a slight difference between utterances, the user's feedback should be used to adjust the AI's ToM instead of maintaining it. In referencing an overview of system-repair strategies, approaching the breakdown with a goal-oriented solution—"solving"—could assist with advancing the conversation towards mutuality [1]. More robust recognition of breakdowns could be achieved through detailed few-shot prompting, which consistently outperforms zero-shot and chain-of-thought methods [20]. Alternatively, a hierarchical system that evaluates potential dialogue disruptions with a subsequent correction by a superior model can also serve to reduce "unsafe" responses [20].

For this scenario, the AI assistant should disambiguate its understanding of the user's intent by asking "Can you clarify: do you want to withdraw your home offer, or your loan application?" Variance in feedback, as opposed to repetition, could augment the user's ToM of the AI by improving the initial belief of confidence. The subsequent answer on the user's behalf, when aligned with the AI's attempt to better understand their request, would support the path to mutuality and an increased belief in the AI's ability, ideally also improving the user's perception of its trustworthiness.

## 2.2 Reduce conversational fatigue

Post-Form Chat currently exists as an opportunity for users to share more detailed financial information in a series of questions asked by AI after previously answering 15 or so questions that were displayed in a "form" format, which required button selections and numerical and text input. Metrics used to evaluate the success of Post-Form Chat are conversational drop-off (with about 10% of users leaving a conversation between questions, on average) and completion rate (around 50%) [16], with the overarching goal of understanding how to balance information collection and value for the user.

We suggest that continued drop-off and low completion rates may be caused by "conversational fatigue"—a form of passive cognitive fatigue resulting from when a user engages in turn-by-turn dialogue with an AI agent for a period of time that exceeds the user's initial expectations—and/or the anticipation of conversational fatigue and potential cognitive load, especially if the end state is unclear, and trust in the AI's ability has begun to erode. Negative past experiences with other AI assistants serve as a reference point for potential frustration with the upcoming task [3]. With such existing biases in mind, the user's formation of the AI's ToM prematurely advances the starting point of conversational fatigue. Additionally, due to the lack of non-verbal feedback cues from the AI, the act of consistently constructing, evaluating, and revising an AI's ToM and adjusting inputs accordingly may likely result in increased mental effort on the user's part. When cognitive resources are strained, "mindblindness" may occur, resulting in more failures to consider and process other's beliefs [11]. This could lead to decreased trustworthiness in the AI chat's competence.

We believe the anticipation of this increased cognitive load—particularly in instances where visibility of system status may be unclear (e.g. the progression of questioning) coupled with the context of providing even more semi-sensitive financial information—could be reduced by using suggested reply buttons [14]. Milana, Costanza, and Fischer (2023) posit that engagement with such buttons "provide[s] a sense that suggestions were generated by the agent itself, which may have demonstrated competence within [the study environment]" [13]. Improving a user's perception of the AI's efficacy and its presumed understanding of the user's ToM—for instance, through offering contextually relevant reply options—has the potential to alleviate elements of conversational fatigue by demonstrating a level of "attention" from the AI assistant, sustaining the presence of trust.

The reply options act as a visual representation of the AI's ToM of the conversation, providing tangible feedback with which the user might shape their perception of the AI as helpfully predictive (even if the options are pre-defined). This heightened perception of competence encourages continued engagement with the AI due to its more seamless integration as an interaction [13], and accordingly, bolster a flow's completion rate by reducing conversational fatigue.

# 3 B2B chat case study

Rocket Mortgage's Business-to-Business chat instance ("B2B Chat") serves as a means for third-party partners and mortgage brokers to submit issue tickets which are resolved by internal team members. This chat experience offers a path for partners to seamlessly identify the correct form required to submit a ticket within ~2-4 conversational turns, in contrast with the previous cumbersome experience, which required partners to navigate through a series of radio button options often involving 6+ mouse clicks. This LLM-powered experience solicits the user's issue type, relevant loan number, and pertinent information required to identify the correct form to surface for use.

## 3.1 Improve trust in system to solve target issues

Our B2B users have anecdotally indicated that they have doubts the conversational system will accurately resolve their target issue, risking lower adoption rates. Here, our definition of "accuracy" comprises of two elements: (a) the correct identification of the indicated issue, and (b) the correct handling of that issue. Følstad, Nordheim, and Bjørkli (2018) found that factors which inspire trust in a user interacting with an AI assistant include the level of the AI's understanding of the user's query paired with the quality of the response [4]. Trust in the AI can be defined by relative measurements including the perception of the AI's competence, reliability, and honesty [18]. Strong levels of user trust in the AI are correlated with user willingness to engage with generative models [18].

Since perceived trust is a precursor to motivation to use an AI-powered chat, it stands to reason that if the end user has a satisfactory experience interacting with the AI, they will be motivated to become a regular user on the basis of perceived trust [10, 18]. We are actively pursuing accuracy improvements by implementing an agentic approach, dividing the knowledge base into chunked categories to encourage improved semantic matching. In addition to improving accuracy ratings, providing a concise explanation of how the chat assistant determined the proposed issue handling may help promote increased trust in our B2B user base. By implementing a mechanism for proving the

system's usefulness through the perception of trustworthiness, we hope to encourage return visits.

### 3.2 Reduce latency and response time

Response time is another salient area of opportunity for our B2B Chat instance. Leveraging generative technology often requires calling Application Programming Interfaces (APIs) and vector knowledge bases, which can result in increased chat response latency that users can perceive to be unnaturally long. A dynamic delay, specially designed to consider how long it takes a human to read a message and write a response, can positively impact a user's perception of both humanness and social presence as a minor response delay is a social cue during human-to-human conversations [6]. Conversely, if a response time is too long, the AI may be perceived as less likable or capable; B2B Chat's user base of third-party mortgage brokers values speedy response times due to the nature of their fast-paced work. Power users who grow more familiar with this AI-powered interface technology may prefer a quicker experience as their mental model is that the interface's role is to efficiently solve a problem rather than to simply hold a conversation [7]. A moderate response, defined as 5-10 seconds, has been shown to best encourage user adoption of a chat interface compared to short (under 5 seconds) or long (greater than 10 seconds) responses times [9]. Presently, our B2B Chat instance experiences 8-12 seconds of response delay for particularly complex query scenarios requiring multiple API calls and knowledge base lookups.

Optimizing response latency is an ongoing goal for our B2B Chat experience. For this user base's mental model of the desired experience, if the chat agent takes too long to respond to a query, the user may perceive the AI to be less trustworthy. As research has shown that perceived transparency is correlated to the overall chat experience, we propose to incorporate more clarity about the process in the interaction [19]. Latency justification may be achieved through visual cues, text-based explanations, or both; additional internal user research will elucidate the direction that will be most impactful to our specific users' perception of subjective transparency and trust. Another technical solution is to explore the use of semantic embedding caching, which has been shown to reduce latency in generative models [15].

## 4 Conclusion

Building and maintaining trust between a user and AI is crucial to achieving a robust, informative dialogue that supports the human user in realizing their goal, particularly in the realm of fintech and financial services.

We highlighted several instances across existing Rocket Mortgage chat interfaces where solutions were formulated through analyzing the specific experiences within a ToM framework: tightening the AI's ability to recognize granular conversational breakdowns and actively disambiguating; enhancing the user's trust in the AI's competence and accuracy through ToM alignment; and providing improved visibility into latency that may arise from the AI composing a response.

This paper has also allowed us to uncover more considerations for the examples outlined above, based on our review of existing research. For instance, what might the latency sweet spot be, and how might it change depending on the AI assistant's conversational system (e.g. goal-oriented vs. open domain)? Would this differ across consumer-facing or B2B audiences in Rocket Mortgage's user base? What about users interacting with AI chat at different phases of the mortgage journey? As industry practitioners of conversation design who are relatively new to the academic ToM framework, we were excited to discover a plethora of opportunities to learn more about its application to conversation design. We hope to explore more of these intersections in the future.

## ACKNOWLEDGMENTS

This paper was completed with the support of Rocket Mortgage and the internal Design and Product organization. We thank our fellow team members and collaborators—especially the wonderful researchers, product designers, content strategists, engineers, and product managers who made these experiences possible — *for* their insights and teamwork. We also extend our thanks to the reviewers for their time and feedback.

## REFERENCES

[1] Essam Alghamdi, Martin Halvey, and Emma Nicol. 2024. System and User Strategies to Repair Conversational Breakdowns of Spoken Dialogue Systems: A Scoping Review. In *ACM Conversational User Interfaces 2024 (CUI '24), July 08–10, 2024, Luxembourg, Luxembourg.* ACM, New York, NY, USA, 13 pages. https://doi.org/10.1145/3640794.3665558

[2] Kalpesh Barde and Parth Atul Kulkarni. 2024. Applications of Generative AI in Fintech. In *Proceedings of the Third International Conference on AI-ML Systems (AIMLSystems '23).* Association for Computing Machinery, New York, NY, USA, Article 37, 1–5. https://doi.org/10.1145/3639856.3639893

[3] Andrea Chin. 2025. Rocket Assist dynamic prompts for rocket.com

[4] Asbjørn Følstad, Cecilie Bertinussen Nordheim, and Cato Alexander Bjørkli. 2018. What Makes Users Trust a Chatbot for Customer Service? An Exploratory Interview Study. In *International Conference on Internet Science, ICIS 2018.* https://doi.org/10.1007/978-3-030-01437-7_16

[5] Asbjørn Følstad, Effie L.-C. Law, and Nena van As. 2024. Conversational Breakdown in a Customer Service Chatbot: Impact of Task Order and Criticality on User Trust and Emotion. ACM Trans. Comput.-Hum. Interact. 31, 5, Article 66 (October 2024), 52 pages. https://doi.org/10.1145/3690383

[6] Ulrich Gnewuch, Stefan Morana, Marc TP Adam, and Alexander Maedche. 2018. Faster Is Not Always Better: Understanding the Effect of Dynamic Response Delays in Human-Chatbot Interaction. In *26th European Conference on Information Systems: Beyond Digitization-Facets of Socio-Technical Change, ECIS 2018, Portsmouth, UK, June 23-28, 2018. Ed.: U. Frank.* 143975.

[7] Ulrich Gnewuch, Stefan Morana, Marc TP Adam, and Alexander Maedche. 2022. Opposing Effects of Response Time in Human–Chatbot Interaction: The Moderating Role of Prior Experience. *Business & Information Systems Engineering* 64, 6 (2022), 773–791. https://doi.org/10.1007/s12599-022-00755-x

[8] Gary Graham, Tahir M. Nisar, Guru Prabhakar, Royston Meriton, Sadia Malik. 2025. Chatbots in customer service within banking and finance: Do chatbots herald the start of an AI revolution in the corporate world?. *Computers in Human Behavior,* Volume 165, 2025, 108570, ISSN 0747-5632. https://doi.org/10.1016/j.chb.2025.108570

[9] Kaeun Kim, Ghazal Shams & Kawon (Kathy) Kim. 2025. From Seconds to Sentiments: Differential Effects of Chatbot Response Latency on Customer Evaluations, *International Journal of Human–Computer Interaction*, 17 pages. https://doi.org/10.1080/10447318.2025.2508915

[10] Fanny Lalot and Anne-Marie Bertram. 2024. When the Bot Walks the Talk: Investigating the Foundations of Trust in an Artificial Intelligence (AI) Chatbot.

In *Journal of Experimental Psychology*, 154, 2, 533–551. https://doi.org/10.1037/xge0001696

[11] Shuhong Lin, Boaz Keysar, Nicholas Epley. Reflexively mindblind: Using theory of mind to interpret behavior requires effortful attention. 2010. In *Journal of Experimental Social Psychology*, Volume 46, Issue 3, 2010, Pages 551-556. https://doi.org/10.1016/j.jesp.2009.12.019

[12] Lui, Alison & Lamb, George. 2014. Artificial intelligence and augmented intelligence collaboration: Regaining trust and confidence in the financial sector. Information & Communications Technology Law. 27. 1-17. https://researchonline.ljmu.ac.uk/id/eprint/8512/

[13] Federico Milana, Enrico Costanza, and Joel Fischer. 2023. Chatbots as Advisers: the Effects of Response Variability and Reply Suggestion Buttons. In *ACM conference on Conversational User Interfaces (CUI ’23), July 19– 21, 2023, Eindhoven, Netherlands.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3571884.3597132

[14] Jakob Nielson. 2005. Ten Usability Heuristics. https://pdfs.semanticscholar.org/5f03/b251093aee730ab9772db2e1a8a7eb8522cb.pdf

[15] Sajal Regmi and Chetan Phakami Pun. 2024. GPT Semantic Cache: Reducing LLM Costs and Latency via Semantic Embedding Caching. https://doi.org/10.48550/arXiv.2411.05276

[16] Rocket Mortgage. 2025.

[17] Qiaosi Wang and Ashok. K. Goel. 2022. Mutual Theory of Mind for Human-AI Communication. In *IJCAI Workshop on Communication in Human-AI Interaction (CHAI),* July 2022, 7 pages. https://doi.org/10.48550/arXiv.2210.03842

[18] Yimeng Wang, Yinzhou Wang, Kelly Crace, and Yixuan Zhang. 2025. Understanding Attitudes and Trust of Generative AI Chatbots for Social Anxiety Support. In Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25). Association for Computing Machinery, New York, NY, USA, Article 1123, 1–21. https://doi.org/10.1145/3706598.3714286

[19] Zhengquan Zhang, Konstantinos Tsiakas, and Christina Schneegass. 2024. Explaining the Wait: How Justifying Chatbot Response Delays Impact User Trust. In *ACM Conversational User Interfaces 2024 (CUI ’24), July 08–10, 2024, Luxembourg, Luxembourg.* ACM, New York, NY, USA, 16 pages. https://doi.org/10.1145/3640794.3665550

[20] Abdellah Ghassel, Xianzhi Li, Xiaodan Zhu. 2025. Detect, Explain, Escalate: Low-Carbon Dialogue Breakdown Management for LLM-Powered Agents. In *arXiv preprint arXiv:2504.18839.* https://arxiv.org/html/2504.18839v2