Evaluating LLM-Based Chatbots through Touchpoint-Driven Process Models

Carlos H. Camillo da Silva[1], Ana R. Cárdenas Maita[1], Gregorio Assagra de Almeida Filho[1], Caio V. Melo da Silva[1], Gabriel Dimant[1], Renato A. Almeida[2], Guilherme M. Lopes Costa[2], Anna M. Cintra Araujo[2], Barbara Correia Dos Santos Santana[2], Enio Alterman Blay[1], Sarajane Marques Peres[1]

[1]Universidade de São Paulo (USP)
São Paulo – SP – Brazil

[2]Instituto de Ciência e Tecnologia Itaú
São Paulo – SP – Brazil

{ccamillo.silva, ar.cardenasmaita, gregassagraf, caiovmsilva, gabriel.dimant, enio.alterman.blay, sarajane}@usp.br

{renato-augusto.almeida, guilherme.lopes-costa, anna.cintra-araujo,barbara-correia-santos.santana}@itau-unibanco.com.br

Abstract. *This article introduces a method for evaluating the quality of large language model (LLM)-based chatbots, focusing on the user journey. Each dialogue utterance, whether from the user or the bot, is mapped to a predefined set of touchpoints that represent key interaction goals. The resulting dialogue sequences are treated as business process activities, enabling process mining to discover and analyze the process model underlying the dialogue context. This method identifies inefficiencies such as bottlenecks, conversational loops, and shifting responsibility to human agents. It was tested on simulated dialogues generated by an LLM-based financial chatbot, with user personas and dialogue goals also synthesized by LLMs. Although based on synthetic data, the results demonstrate the potential of process mining to uncover structural strengths and weaknesses in LLM-based chatbots.*

1. Introduction

The rise of large language models (LLMs) has enabled the development of increasingly sophisticated conversational agents, transforming the way users interact with digital systems [Zhang et al. 2024]. Furthermore, LLM-based chatbots are now widely adopted in sectors where personalization, empathy, and secure exchanges are critical, such as customer service, healthcare, and finance [de Arriba-Pérez et al. 2022, Rana et al. 2022]. Although current LLM-based chatbots seem highly effective, their performance is still difficult to evaluate from user experience (UX), software engineering, and business decision-making perspectives. Because conversations are inherently fluid, decomposing them into fundamental units may be essential to uncover structural inefficiencies that harm user experience [Ozuem et al. 2025]. In business contexts, process mining can enhance tools for mapping user journeys [Aalst 2016]. This enables organizations to detect friction points and optimize automated interactions like those promoted by chatbots and digital workflows. By analyzing event data, companies gain actionable insights into customer behavior, supporting personalization, operational performance, and data-driven decision-making. Such integration aligns automated interactions with user needs and expectations while revealing critical patterns that influence experience quality. Prior studies show how process models unify data from multiple touchpoints, enabling systemic analysis and optimization [Bernard and Andritsos 2017, Evermann et al. 2017].

In this article, we propose an approach centered on the conversational touchpoints concept, i.e., discrete and meaningful units that, in sequence, shape user experience. Each dialogue turn, whether a chatbot prompt or user request, represents a touchpoint reflecting an interaction moment. Without mapping and analysing these touchpoints, issues like critical bottlenecks, repetition loops, or abandonment triggers remain anecdotal rather than systemic insights for managerial decision support. To fill this analytical gap, we present a method that uses touchpoint mapping as a basis for process mining techniques. The goal is to transform conversation flows into event logs for supporting process modeling and analysis. This allows dialogues to be analyzed, visualized, and optimized as business processes. The main contribution is an integrated approach that classifies interactions into predefined touchpoints, enabling structural analysis of user experience to provide insights for improving the chatbot system and aligning it with the business process. Given the financial context of our work within a banking environment, access to real customer conversations is restricted due to confidentiality and regulatory constraints; therefore, we employ carefully simulated dialogues that replicate the characteristics of real interactions as closely as possible, ensuring both data privacy and contextual relevance.

The remainder of this article is organized as follows: Section 2 reviews background and related work; Section 3 presents the evaluation method; Section 4 describes the experiment and results; Section 5 outlines limitations and future work.

2. Background

2.1. Chatbots and Large Language Models

Chatbots are systems designed to mimic unstructured and often lengthy conversations characteristic of typical human interactions. In industrial contexts, they often serve as digital assistants or frame-based dialogue systems [Jurafsky and Martin 2025]. As LLMs have evolved significantly, they have become the core processing units of modern chatbots, enabling these systems to generate contextually coherent responses and manage complex and multi-turn dialogues. LLMs are advanced neural language models trained on massive datasets and built on the Transformer architecture, which processes token sequences and supports a wide range of natural language processing [Minaee et al. 2024]. Through optimization, LLMs can be further exposed to context-response pairs from real or synthesized conversations over large corpora, and can therefore gain the ability to generate context-aware responses and exhibit emergent conversational capabilities.

LLMs offer a data-centric basis for intelligent dialogue systems, supporting intent recognition, emotion tracking, and journey-based classification with minimal task-specific training. This enables integration into stages like customer journey analysis and personalized recommendations in real or simulated contexts. Combined with advanced architectures, such as autonomous agents [Wang et al. 2024], and semantic retrieval methods like retrieval-augmented generation (RAG) [Gao et al. 2024], they position chatbots as key tools for optimizing customer–business interactions

2.2. Touchpoints

According to Merriam-Webster [Merriam-Webster Dict. 2025], a touchpoint is "an interaction between an existing or potential customer and a business". We extend this to define touchpoints as user–service interaction moments that form the building blocks of the user journey. In chatbot systems, touchpoints include actions like authentication, information retrieval, issue resolution, and intent clarification, all embedded in the exchanged utterances. Extracting these touchpoints from utterances is crucial for converting raw interactions into structured data that enables analysis, interpretability and service improvement.

Rather than following a strictly linear flow, touchpoints in user–system interactions often appear in non-linear, cyclical, or repetitive patterns. Users may revisit, abandon, or re-enter specific touchpoints within a session, making their mapping and analysis a complex yet essential task for accurate behavioral modeling [Bernard and Andritsos 2017]. This complexity becomes a strategic advantage when paired with analytical techniques that reveal patterns and inefficiencies in the user journey. Touchpoints are central to user experience analysis, especially in services involving intelligent agents. Understanding these interaction points supports the enhancement of personalized experiences through artificial intelligence [Rana et al. 2022]. Additionally, moments of service failure and recovery strongly affect user satisfaction and trust in chatbot interactions [Ozuem et al. 2025].

From a functional perspective, touchpoints support four main objectives: identifying critical moments that define the user journey, evaluating service efficiency and responsiveness, enhancing user experience through targeted improvements, and ensuring the reliability and security of the interaction [Rana et al. 2022]. Their exploration may be useful for any rigorous approach to conversational system design and evaluation.

2.3. Process Mining and User Journey Analysis

Process mining [Aalst 2016] connects data science and process science by using event logs to monitor and optimize processes throughout the BPM lifecycle. It models and analyzes event sequences from real data, providing accurate representations of business processes. Here, we define key process mining concepts and relate them to chatbots and user journey analysis. An *event* is the execution of an activity within a process instance; in chatbots, this maps to a touchpoint extracted from an utterance, whether a bot prompt or user input. An event log records sequences of such events, corresponding to touchpoint sequences in dialogue sessions. Each dialogue session forms a *process instance* or *case*, with unique, ordered events. Events may include attributes like *resource*, e.g. a person in traditional processes, but in LLM-based chatbots, this can be an agent component, a specific chatbot persona, or even a user. Crucially, the order of the events reflects the chronological occurrence, tracked by the *timestamps* attribute [Aalst 2016].

User journey analysis is a common process mining task that provides actionable insights through structural examination of interactions [van der Aalst 2022]. Variant analysis reveals different user pathways, including completion or abandonment. Bottleneck analysis spots stages with long response times, indicating inefficiencies. Engagement duration measures identify where users lose interest. Drop-off analysis pinpoints premature exits, while loop analysis detects repetitive exchanges signaling confusion, guiding improvements in intent handling and clarification [Evermann et al. 2017].

2.4. Related Work

The classification of dialogue touchpoints in chatbot interactions builds on research in process analysis and user experience. This review links the literature on discovering interaction patterns to challenges of modeling touchpoints within conversational journeys.

Our core means for analyzing LLM-based chatbots through touchpoints is rooted in process mining. This discipline provides the essential tools to convert unstructured event logs from user-chatbot interactions into structured process models [Aalst 2016]. By doing so, we can automatically discover significant interaction steps and map the flow between them, effectively revealing the fundamental touchpoints of the conversational process. The comprehensive overview by [van der Aalst 2022] frames the advanced techniques applicable to this discovery, allowing for the analysis of complex and variable user dialogues. Once discovered, touchpoints can be sequenced to model the user's journey. The work of [Bernard and Andritsos 2017] is related to our goal, as they use process mining to map customer journey paths, which we conceptualize as sequences of touchpoints. The context for enhancing these journeys is provided by [Rana et al. 2022], who reviews how AI can enrich the experience at each touchpoint. Predictive methods, such as those using deep learning [Evermann et al. 2017], suggest a powerful extension: forecasting the user's likely next touchpoint, enabling proactive intervention at critical moments of the interaction. Furthermore, the literature covers the critical task of classifying the nature of touchpoints. Studies show how dialogue analysis can define touchpoint categories. For instance, [de Arriba-Pérez et al. 2022] shows how conversational cues can be classified as touchpoints indicating a user's cognitive state. In service contexts, [Ozuem et al. 2025] provides a theoretical basis for identifying and classifying critical "frustration" or "service failure" touchpoints. Finally, [Samuel et al. 2024] contributes to the idea of "persona-defining" touchpoints, which are crucial for evaluating whether an interaction reinforces or undermines the chatbot's intended character and user trust.

Finally, since our focus is on LLM-based technology, it is worth noting that [Zhang et al. 2024] emphasizes the importance of systematically analyzing conversation data, particularly given the scale and complexity introduced by LLMs. In their studies, the authors present a framework that structures conversation analysis around key stages such as scene reconstruction and causality analysis, providing methods to extract actionable insights from interaction logs. This supports more accurate and data-driven evaluations of user journeys in complex conversational systems.

3. Method

Our approach to evaluating LLM-based chatbots focuses on analyzing the user journey mapped onto a process model. The method for this is structured into three stages designed to transform unstructured dialogues into analyzable data, enabling a deeper understanding of the user journey in alignment with the principles outlined in [Bernard and Andritsos 2017] and process mining. The execution of the approach requires the extraction of a set of dialogues from a communication platform used by an organization and its clients. For the purposes of this study, both the chatbot system and the (simulated) users were modeled in an experimental environment for evaluating our approach. Figure 1 illustrates the method used to test our approach, along with its constituent phases.

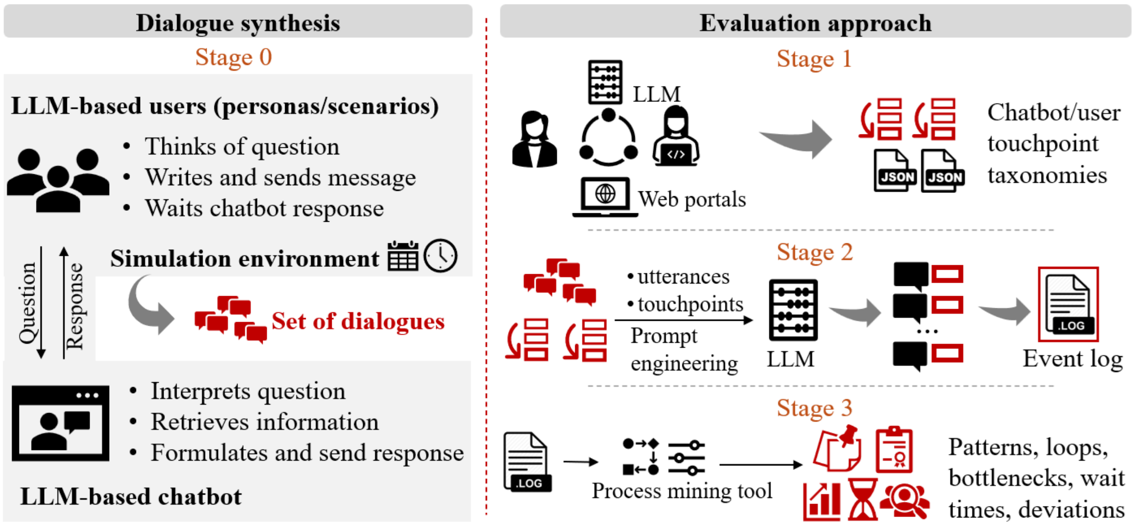

Figure 1. LLM-based chatbots evaluate approach - overview

Stage 0: Synthetic Dialogue Generation – The generation of synthetic dialogues in our framework combines prompt engineering with persona simulation to reproduce realistic user interactions within a financial business process domain. An LLM-based chatbot was implemented using an agent-based architecture and a RAG approach, leveraging publicly available descriptive materials about credit card products and services. To simulate users, personas were designed via LLMs based on interrelated characteristics including demographic data (age, location, education), behavioral traits grounded in the Big Five personality model (e.g., impulsivity, anxiety), and financial profiles (credit history, card usage, payment behavior). These attributes enable LLMs to generate coherent, context-rich dialogues reflecting diverse user needs in typical financial service scenarios. Personas were associated with scenarios (like in [Samuel et al. 2024]) representing concrete user intentions (e.g., account access issues, credit limits, or dissatisfaction with service). Simulated interactions with the chatbot were carried out through dialogues, where personas expressed concerns and expectations via structured one-shot or few-shot prompts. Additionally, dialogue generation was instrumented to vary parameters such as the time and day of occurrence (business hours vs. off-hours; weekdays vs. weekends) and interaction speed, simulating human response times[1]. This setup imposed realistic constraints on both users and the chatbot, enabling the generation of diverse and complex dialogue flows that reflect varied user behaviors and service demands.

Stage 1: Definition of the touchpoint taxonomy – The first step involved developing a touchpoint taxonomy divided into two categories to capture the bilateral nature of interactions: those directed toward the chatbot and those associated with user behavior. The initial model was created experimentally and iteratively, starting with the identification of the most common user complaints about credit card services, based on a general search in public consumer complaint platforms, which provided a ranking of the most frequent

[1]User response time was set mainly according to the persona's typing speed.

issues and originated the initial set of touchpoints. Subsequent the analysis of sample dialogues generating by a LLM revealed that some touchpoints were irrelevant or some in need of refinement. This evaluation led to the addition of new touchpoints and adjustments to existing ones to better capture the diversity of conversational patterns in the dialogue flows. Importantly, the creation of this touchpoint taxonomy is domain-specific and tailored to organizational needs, limiting its generalizability across contexts. Table 1 presents examples of the established touchpoints, organized by category, description, and sample utterances expected to be classified under each touchpoint category[2].

Table 1. Examples of chatbot and user touchpoints

Category	Interaction Description	Example Utterance
Chatbot greeting	The chatbot starts with a formal greeting.	Hello! I'm the virtual assistant for Bank X. How can I help you?
Intent detection	The chatbot identifies the user's need.	I understand you want to know your card limit.
Data response	The chatbot returns the requested information.	The available limit on your card is $2,500.00.
Repetition of intent	The user repeats the same request due to not feeling understood	I said I want to know the limit.
Stress detection	The user exhibits high stress or panic.	Someone is using my card! Help me!
Spontaneous Thank You	The user ends the conversation with gratitude	Thank you for your help!

Stage 2: Automatic extraction of touchpoints from utterances and event log generation – In this stage, the identification and labeling of each utterance in a dialogue were automated using an LLM guided by prompts containing the information summarized in Table 1. This approach ensured contextualized, consistent, and scalable classification. The prompt for touchpoint extraction was iteratively refined to minimize hallucinations, preserve contextual integrity, and enhance reliability, enabling accurate mapping of utterances to touchpoints across diverse dialogue scenarios. Each utterance was then assigned to exactly one touchpoint, transforming the conversational flow into a structured sequence of events. This procedure was implemented within our framework through a dedicated script that receives three inputs: JSON files containing the definitions of the touchpoints (chatbot and user) and a dataset of dialogues. Based on these inputs, the script generates a structured CSV output that conforms to the standard format of an event log, making it suitable for subsequent process mining analyses. The event log includes the following attributes: dialogue identifier (case), event identifier (utterance identifier), activity (touchpoint), start timestamp (message received), end timestamp (message sent), chatbot agent indicating the RAG modules used to generate the response (resource), and utterance type specifying whether it originated from the chatbot or the user (event description attribute).

Stage 3: Analysis of the process model underlying the user journey in the dialogue – In this stage, we applied process mining techniques to the event log derived from chatbot and user interactions. Process mining discovered a touchpoint-driven process map char-

[2]Touchpoints taxonomy, event log, and analytical results were originally developed and the experiments conducted in Portuguese, then translated into English for this paper.

acterizing the process model underlying the dialogues. This map captured the actual sequence of user interactions and system responses, providing a data-driven representation of the emergent behavior of the LLM-based chatbot systems. Using process mining, we extracted insights into how users navigated the conversational flow and whether the chatbot's objectives were achieved. Both functional and non-functional requirements could thus be verified or even elicited from the structured exposure of such behavior.

4. Experiment setup, results and analysis

4.1. Dialogue synthesis, touchpoints taxonomy and extraction of touchpoints

The simulation environment[3] built for generating dialogues was executed using two parallel instances of the GPT-4.1-2025-04-14 model, being one for the chatbot and the other for simulating the user. A set of 25 personas/scenarios was used to generate 300 dialogues. The simulation distributes the dialogues over time and also forces them to have different durations. The dialogues contain varying numbers of turns depending on the decisions made by both LLM instances during dialogue generation. The structural quality of the dialogues (correct interchanging between chatbot and user turns) was checked automatically. The semantic quality was verified manually through sample inspection.

The touchpoint taxonomies consist of 29 categories (chatbot) and 35 categories (user), covering concepts as greetings, intent understanding, request processing, engagement, and security. The touchpoints extraction for each utterance was performed using a third instance of the GPT-4.1-2025-04-14 model. The touchpoint extraction process depends on the LLM's ability to interpret utterances and touchpoint descriptions and to correctly associate them. Since subsequent process mining analysis relies on the event log derived from this extraction, its quality is critical. Despite iterative refinement of the prompt engineering for touchpoint extraction, the process remains sensitive to the model's performance. We inspected a sample of the extracted touchpoints and found the results generally satisfactory, though some few associations did not meet our expectations.

4.2. Process mining analysis

The Disco tool (Fluxicon[4]) was used for process mining: model discovery of user journeys, extraction of descriptive statistics, and analysis of drop-offs, loops, and bottlenecks.

Process discovery and user journey modeling: A process model was discovered using the event log generated from the 300 dialogues, and it serves as a *de facto* mapping of the different paths the dialogues followed during the simulation. Figure 2 shows two versions of the process model: the first one (top left) is a simplified version generated displaying 25% of the activities present in the original model, and 0% of the paths, meaning it retains the minimum number of paths required to ensure a sound process map; the complete process model (bottom right) displays the full complexity of the user journey in the dialogues. In the figure, rectangles represent touchpoints and arcs indicate their sequencing within the dialogues. Darker colors and thicker arcs denote higher frequencies of occurrence for a given touchpoint or sequence. The numbers show absolute occurrences. The simplified

[3]The artifacts, all dialogues, touchpoint definitions, and persona designs used in this study are published in the project repository, while any material containing references real context remains excluded: `https://github.com/c2d-usp/analise_jornada_usuario_chatbot.git`

[4]`https://fluxicon.com/disco/`

model allows us to visualize the most frequent sequences in the set of 300 dialogues. For example: in 29 dialogues the user's first utterance after the chatbot's initial greeting was identified as a stress manifestation; in 61 cases of stress manifestations, the chatbot responded with proactive clarifications; the chatbot provided 83 no-solution responses, of which 53 were followed by the user ending the conversation with a complaint.

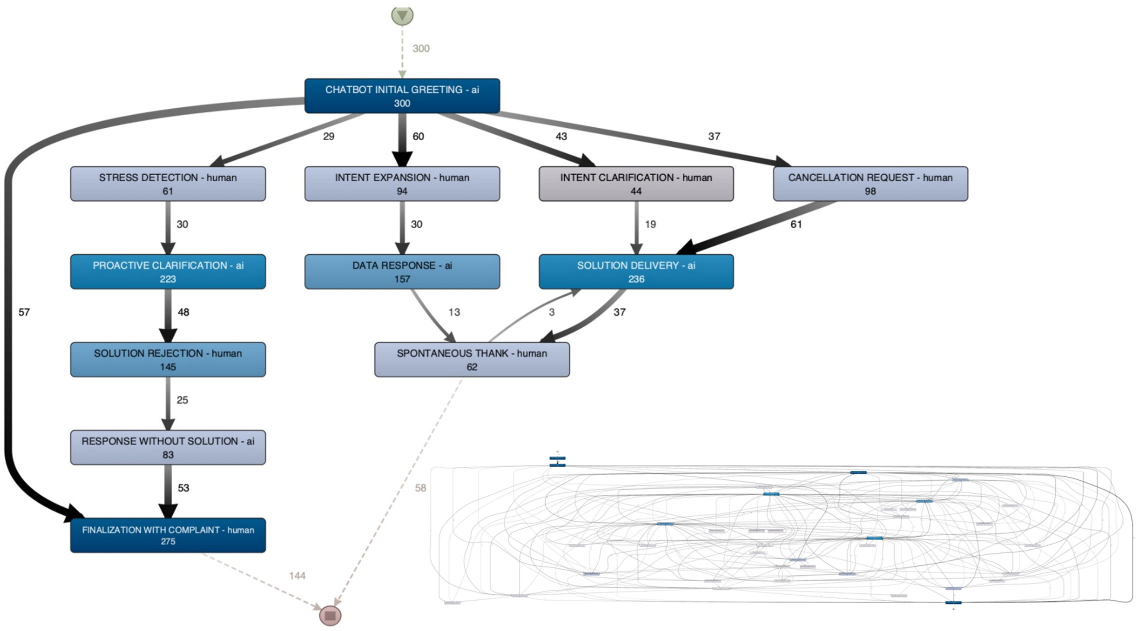

Figure 2. Process models: simplified view (top left); full view (bottom right)

Descriptive statistics: Table 2 lists the main statistical and temporal information extracted from the process mining analysis. It is noteworthy that the user journey exhibits high complexity, as the number of different paths (variants) accounts for 79% of all dialogues. Moreover, the most common variant occurs only 9 times (3.33% of the dialogues), characterizing the process as unstructured and indicative of a knowledge-intensive process, which is inherently difficult to analyze.

Table 2. Main descriptive statistics

Statistics description	**Value**	**Statistics description**	**Value**
Total of dialogues	300	Time period	30 days
Total unique paths (variants)	263	Maximum duration	1h34″
Main variant coverage	3.33%	Minimum duration	1′44″
Longest variant	18 steps	Shortest variant	4 steps
Total of touchpoints occurrence	2.084		

Statistic description	**Value**	**Label**
Most frequent chatbot touchpoint	13.2%	"Finalization with complaint"
Most frequent user touchpoint	11.3%	"Solution delivery"
Rarest chatbot touchpoint	0.05%	"Partial solution"
Rarest user touchpoint	0.05%	"Security confirmation"

Drop-off analysis: The drop-off analysis has the potential to reveal process points where user abandonments are highly concentrated. As an example of this analysis, we tracked abandonment by monitoring progressive user loss across touchpoints. For instance, if 60 touchpoints transition sequence from “Chatbot Initial Greeting” to “Intent Expansion”, but only 30 proceed to “Data Response”, this indicates that 30 dialogues (50%) dropped off at that point. Following this strategy, the main touchpoints with the highest drop-off rates can be identified. In the simplified version of the process model, we highlight the following critical touchpoints for drop-off situations: “Intent expansion” - addition of information, “Intent clarification” - correction of information, “Solution delivery”, and “Solution rejection”. Mitigation strategies could include refining chatbot prompts and the knowledge base to provide clearer explanations, incorporating contextual confirmations, or delivering user-aligned solutions.

Loop analysis: The loop analysis revealed patterns of repetitive interactions within the chatbot conversations. The frequency of each loop and the number of dialogues in which they appear provide valuable insights into points where users may experience frustration or confusion. To illustrate this analysis, we identified two loop situations in our experiment: chatbot repeatedly asks for missing or incomplete information (in 90 dialogues); unnecessary confirmations before proceeding, causing user impatience (in 45 dialogues).

Performance analysis: Process mining allows the analysis of turn durations. The performance information highlights bottlenecks by indicating dialogue turns with prolonged wait times that may degrade the user experience. In our simulation, “Cancellation request” and “Finalization with complaint” emerged as stages with significantly extended durations, suggesting the chatbot struggles with complex cancellation processes and user complaints. Additionally, dwell time analysis could reveal how long users remain stuck at specific activities or conversation blocks within the interaction. By comparing long-duration and short-duration cases, this analysis reveals which parts of the journey effectively maintain user attention and which may cause delays or disengagement. For example, user turns with “User Feedback” were very fast in our simulation, while turns involving “Stress Detection” were among the slowest.

5. Conclusion

The core contribution of this work lies in introducing a method for evaluating LLM-based chatbots through the abstraction of conversations into sequences of touchpoints. We define touchpoints as discrete points of contact that classify the purpose of each dialogue segment. This approach transforms the inherently unstructured conversational flow into an event log suitable for formal analysis. In this way, we leverage the full analytical capabilities of process mining. Analyzing sequential touchpoint patterns makes the abstract notion of the user journey tangible and measurable. In practice, this analysis has proven to be a valuable diagnostic tool for identifying issues in LLM-based chatbot systems and optimizing them to align more closely with business process objectives.

Certain limitations should be considered. Although synthetic data was generated in a controlled and diversified manner, it may not fully reflect the spontaneity and variability of real-world human interactions. Also, the accuracy of touchpoint classification depends on the performance of the third-party LLM. Thus, prompt engineering plays a critical role, requiring clear and linguistically nuanced formulations to ensure each utterance is correctly assigned to its touchpoint category. Next steps in this research comprises a validation experiment for the touchpoint extraction and enhancing the simulator with adaptive time dynamics, varying user response times based on persona-emotional states.

Acknowledgements

The authors would like to thank the Programa de Bolsas Itaú (PBI) of the Centro de Ciência de Dados (C^2D), supported by Itaú Unibanco S.A

References

Aalst, W. M. P. (2016). *Process Mining – Data Science in Action*. Springer.

Bernard, G. and Andritsos, P. (2017). A process mining based model for customer journey mapping. In *Proc. CAiSE Forum and Doctoral Consortium*, pages 49–56. CEUR-WS.

de Arriba-Pérez, F., García-Méndez, S., González-Castaño, F. J., and Costa-Montenegro, E. (2022). Automatic detection of cognitive impairment in elderly people using an entertainment chatbot with natural language processing capabilities. *J. Ambient Intell. Humaniz. Comput.*, 14:16283–16298.

Evermann, J., Rehse, J.-R., and Fettke, P. (2017). Predicting process behaviour using deep learning. *Decis. Support Syst.*, 100:129–140.

Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., Dai, Y., Sun, J., Wang, M., and Wang, H. (2024). Retrieval-augmented generation for large language models: A survey. *Preprint arXiv:2312.10997*.

Jurafsky, D. and Martin, J. H. (2025). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition with Language Models*. 3rd edition. Online manuscript released Jan. 12, 2025.

Merriam-Webster Dict. (2025). Touchpoint. `https://www.merriam-webster.com/dictionary/touchpoint`. Accessed: 2025-07-19.

Minaee, S., Mikolov, T., Nikzad, N., Chenaghlu, M., Socher, R., Amatriain, X., and Gao, J. (2024). Large language models: A survey. *Preprint arXiv:2402.06196*.

Ozuem, W., Ranfagni, S., Willis, M., Salvietti, G., and Howell, K. (2025). Chatbots, service failure recovery, and online customer experience through lenses of frustration–aggression theory and signaling theory. *J. Serv. Mark.*, 39(5):493–512.

Rana, J., Gaur, L., Singh, G., Awan, U., and Rasheed, M. I. (2022). Reinforcing customer journey through artificial intelligence: a review and research agenda. *Int. J. Emerg. Mark.*, 17(7):1738–1758.

Samuel, V., Zou, H. P., Zhou, Y., Chaudhari, S., Kalyan, A., Rajpurohit, T., Deshpande, A., Narasimhan, K., and Murahari, V. (2024). Personagym: Evaluating persona agents and LLMs. *Preprint arXiv:2407.18416*.

van der Aalst, W. M. P. (2022). Process mining: A 360 degree overview. In *Process Mining Handbook*, volume 448 of *Lect. Notes Bus. Inf. Process.*, pages 3–34. Springer.

Wang, L., Ma, C., Feng, X., et al. (2024). A Survey on Large Language Model based Autonomous Agents. *Front. Comput. Sci.*, 18:186345.

Zhang, X., Yu, H., Li, Y., Wang, M., Chen, L., and Huang, F. (2024). The imperative of conversation analysis in the era of LLMs: A survey of tasks, techniques, and trends. *Preprint arXiv:2409.14195*.