A chatbot developer’s guide to handling erroneous situations

Jeffrey Eboreime

Laurea University of Applied Sciences

A chatbot developer's guide to handling erroneous situations

Jeffrey Eboreime

Business Information Technology

Bachelor's Thesis

May, 2023

Laurea University of Applied Sciences **Abstract**

Business Information Technology

Bachelor's Degree

Jeffrey Eboreime

A chatbot developer's guide to handling erroneous situations

Year 2023 Number of pages 46

The aim of this research was to evaluate chatbots' conversation flow designs, especially their error handling mechanisms and strategies. The purpose was to suggest practices to improve conversation quality and increase user retention and chatbot interactions by attempting to directly resolve common causes of failing chatbot conversations. The concern was that many conversations fail and result in escalation to second-line support channels. These types of conversations should be actively managed rather than using generic responses that may not work to restore conversation flow.

The project outcomes presented chatbot design solutions and communication strategies to improve ways of creating conversation flow. The focus was not on how chatbots ought to be built, but on how to improve responsivity when it did not know how to communicate with its users.

The studied chatbots were deployed in the finance sector. All data and findings, though anonymized, were based on real conversation history. The research showed how changing different elements of chatbots' ways of communicating with their users affected conversation flow and various indicators of conversation quality.

How a chatbot should behave in an erroneous situation may not be easily defined, as not all chatbots are deployed to serve the same function, such as a self-service chat or a contact deflection method. Methods that work for one instance of a chatbot, may result in adverse effects on another. Therefore, conversation design, system responses, and overall chatbot behavior should be tailored for each deployment.

Keywords: Chatbot, NLP, Artificial Intelligence, errors

Abbreviations and terms

AI	Artificial Intelligence
Chatbot	A conversational AI
KPI	Key Performance Indicator
HCI	Human-Computer Interaction
IVR	Interactive voice response
LLM	Large language model
NLP	Natural Language Processing
SaaS	Software-as-a-service
UI	User Interface
VUI	Visual user interface

Contents

1 Introduction

This research is done in partnership with an IT consulting firm that provides conversational AI (Artificial Intelligence) solutions for various enterprises. The client wishes to remain anonymous but has provided access to actual conversation history and platforms to better allow this research.

There are numerous available development guides for deploying chatbots in a service environment and how to create meaningful conversations with them, however, developing chatbots for dysfunctional situations is an often-overlooked matter. Few literary works directly address the issue of what to do when a chatbot does not work as desired during a conversation and faces communication errors with its human end-users.

The often-ignored truth is that not all conversations with chatbots are successful and may leave their end-users dissatisfied. Though chatbots often have systems in place to catch these errors or exceptions, *how* those situations are managed are just as important to a bot's functionality as are the main topics it is designed to assist with. This study explores common chatbot errors and how conversation design in erroneous situations can help recover a deteriorating conversation.

1.1 Objectives and scope

This research aims to improve chatbot error handling mechanisms and system responses and evaluate how their design affects overall conversation quality, user retention and interactions. This will be done by analyzing historical conversation data to find root causes of unsuccessful conversations and suggesting methods and practices for improvement of both user and chatbot interactions and experience.

Research will be conducted on different chatbot instances deployed in personal financeoriented services. The chatbot should be available upon landing on the company or brand's homepage, though are not required to have human-to-human chat capability but must at least be able to accurately respond to users asking for human assistance.

2 Background of Conversational AI

When a person calls a company or brand, it is because they want to gain information about something, conduct some type of action, or because something has happened and troubleshoot a situation. This contact request is answered by first-line support - customer service or a ServiceDesk, depending on the business. They're tasked with providing common assistance with customer inquiries. This includes collecting necessary information, diagnosing an incident, and possibly resolving the issue (Zitek n.d.). Rather than hiring human labor to assist with minor inquiries about repetitive or mundane information, such as business operating hours, common service charges, or suggesting a device restart; that employee may be more valuable handling escalated issues in second- or third-line support.

Technology has changed the way people communicate and has forced businesses to adapt to a new type of communication to achieve success. In an "always connected" society the more a business grows, the more it costs to deliver customer service (Zabój 2022). By deploying chatbot first-line services, brands can significantly scale up their ability to handle incoming contact requests, while reducing the costs and need for human labor.

Each time a human service agent is required to assist in a phone transaction, a cost per call is incurred. This cost can be calculated as: *Cost per call = Total cost of all calls (human labor)/Total number of calls*. The industry standard ranges the cost per call between €2.50€6, depending on the type and size of the business (LiveAgent n.d.). This concept also applies to chat-based services as cost per conversation, or generally as cost per transaction.

Chatbots, including voicebots, and conversational AI (Artificial Intelligence) are "a synthetic brainpower that makes machines capable of understanding and responding to human language" (Boost 2023). They are efficient conversational bots capable of having many simultaneous conversations at any time of the day with minimal downtime.

Chatbots utilize a branch of AI called Natural Language Processing (NLP) to comprehend, generate, and manipulate human language with natural language text or voice (Oracle n.d.). NLP technology provides an advanced language engine and chatbots' service platforms provide visual interface with varying features and solutions to manage and deploy chatbots.

2.1 Limitations of chatbots

Despite utilizing advanced technologies, AI and other software applications including chatbots have their limitations and shortcomings. Much of chatbots' functionality comes not

just from their technologies, but from their conversation design, and their dysfunctionality often comes from lack thereof.

Bouzid and Ma state that “the quality of a chatbot is judged not only by how the chatbot manages interactions, but also how it navigates situations when things don’t proceed as expected” (2022, 89). This can be related to the common idiom “a chain is only as strong as its weakest link”. Not all problems are necessarily a fault of the chatbot, though whether or not the chatbot states that to users is a part of conversation design.

2.2 Common types of problems

Due to numerous technical and operational problems before a chatbot is even deployed, the focus of this research evaluates existing chatbots’ various interaction problems, limitations, and root causes. Hardware, software, and platform specific problems will be excluded.

Many chatbots are very sophisticated input-output applications. Just like they are made to accurately respond in a relevant manner to the input, they are programmed a set of system responses to catch specific types of errors and report it to the user. In common problem situations, the bot is usually self-aware of its own state and can communicate itself at least in a generic fashion. Four problems with chatbots as identified by Bouzid and Ma (2022, 9091) are:

- No-input: The bot did not receive any input from the user.
- No-match: The bot received input but does not know what to do with it.
- Misrecognition: The bot misunderstood the input and output the wrong response.
- System failure: Internal resource, platform, or other technical error.

As these are commonly known problems, they are very important to be addressed appropriately to allow for successful conversations between chatbots and their users. As with common problems identified, the respective human error causes have too:

- The user has not responded.
- The user said too much or too little, or the wrong way.
- The user was not concise or clear enough in their expression.
- Technical limitations or programming errors.

Even with chatbots’ AI technology and the highlighted features, expectations are not always met when their limitations and shortcoming are masked by humor and counter questions (Marttinen 2020, 106). With the ability to communicate error provided, *how* it communicates its errors to users is a matter of conversation design, even when not working as expected.

2.3 Software applications

Hong Zhu (2005, 48-50) states that software design is complicated and one of the most difficult tasks in software development. All designers face difficulties, but the question is if there are specific reasons that make software design even more difficult. He identifies two types of causes of software development difficulties: *the essences* and *the accidents*. Essences are the inherent difficulties in software. They are irreducible, will not disappear, and cannot be solved through technical solutions -they *are* technical problems. Accidents are difficulties that attend its production but are not inherent – the design problems.

Zhu describes four essences of difficulties in software development as:

- **Complexity**. Software entities are complex in terms of the sizes of their state spaces and how they function, which makes it hard to conceive, describe, and test software. Digital computers are more complex than most things people build and more so is the software they operate.
- **Conformity**. Software is expected to conform to the standards of other existing software, or external bodies – people. Much of the complexity is arbitrary, forced without reason to conform to many human institutions and systems to which the interface of the software must conform.
- **Changeability**. Software suffers constant need for change. It embodies the function of the system that feels most of the pressure of change. The most fundamental reason why software is under pressure to change is because software brings profound changes to our life toward the age of information. Such changes in turn demand more changes in the software.
- **Invisibility**. Unlike most human-made artefacts, software is invisible. The representations used to describe software lack visual links that can provide an easily grasped relationship between the representation and the system. This impedes the process of design and communication among others (2005, 49).

Zhu's identified essences apply well to chatbot development because chatbots are complex software designed for hardware that is designed for human input. The platforms they operate on usually do not run on local computers, rather are deployed and managed on remote servers through cloud applications.

AI and chatbots are often advertised to utilize the latest technologies, and they do, but creating software to conform to human needs, behave in familiar ways yet offer something entirely new makes designing software and the operations they perform difficult. It requires

constant change from those that use it and develop it. Often once a software is released, the next step is to iterate and begin working on the next version with updates or fixes.

2.3.1 Error handling in software

Error handling in a software application refers to the response and recovery procedures from present error conditions. It is the process comprised of anticipation, detection and resolution of application, programming, and communication errors. Its purpose is to maintain normal flow and gracefully help to resume program execution when interrupted (Rouse 2017).

Error-handling applications can resolve logical and runtime errors or have their impact minimized by adopting reasonable countermeasures. Rouse claims that “error handling is one of the crucial areas for application designers and developers, regardless of the application developed” (2017) and in worst-cases forces the application shut down.

Chatbots’ system responses are their error handling mechanisms. When the NLP, platform, or connected third-party applications return an error, chatbots should have effective recovery procedures prepared.

2.4 Challenges

Along with human error and software limitations, one of the challenges chatbots, particularly voicebots, arise from the flow of natural conversation. While interacting with a text-based chatbot, previous messages are usually visible to the user; they can scroll up and see what was said earlier. Spoken conversations function differently.

Without a visual user interface (VUI) callers need to remember more information to progress throughout the conversation. Bouzid and Ma state that a voice-based conversational interface is unidirectional, ephemeral, and invisible. These linear types of conversations demand more attention from users, force them to speak up, and engage in a more focused way (2022, 1821). With voicebots concise inquiries and responses work well, though chatbots may struggle if the conversation de-linearizes or deviates from designed conversation flows.

2.4.1 Linear conversations

A simple linear conversation can be summarized as follows: a human greets the chatbot, the chatbot greets the human; the human asks a question, the chatbot answers; the human may ask a follow-up question or says farewells, and the bot replies in a relevant manner.

Figure 1: A linear conversation with no deviations.

Conversations may not always follow a perfect path from start to finish, and situations may arise where users ask follow-up questions or give input that chatbots may have trouble understanding. In other words: the human greets the chatbot, the chatbot responds; the human asks a (another) question, and the chatbot does not understand or misrecognize the input. This causes a critical moment in a conversation because the conversation flow is offtrack, or de-linearized. Chatbots can inform their user that they didn't understand and ask them to rephrase, but that alone may not resolve the issue. Figure 2 below illustrates delinearization and a generic attempt to restore the conversation.

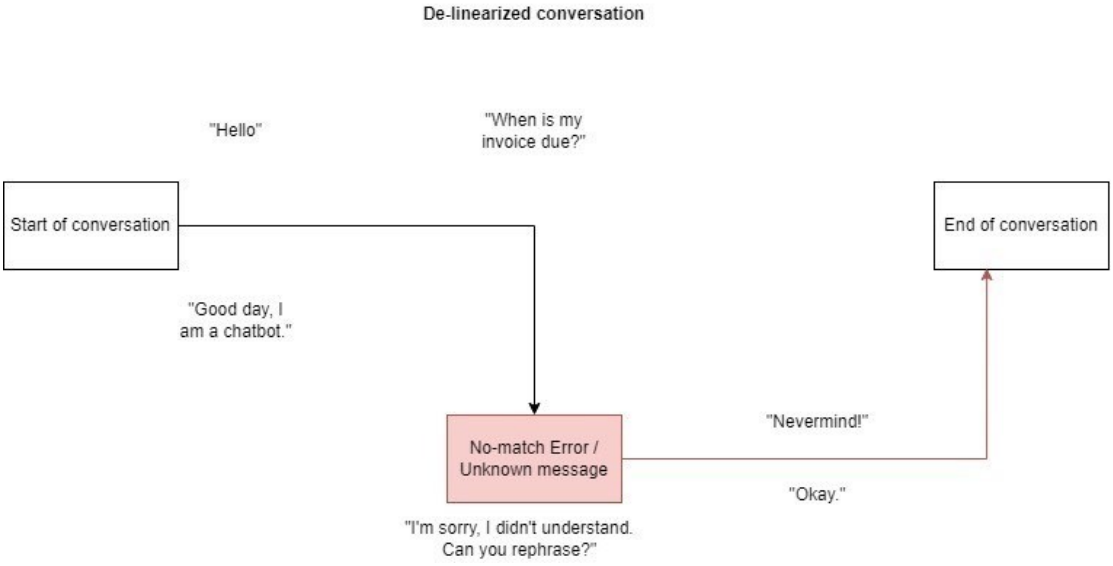

Figure 2: A linear conversation with a no-match error can cause a conversation to end prematurely.

These de-linearized conversations are commonly identified sources of failing conversations that may lead to escalation to second-line support or premature termination. Premature termination suggests a user exited the conversation, for example by hanging up, before their

reasons for contacting a brand was clearly resolved. Though chatbots can communicate to users what went wrong, it is important to try to *actively* restore conversations and steer them back on track. One of the objectives of this research is to find solutions to effectively manage these types of conversation situations.

3 Research method and theories

This research will be conducted using an action research cycle to best produce outcomes. Action research is a highly interactive method of conducting research and taking action at the same time to resolve an issue. It prioritizes reflection and bridges the gap between theory and practice (George 2023).

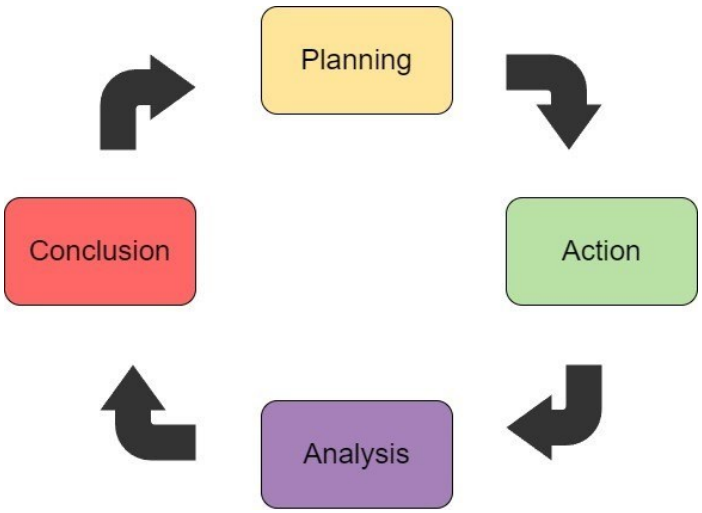

Figure 3: The action research cycle illustrated.

Much of the applied theory and practices are based on studied literature regarding software development and chatbot development guides. Though not all suggestions may apply to all chatbots included in the study, the theoretical framework to produce progressive and meaningful conversations with human users remains.

The research is based on existing software automation and ServiceDesk practices, knowledge, and theory. It is to find optimal ways of handling chatbot conversation errors gracefully and effectively and evaluating how conversation design has an impact on chatbot usability. Outcomes ought to show that developing proper chatbot error handling guidelines and practices will yield increased successful conversation rates and decreased escalation rates, and suggest best practices chatbot developers can use to improve new or existing chatbot solutions.

Several key performance indicators (KPI's) will be considered when analyzing data, none of them are necessarily indicators of shortcomings but act as points of reference. Brands may not give equal significance to all KPI's but utilize them as a basic measure of chatbot performance.

3.1 Defining metrics

From May 1, 2022, to May 1, 2023, the client's chatbots had an average of 110,750 conversations per month. With such a high volume, it would not be viable to review each individual conversation. The client suggests their customers set internally agreed quotas on how many conversations to review, but each customer or brand may review conversations according to their own business strategies.

It is important to note that information given to users, or actions performed may not necessarily be to the users' satisfaction, though are otherwise accurate or successfully performed in terms of the chatbots functionality. For example, if a user asks about an invoice's due date and the chatbot responds by telling them it's past due, it is not necessarily the wrong answer despite the users' dislike of the response. Chatbots users are often allowed to give conversation feedback, and though end-user feedback is valuable, it is not an accurate measure of chatbot performance for this research and will not be considered in analysis.

3.1.1 Chatbot performance

Conversation quality ratings are one of the metrics targeted for improvement in this research as they are one of the main KPI's developers use to measure chatbot performance. Platform providers and the client recommend partnered brands to use specified guidelines when reviewing conversations. Successful conversations can be generalized as conversations where the chatbot responded to users to the best of its abilities about relevant topics, even if the conversation was not entirely without error. Brands may have different standards for defining their meaning of successful conversations and its significance as a KPI.

3.1.2 Communication terms

A message is the single input from a human user to a chatbot, or from a chatbot to a human user, in either written or spoken form. Questions, statements and remarks, button-clicks, or other interactions, and even typos or mumbled speech received by the chatbot are also considered an input message.

Conversations are interactions, or sessions, with chatbots. A conversation might consist of a single or many messages or interactions with a chatbot. Generally, one conversation comes from one human at a time, though they can reinitiate contact and have multiple conversations in a day.

On chatbot platforms, when a user provides a message and receives a trained response this is often referred to as making a prediction to an intent. Intents can be thought of as units of knowledge, often in a hierarchical structure, but depending on the platform may be an umbrella term for broader concepts or groups of knowledge. Predicting is a term used to describe the NLP processes that translate user messages into trained chatbot responses.

3.1.3 Definition of success

To consider a message or prediction successful, the users' message should be interpreted, and the chatbot respond with a relevant or at least trained response to the inquiry, provided what they're requesting is within the chatbots' designed and available knowledge, or scope.

A successful conversation can be defined as a conversation where the user makes successful prediction(s) to intents and receives accurate or acceptable responses without too much trouble. In simpler form, they ask a question(s) and get enough correct responses.

For this research, success rate can be calculated as:

*Success rate % = amount of successful conversations/ all reviewed conversations *100*

3.1.4 Definition of failure

A failed message can be defined as a single input from a user that the chatbot was unable to interpret and failed to make a prediction to an intent. This often happens when a users' message is too short, long, or too obscure, and often returns a No-match type error called an unknown message. Unknown messages are valuable KPI's for measuring chatbots performance and will be an important element in this research. Unknown messages result in specified system responses for chatbots to communicate their inability to understand to users. Though it is technically a type of trained response, it is not a favorable one or considered a prediction.

A failed conversation is where a user gives relevant input, but the chatbot fails to provide a relevant response(s) enough to resolve the user's matter. Like unknown messages, repeat failed predictions are a commonly identified reason for conversations being terminated prematurely. This only applies to conversations where the bot *should* be able to make a

prediction to an available intent. Conversations where users interact with chatbots in ways that they were not meant to be able to respond to are considered out-of-scope. These can be a useful KPI for brands to plan their chatbot strategies and out-of-scope messages may affect the overall rating given to a conversation. Out-of-scope conversations will be omitted from this study and are not otherwise identified as a type of failed conversation.

The client does not measure a failure rate but does measure an unknown message rate. It can be calculated as:

*Unknown Messages % = all unknown messages/ all messages *100*

3.1.5 Escalation

For the purposes of this research, conversations escalated to human-agents in second-line support will be considered failed conversations to better define metrics. In practice, different brands give varying significance to escalation as a KPI. Escalation is not always an undesirable event, but generally what is being thwarted by deploying a chatbot.

Escalation rates are not calculated directly, instead rates of conversations handled only by the chatbot are used. Conversations not handled solely by a chatbot infer escalation.

3.1.6 System responses

Other system responses will be considered when calculating metrics but may not otherwise be significant to the objectives of this study. Included system responses include statistics and methods for managing greetings, unknown messages, excessively long messages, as well as conversation escalation strategies.

3.2 Data collection

Most conversation data evaluated is available on the chatbots' platforms and are filterable and sortable to highlight various elements, such as specific error types or intent predictions. This data can be further refined to make comparisons and draw conclusions from development efforts. How each brand utilizes this statistical information is a part of their business strategy, however, most data is generated consistently by the platform itself.

Qualitative data is comprised of the number of conversations and the different refinements of them to create relevance for this study. This includes amounts of conversations, specific messages, interactions and responses, and intents, and all data evaluated will be from a defined date range. As there is a lot of conversation history available, reviewed

conversations quality relies on conversations rated by various brands' developers and their strategies. Though this may aggregate data, rating conversations follows a suggested guideline and is still an accurate measure of chatbot performance.

Qualitative data includes end-users' sentiment and demeanor towards the chatbots, and how their interactions may change with modifying conversation design. Prematurely ended conversations may be considered qualitative data, however, may not be a strong indicator of an error situation. This also includes the strategies utilized to navigate conversation, whether conversation design was generic or proactive.

3.3 Applied theories

Software applications are abstract concepts designed to do virtually anything we want them to do. They are virtual realities that humans have created and are governed largely by developers' design with parameters set to specification. Chatbots too are software applications and can be governed and diagnosed with rules and conditions.

3.3.1 Sermorobotosis

Sermo, a Latin word for speech and *-osis*¸ denoting a condition, give birth to the authors concept of sermorobotosis, a dysfunction of the conversational robot. It is the inability of a conversational AI to prevent itself from progressing into meaningful conversation, or from regressing into meaningless conversation, until critical failure. In other words, it is the chatbot's inability to stop giving valid answers, or stop giving wrong answers, until the conversation ends.

This can mean providing too much (right or wrong) information to the user, or not enough (right) information. Chatbots can only respond to the best of their trained abilities but do not necessarily know when or if they have given a wrong answer. They might be able to interpret what a user is saying, but simply are not able to give any better answer due to limited knowledge about the subject. However, it can still understand users well enough to give an answer at all, even if wrong. The chatbot tries its best and essentially fails upwards.

For example, if a user repeatedly asks about their previous month's credit card invoice, and the chatbot only responds with the current month's invoice (due to its limited knowledge about invoices), the user is getting the best possible answer, which is meaningful within the right context, but meaningless to the user's request. Enough failed attempts to get the right answer would eventually lead to either escalating the conversation to a human agent, if possible, or the user terminating the conversation.

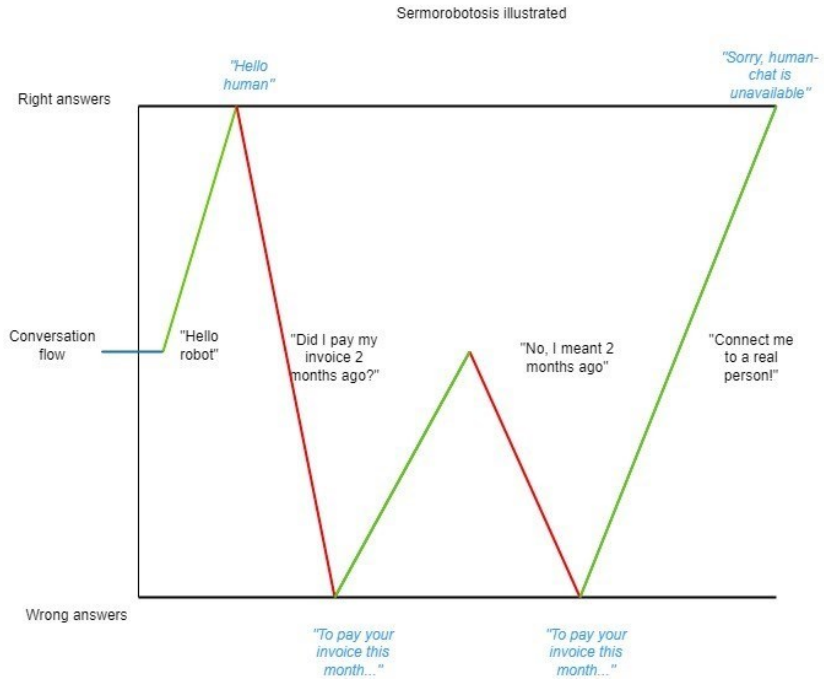

Figure 4: Sermorobotosis illustrated. Wrong answers are also meaningful ones with the right context.

Though the bot successfully answers to the best of its abilities, did exactly what it was supposed to do, and the caller may have even given indicators of terminating the conversation by for example saying, “never mind”, “thanks”, or “bye”, the conversation had reached a critical point from failing upwards and left the user’s matter unresolved.

The opposite also holds true; if the person inquiries about this month’s invoice, but the bot simply fails to understand the user or gives otherwise wrong or inaccurate responses, which is also meaningful within the right context, the conversation will reach a point requiring escalation to a human agent or the person terminating the conversation, likely without any thanks and their matter unresolved.

From a chatbot’s perspective, their objective is to keep interacting with users and providing them with (ideally) accurate responses to their queries until they indicate readiness to terminate the conversation. As an input-output application, the only thing a chatbot can do is try to speak when spoken to, even if it is unsure of what to say.

3.3.2 Rules of conversation

Unless a human interacting with a chatbot indicates that they are ready to end the conversation, the conversation likely should resume. The way a user might indicate readiness would be by thanking the robot or by saying farewell. If the user has not indicated readiness to end the conversation and it is terminated by the chatbot, this could be considered a point of failure and is often the result of an error or problem. However, this does not always hold true.

Bouzid and Ma state “conversations are highly structured interactions and observe a well-defined set of rules. Adherence to these rules is expected, while deviations from them become a source of meaning” (2022, 38). The source of meaning refers to the identifying of an error or problem. They also suggest that though rules of conversation apply to people conversing with one another, they do not necessarily apply towards chatbots, but are expected to be abided *by* chatbots. A few of their suggested conversation rules are:

1. Expectation to cooperate with each other.
2. Tell you the truth and only the truth.
3. Give as much information as reasonably expected.
4. Stay on topic.
5. Speak clearly.
6. Do not interrupt.
7. The goal is to be helpful and advance the exchange forward.
8. Avoid saying things that would obviously lead to wrong conclusions.

A human in a conversation with a chatbot always has the right to terminate the conversation at any point without indicating they are about to do so. However, a chatbot should never terminate a conversation without informing the user that they are ready to end it or have completed the user's request. It is noted, that in a majority of successful and failed conversations evaluated, users did not indicate intention to end conversation and exited without prompt. Therefore, a conversation prematurely terminated on behalf of the human user may not always be an accurate indicator of conversation failure.

Humans also have the right to violate any or all the rules of conversation at any point with a chatbot, though it may be counter-productive or pointless. While chatting with one, humans may lie or exaggerate, withhold information, change topics at random, or otherwise chat aimlessly or even disrespectfully. Chatbots should not violate any of the rules of conversation or disrespect humans. Doing so is likely to cause miscommunication, misinterpretation, distrust, or otherwise compromise the chances of a successful conversation and possibly a brand’s public image.

Bouzid and Ma suggest when designing a chatbot, one should remember even if the human user a chatbot is being designed for understands that the chatbot is a machine and doesn't mean to be disrespectful, if the chatbot behaves disrespectfully, would the human be slighted? If the answer is yes, then design some other behavior (2022, 52).

The rules of conversation do not provide accurate ways to measure chatbot performance but can be used as guidelines while designing conversations. Usage of these rules in this research is not a metric to measure, but their presence or absence will be considered during analysis.

3.4 Research limitations

Implemented processes may not be set up on all chatbots instances included in this research. Though the collective data sample of 375,100 conversations represents all the conversations included in this study, not all conversations are affected by the same tested activities or are performed over the same duration as activities were set up at different times during the research period.

All testing conducted is approved by the brands involved though are limited in theories and practices suggested in this text. The tested activities, primarily system responses, are specific to each brand's needs and provide valuable information that can be generalized or applied to other chatbots.

The data analyzed includes conversation history rated by multiple professionals, and though rating methods are agreeable and similar, slight variations and discrepancies are inevitable.

Though this poses the risk of obscuring data, it accurately reflects chatbot performance as measured by the same group of people.

Because conversation quality ratings vary greatly by brand and the individuals who perform the rating, the aggregate data studied may not yield entirely consistent information across the whole sample size. Sample sizes will vary to highlight the effects of any implemented activities still within the same research period.

4 Research implementation

All testing will be conducted in a production environment and requires communication, planning, and consent from the parties involved. Ideally any processes, practices, or other activities be started at the same time, however, due to time and schedule constraints research activities may not all start at the same time.

All chatbots are unique in the sense that they have all been developed by different brands and people and may present different requirements or prioritization of errors to address. Therefore, not all research activities may necessarily be implemented in all instances, but that too provides perspective and points of comparison.

4.1 The study subject

The client partners with many brands serving multiple industries; to narrow down the scope of this research, the data analyzed and applied against theories is based on chatbots in the finance sector.

Data analyzed will only be from May 1, 2022, to May 1, 2023, and evaluate primarily failed conversations as previously defined. The reviewed information will include filtered statistical information to measure conversation quality indicators and system responses before and after research activity implementations. To measure error handling mechanisms the primary metrics and strategies considered are:

- Conversation quality ratings, a main KPI the client measures overall success of conversations and chatbot quality.
- Amounts of conversations requiring escalation to humans.
- System responses, and different types and amounts of errors.
- Number of available intents at given periods.
- Current error handling strategies.

Privacy is of most importance, therefore no actual conversation details or specific brand names or strategies will be disclosed in this text. Brands may be compared to one another, and against applied theories as defined in this research, therefore claims and results in this text are specific generalizations and do not reflect any individual brand's business performance or strategies. Any portrayed conversations within this text are simplified reenactments of common chat scenarios identified during research. The client does not suggest its partners apply these research theories or practices currently, because they have yet to internally specify error handling or conversation design guidelines.

4.2 Identifying appropriate issues

Through cooperative efforts with the studied brands, commonly identified problem areas and general areas of development should be identified first. Common KPI's brands evaluate to measure their chatbot's performance generally all reflect how frequently the chatbot has given specific responses in specific situations. This includes the very first messages a chatbot

gives to its users, how it handles erroneous situations, and how it handles escalation to second-line support.

There is no predefined method for handling these types of situations, therefore default methods can be generalized as simple and generic, or advanced if it leverages additional information or processes to progress the conversation.

4.3 System responses

The system responses seen by many, if not most chatbot users, is the greeting message, as well as the system response during an unknown message. These responses will be evaluated to determine how they play a role in conversation flow. Escalation, or users asking for a human agent to join the conversation, is not a system response, but can be considered to be one for the purpose of this research.

4.3.1 Greetings

The chatbots' greeting will be updated to teach it page-awareness. If for example, a user begins a conversation on a bank's website specifically about loans; the bot should be aware or prepared to assist them in matters related to loans. The greeting should then acknowledge to users that it can assist with specific, or ideally the most common, loan-related matters. This also means increasing the chatbot's visibility on the website it serves to sub-pages, if possible.

The chatbot should provide a short message to convey this awareness, provide relevant buttons to the topic, or even a mix of both. The client of this study encourages *conversational* solutions; however, each brand may decide the appropriate ratio of text to buttons in their chatbot strategies. Whether the opening message is reliant on clickable interactions or text only, the stated awareness to the user and its possible effect on conversation success is what is being evaluated.

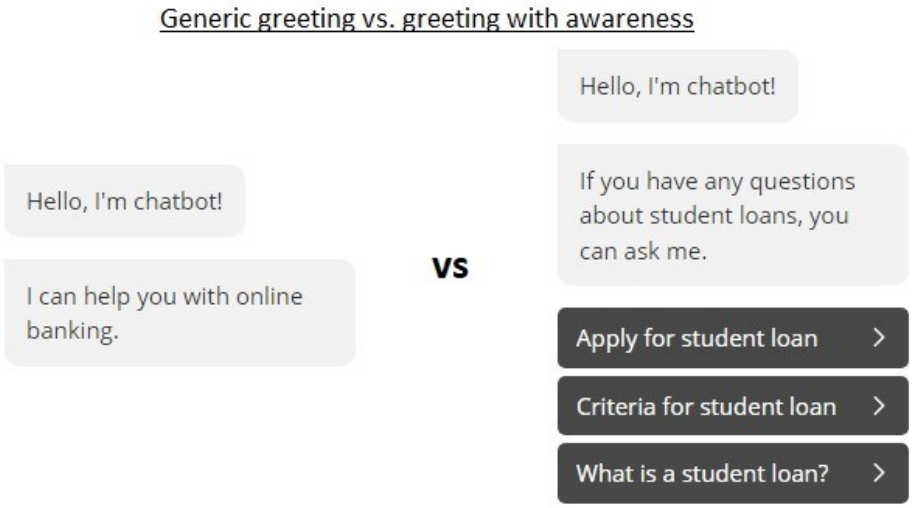

Figure 5: A comparison between greetings with and without page awareness and buttons.

4.3.2 Unknown messages

The chatbot's method of handling erroneous messages, specifically unknown messages, should be noted. A generic attempt to recover the conversation is by having the chatbot simply state a variant of "I didn't understand, can you please rephrase?". This is often a platform default but can be reconfigured. Much like the opening message being specific to users' location on a banking website, a chatbot's ability to manage unknown situations should also include awareness of a user's location on the site or previously discussed topics when things go wrong.

If no solid information can be utilized to handle the situation, the chatbot should rely on a generic approach, but offer ways to get the conversation either started or back on track. If buttons are to be listed about what the chatbot can help with, they should be prioritized according to the most common topics users ask about.

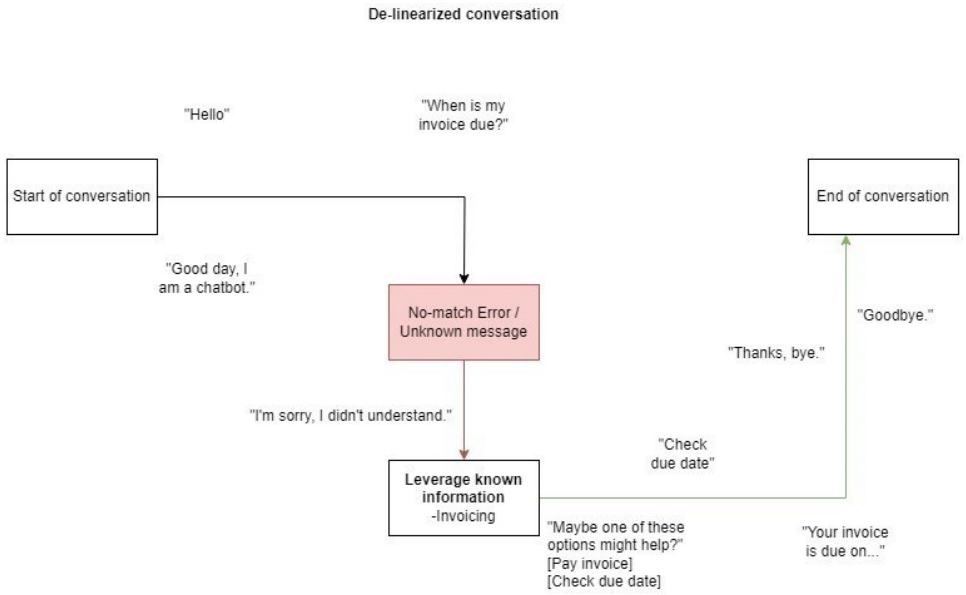

Figure 6: A de-linearized conversation with an active attempt at restoring flow.

4.3.3 Escalation handling

One should make note of what happens when users ask for a human agent. Does the chatbot immediately comply, or does it try to deflect, or is human escalation even possible? If escalation is not possible, the chatbot should clearly state this to users and try suggesting relevant topics that the user may be inquiring about. This may be challenging, because it is already identified that users often do not state their matter clearly before asking for human assistance.

If human escalation is possible, and especially if escalation is not a regularly desired outcome, the chatbot should try to deflect at least once. This can be done by acknowledging the user's request but still suggesting conversing with the chatbot, or by other means in attempt to dissuade the user from wanting their conversation escalated. If any information is available, such as a user's location on a website, it should be leveraged to decrease chances of escalation.

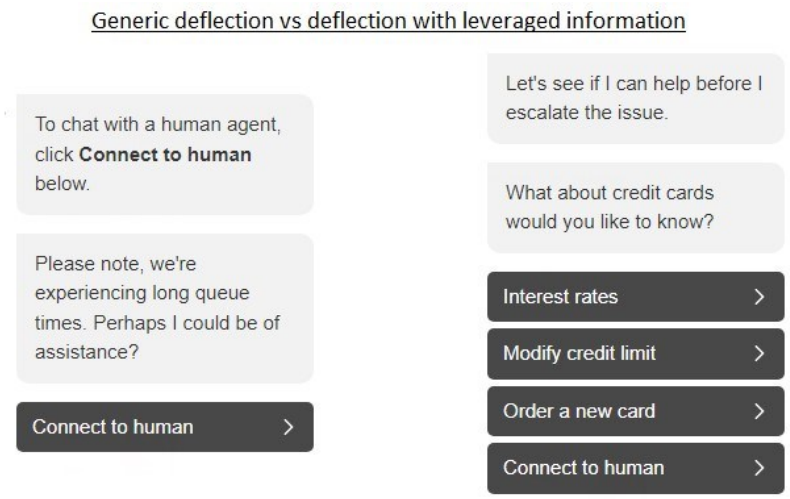

Figure 7: Deflection attempts to try to retain the conversation with the chatbot.

5 Findings and Analysis

Developers can design a bot to perform the tasks that they think end-users will want to be able to perform through the transactions. Though there is likely already basis for the necessities and scope of the chatbot, the only way to test viability is to let real people use it and break it.

Software testing is a great way to find and fix problems before moving an application into production. Though clients perform rigorous testing of chatbots during development, actual conversations with actual people provide the most valuable information about a chatbot's performance.

5.1 Comparing results with objective

The results of this study show that creating a chatbot's dialogue to be more engaging and interactive with its users results in increased activity and engagement. Though this is a desirable result, it also increases the chances of failure. This makes it difficult to accurately measure KPI's when conversation success rates should theoretically be increasing, but conversation amounts are too, including the failed ones. Increased conversation amounts may also inadvertently put strain on the development teams' ability to review and improve conversations.

5.1.1 Results: Greetings

Changing the chatbots opening messages showed positive change in user interactions. Three months prior to adding multiple button options in the greeting messages, users clicked an average of 17,290 buttons per month. After buttons were added, the following three months showed button interactions increased to an average of 46,005 per month. Success rates, however, could not be deduced from that change alone. Users clicking buttons reduced the amount of written input required, which in turn reduced chances of unknown messages, thus reducing the chance of something going wrong. As previously theorized, each message from a user is potential to fail.

Adding buttons to the chatbot's greeting was a strategy to start the conversation by offering clear conversational paths right from the start and increase engagement with the chatbot. Proactively this was successful in initially guiding the start of a conversation but did little to address situations *after* an error had occurred. The increased visibility of the chatbot on a brand's website increased the amount of monthly conversations it had with human users by 31.2%, which may reduce the load from other more costly communication channels. Though conversation amounts increased, so did the potential for error and failed conversations, though it cannot be ruled out as the single cause of reduced successful conversations.

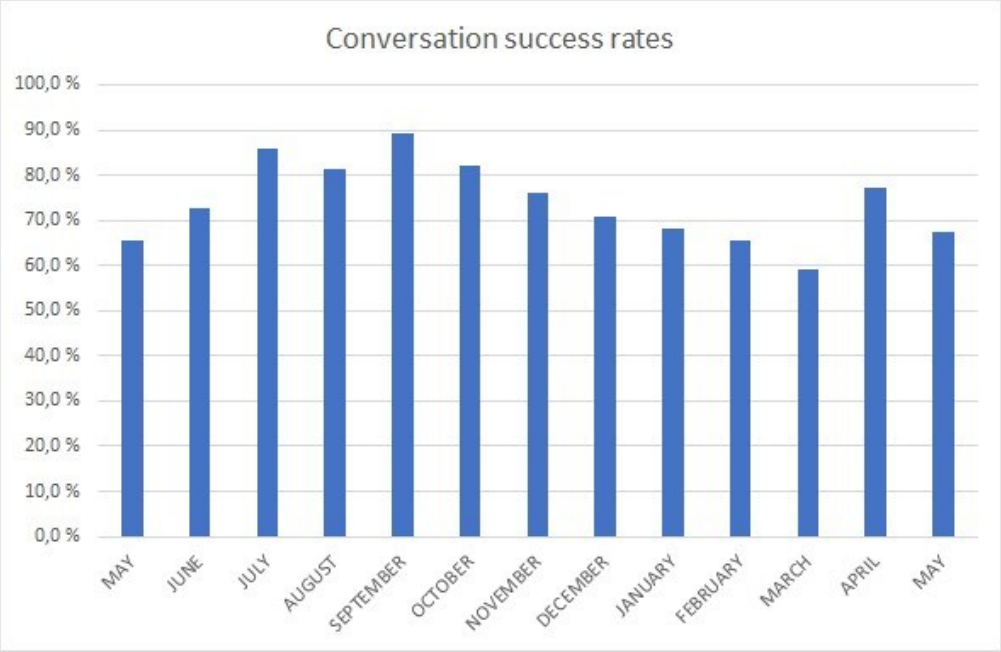

Figure 8: Increased visibility and awareness in September increased conversation amounts but reduced successful conversations.

Instead of generic greetings regardless of users' location on a website, tailoring the greeting in accordance with the page users were on when they initiated the conversation resulted in users' messages being more specific. Though the chatbot was able to greet users based on its location on a website, it was later identified that the chatbot lacked contextual awareness without users first interacting with the chatbot. The chatbot essentially pretended to know where it was, without really knowing. This issue was resolved by rephrasing answers and adding more context specific buttons.

However, it was found that users seem to prefer using buttons to interact with chatbots. When provided with multiple buttons to choose from during the opening messages, users frequently clicked the provided buttons. Though a desired outcome, it was noticed that there were less written messages at the start of the conversation. This resulted in users providing less context or specific input regarding their matter and made it harder to interpret the reason for their using the chatbot. When provided with only a few buttons, users seemed unsure of what to say, or how to chat with the chatbot. When provided with more button options, users relied on them heavily and said even less.

Development, including training others, and test deployment of a process with updated greeting messages with page-awareness took less than two hours. Deploying the process should take a mere 30 minutes, but the testing process prior to go-live may take weeks due to brand strategies and priorities.

5.1.2 Results: Unknown messages

Even though unknown messages are notoriously common with chatbots, the common approach seemed to be expanding the chatbots knowledge, or the number of intents. This too, is an effective way of handling unknown messages, though an indirect one. Teaching the bot more subjects is a way to prevent specific unknown messages from occurring.

The processes necessary to build advanced methods of handling unknown messages were restrictive due to existing methods clients had unrelated to error handling. It would have resulted in any amount of downtime, which was not acceptable at the time of research. Though the purpose and feasibility of advanced methods was clear, it was not in the immediate scope of chatbot development teams to implement theoretical and client unsupported methods of handling unknown situations. Though counter-intuitive to not research or develop a tailored solution to a common problem, it was deemed safer to have a generic response rather than risk causing a problem that could affect a significant number of users.

Even though unknown system responses were not revamped to include advanced methods, two buttons were added to the response after asking users to rephrase: “What can I help with” and “How does chat work?”. This small addition resulted in increased user interactions as well as an increase in successful conversations. Though the increased success rate could not be easily calculated from that single change, it was observed that in over 10% of conversations with updated unknown messages users were seemingly able to resolve their matter without escalation or frustratingly exiting the conversation.

5.1.3 Results: Escalations

After implementing processes in September 2022 to include page awareness, increase chatbot visibility, and offer topic relevant button options, conversation amounts increased greatly. However, this also resulted in the chatbot temporarily being able to handle less conversations autonomously, meaning a temporary rise in escalations. Chatbot visibility was reduced shortly after due to excessive amounts of conversations which strained resources and the ability to review enough conversations in a timely manner.

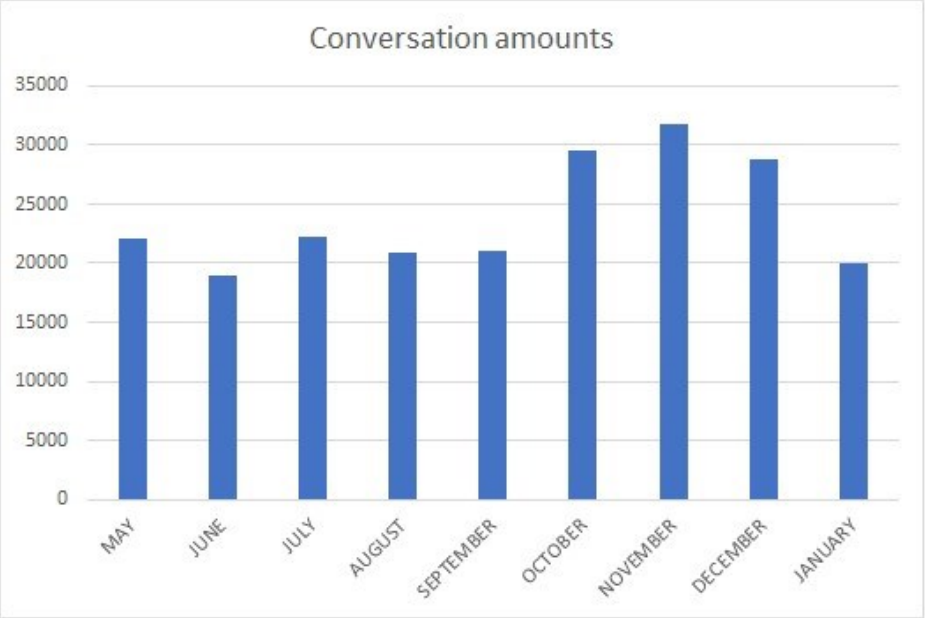

Figure 9: Chatbot visibility increased conversation amounts after September.

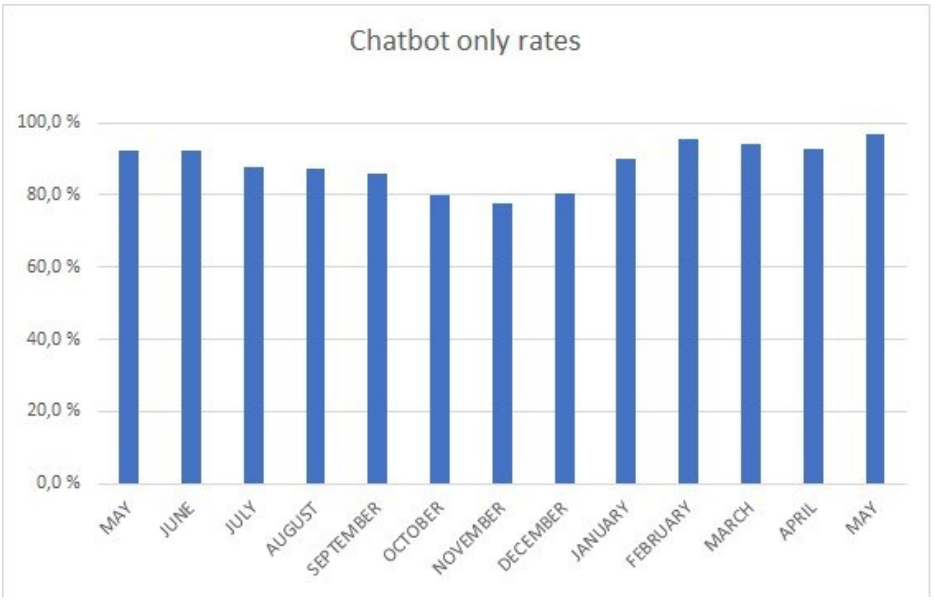

Figure 10: Conversations rates involving only the chatbot.

With a process designed to try to deflect users from being escalated to human agents, findings suggest that not all users are willing to "try again" and insist being transferred to a human agent. It was also noted that many users who insisted on escalation also expressed their dislike of chatbots during their short conversation with one.

5.2 Erroneous by design

Causes of error situations are almost always inadvertently designed into the bot. One can only teach a bot to understand as much as people do themselves, more specifically, the small group of people developing the bots. Regardless of how well a chatbot is designed, users seem to find a way to push it to a breaking point. Whether or not, and even *how* a chatbot admits this can influence the rest of the conversation.

The most common reason for failed predictions is that the chatbot simply did not understand what was said - unknown messages. Statistics from the data pool showed that 11-15% of conversations contained unknown messages. It was observed that the common way to handle unknown messages is by having the chatbot simply apologize and ask the user to rephrase their question. Though an appropriate strategy, it is a generic attempt to recover from a failing conversation. Before adding the two buttons "What can I help with" and "How to use chatbot" being added, the only other button options provided to users was one to display brand contact information to call or visit them. No other restorative measures were utilized to handle these situations.

Though polite, chatbots do not necessarily have to apologize or even admit mistakes. Bouzid and Ma claim that “a chatbot that apologizes a lot is a chatbot that fails a lot” (2022, 96). They state that chatbots should apologize when they fail but should be quick to correct the failures. Alternatively, the chatbot could use affirmative language to encourage progression or build trust with its user.

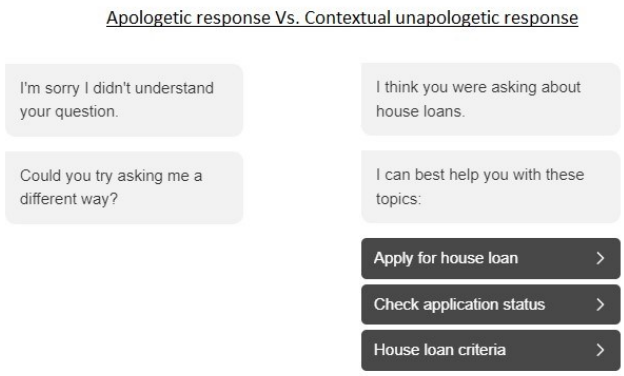

Figure 11: Instead of apologizing, affirmative language may be effective.

5.2.1 Uncanny Valley

It is logical to assume that making a chatbot seem like a human would make it easier to talk to, but making it sound too human can make it more difficult. Chatbots, by default, do not understand very many expressions, idioms, and wordplays that are otherwise very common to many people. Should a human user not realize they are talking to a bot and use an unfamiliar expression, it easily steers the conversation off course and possibly confuses the user.

Chatbots should avoid trying to sound too human, or otherwise talk like a human and use sarcasm or idioms. Overly excited or human responses such as *"Hi! What’s your name? Oh, and where are you from? My name is Bender, could I have that girder?"* may either sound bad, be bad in taste, or give a false sense of security. It can also give a feeling of uncertainty, much like talking with a robot that looks realistically human; you can tell that something is off, but you can’t quite tell what it is, but it makes you uneasy. This is called the uncanny valley, it suggests that human-like behaviors or appearances can make an artificial figure seem more familiar and human to people, but only up to a certain point. The familiarity drops sharply once the figure tries and fails to realistically mimic a human (Hsu 2012). Oddly, familiarity increases after the plunge indicating true human-likeness or a healthy human person.

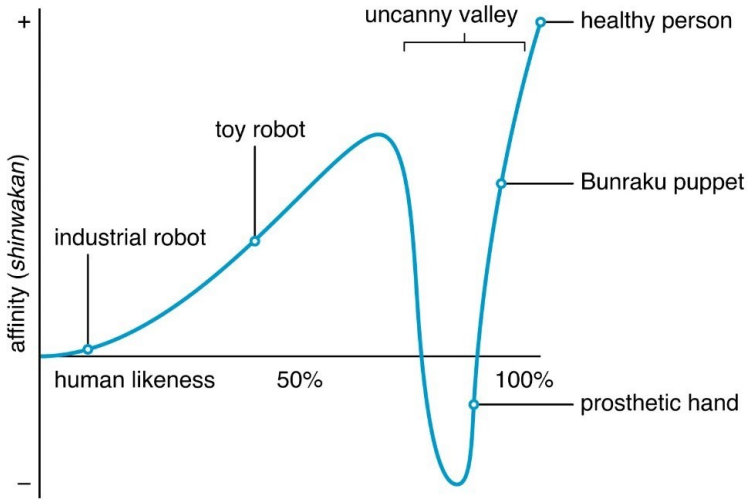

Figure 12: The Uncanny valley, by roboticist Masahiro Mori in 1970 (Kendall 2022)

Voicebots too are prone to the uncanny valley. When it is clear to users that they are speaking with a robot, its limitations are implicitly being told to the human. A chatbot also should avoid speaking in an overly robotic fashion, for example by saying, *"Say your name. What is your country? I am bender please insert girder"*.

It is important to find a good balance but err on the side of robotic; better safe than sorry. A balanced version of the previous examples could be *"Hello, please state your name. Where are you calling from? I'll need a girder to continue"*.

5.2.2 Familiarity may cause confusion

Fluently speaking voicebots try to imitate human speech, which inadvertently creates unrealistic expectations about their abilities. The more human-like the bot's behavior, the higher the expectations rise (Marttinen 2022, 107). Marttinen references a Cassell case where they identified that when speaking to a robot, when it would acknowledge a user's presence in a human-like way, the human would speak faster and less clearly to the bot which resulted in poorer performance. The best performance was achieved by using clear and limited commands with the bot.

The chatbots evaluated during this research all explicitly exclaimed that they were bots, or virtual helpers, in their opening messages. They have been deployed for a number of years, which has created familiarity amongst users, however, some users still were not immediately

aware they were chatting with a chatbot. It was noticed that, primarily during the initial release of a chatbot onto a service channel, users were less aware they were conversing with a robot.

Bouzid and Ma (2022) suggest not using a phone ringing sound to connect a human to a voicebot. A ringtone invokes a conditioned response, indicating to a person that another human is about to pick up on the other end. By connecting a human caller to a voicebot with a ringing sound, it sets the false pretense and prepares them to speak as if they were about to talk to another human. When the call is answered, the caller may:

- Speak to it as if it were a human.
- Not realize they are talking to a robot.
- Realize they are talking to a robot.

Speaking to the chatbot as if it was a human may immediately lead to errors. For example, a person calling and starting dialog with non-sensical terms or phrases: *"Oh, yeah, hi uhh, I was wondering what I should do if my card is acting up."* Though the chatbot may identify the word *card*, the rest of the user's statement does nothing to explain the situation or is otherwise helpful to making a prediction to an intent.

Not realizing they are chatting with a bot can also lead to a person providing too much input that the chatbot cannot process all at once, if at all. On top of data overload, users may develop expectations that the perceived human counterpart should recognize context, figures or numbers, or other vital information.

Conversely, a person may also not provide enough information for chatbots to make a prediction, though a human counterpart would likely have understood. This may be peoples' names, geographical locations, abbreviations, or common expressions and other remarks.

If users realize only afterwards that they are talking to a robot, they may be stunned, upset, or confused because they expected a person to respond. By letting them know as soon as possible that the answering party is a machine, the caller will probably be more willing to cooperate instead of demanding a person immediately, as they may have been expecting.

The usage of a ringing sound is possibly a built-in mistake which leads to the whole conversation starting off on the wrong foot, especially if the voicebot also sounded too human. Instead, avoid using a ringing sound and have the bot answer before any ringing begins. If that's not possible, then the answer should begin with something to immediately indicate that it is not a person. This can be a sound, a chime, or slogan, which immediately tells the caller that the answer was automatic.

Bouzid and Ma suggest when a person enters a situation where they expect to engage in conversation, they mentally prepare themselves and choose a “mode” to communicate in. When they speak with another human, they use a “human mode” and generally have positive expectations, and should they speak to a robot - they use a “robot mode” and are more likely to expect a negative, or less positive experience (2022, 119).

When in robot mode and speaking with a very robotic robot, human-like features are more noticeable usually in a positive manner because people have the tendency to like other humans. However, if a robot is too human-like, people tend to operate in “human mode” with the bot, and any robotic nuances are highly noticeable and promote negativity.

This also applies to text-based chatbots; if the chatbots message to the user convinces them they are chatting with a human counterpart, they may face the same complications as voicebots.

5.3 Personality and style

Chatbots ought to have a consistent personality. The way that a chatbot communicates with users is important in creating a meaningful and successful conversation. For voicebots, this is true about which gender the bot sounds like.

According to research by Mitchell, Ho, Patel & MacDorman in 2011, they found that when comparing *human vs. machine* speech, women and men agreed a human voice to be more pleasant. In a *female vs. male* speech test, results showed that both women and men agreed that a female voice is more agreeable and pleasant to hear. A male voice seems to be preferrable when seeking affirmation, authority, or answers - whereas a female voice is preferred when seeking information, assistance, or clarity, and is easier to distinguish in a pretentious situation (Marttinen 2020, 109).

There are exceptions however, the gender of the voice should be appropriate to the service being provided. In the late 1990’s, it was widely speculated that German automaker BMW recast the voice of the female-voiced GPS navigation system installed in their vehicles because of the high number of complaints from men that they didn’t want to take driving instructions from a woman, despite knowing that the GPS voice was synthesized (Marttinen 2020, 110). Moral judgements aside, one could speculate that people (men) preferred an authoritative voice telling them what to do, contrary to earlier findings that a female voice is preferential.

However, this bias is not only true for female voices. Marttinen (2020, 110) explains, typically (voice) commands work best when being given by a Caucasian male because the commands are heard clearer. However, the underlying problems; dialects, accents, or other

ways of speaking that are not native to the typical American Caucasian male, or otherwise local target audience, can easily cause confusion or misunderstanding with a chatbot. Part of the problem simply arises from the fact that most machine learning material has been collected from Caucasian males. A diverse chatbot development team with a good understanding of the target audience can improve communication with users.

Chatbots have many different uses, one should consider what type of tone or voice to use when designing a bot. In text-based bots, the tone may not be as apparent, but certain sentences or phrases may carry connotations which may conflict with the bot's intended personality. Sarcasm especially rarely translates well over text, much less from software.

The chatbots included in this study had formal styles of communicating and consistent personality in their responses. Some of the clients chatbots introduced themselves with human-like names not immediately suggesting they were a bot, while others' names included the word "bot" to highlight the obvious. Despite these naming conventions, it seemed that not all human users noticed this naming technique and were not immediately aware they were chatting with a chatbot.

6 Discussion and solutions

All brands in this study appropriately utilized a chatbot on their website as a first-line support channel. Also, they all utilized automated for their phone support. Both are effective ways of handling a large amount of incoming contact requests.

Specifically with their online chatbots, though some approaches were rather generic it was not necessarily due to their disinterest in improving services, but also to their resource limitations. Software and chatbot developers alike, generally do not have the freedom to build and develop anything they want without proper justification and approval from multiple stakeholders. Though brands may recognize a gap in their chatbots' knowledge or have identified a common root problem, the implementation of a solution is not always straightforward.

Even with a well thought out system in place to handle error situations, human error is difficult to predict and plan for. Because chatbots are developed by humans, those same human errors may inadvertently be programmed into the chatbot itself. That's why constant testing and implementation of fail-safes and new methods are important.

6.1 Restorative system responses

Error situations and unknown messages should be handled in a manner that tries to restore a conversation back into a linear flow to progress through the conversation. When a conversation deviates from its designed flow, it is at a critical point and any available information should be used to try and understand the user's perspective. If for example, a user was on a website about loans and produced an unknown message, the chatbot could leverage that knowledge to formulate a response likely relevant to the user's issue.

Bouzid and Ma suggest establishing safety points to keep track of where in the conversation something went wrong and returning to that point to try again (2022). This happens naturally with human-to-human speech; if one loses track of what the other person was saying, then they can think back to what was said earlier. Though each chatbot platform operates differently, safety points may be set or identified by referring to the constant data being generated by a conversation at any given time. This includes utilizing any platform specific features or definable variables.

6.2 Prevention is key

As previously mentioned, the best way to find out if a chatbot is viable is to let real users test it and break it. Identifying problems with chatbots is the first step in finding its solution.

Though the focus is how to handle *failing* conversations, perhaps the most important steps are to first identify problems at their root causes. Only then is it logical to try to fix problems at their source to prevent them from ever occurring.

A poorly started conversation is likely to lead to trouble later in the conversation. Chatbots should be specific in stating their abilities, at least to some extent, rather than broadly offering assistance. The brands included in this study tested how changing a chatbot's greetings would affect conversations. Though not conclusive enough to accurately say that it directly reduced error situations, it did yield in increased interaction with the chatbot.

Chatbots should be designed to have awareness regarding where it is communicating with its users, for example a specific page of a website. If awareness cannot be implemented, providing options to the user can also be an effective way to initiate conversation. Interactive voice response (IVR) used in automated telephony systems often practice this by listing menu options during a call.

Though IVR's function differently and do not necessarily have awareness, the numbers they can be reached at already set a pretense. Chatbots should try to utilize available information about the user's session to predict what matters they may be likely to ask about.

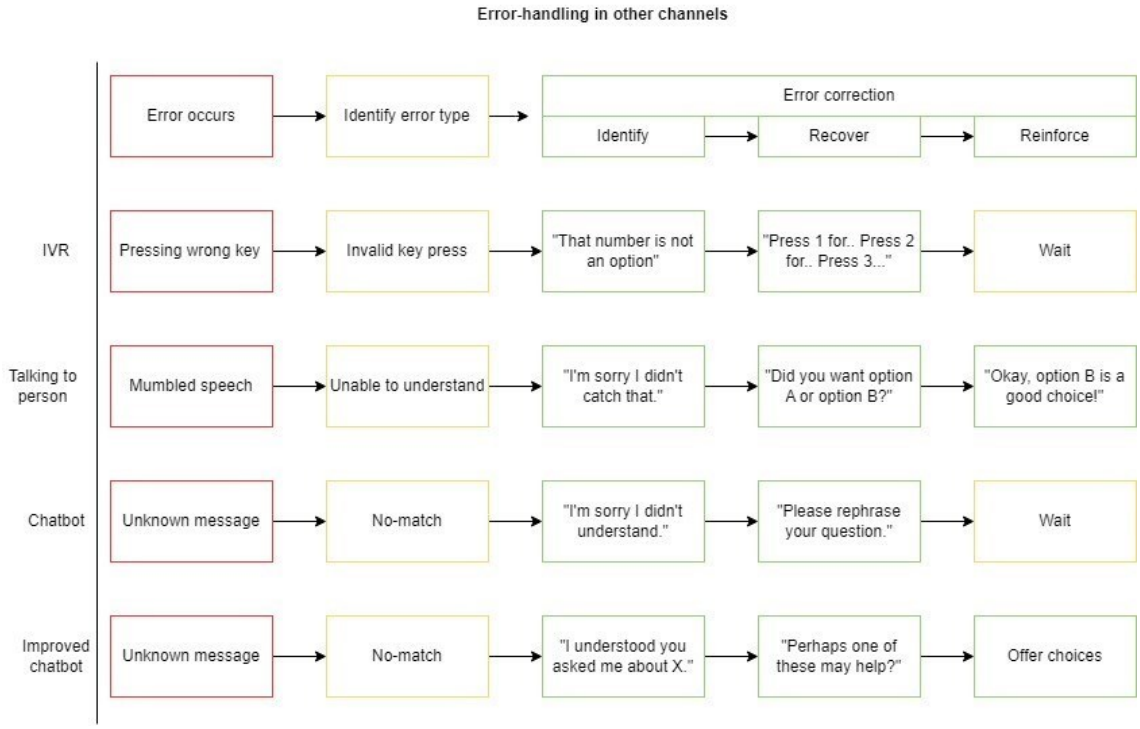

Figure 13:Error handling in other familiar scenarios.

6.2.1 Build for failure

Conversational AIs are sometimes seen as entities and referred to as if they were a person. In truth, they are programmed applications with lots of possible outputs. Though teaching a chatbot to be human-like can have a positive result when done in moderation, it should also be developed to be like any other transactional software application and include common programming logic such as If-statements.

If-statements evaluate an expression and return a result of "this or that". Though not always limited to two options, they serve the purpose of making a decision based on the input. Chatbots, and other AI, are essentially a compilation of very complex If-statements constantly evaluating the input. With that logic, building more If-statements, or teaching the bot new intents and system responses support error handling processes.

6.2.2 Keep it short

The conversations should be quick, clear, and concise. Human callers are calling because they have a matter to resolve and likely do not want to call potentially costly service numbers only to listen to a voicebots monologue. Though the goal of the voicebot is to interact with the user for as long as necessary; from the voicebots perspective, the user is the problem, and it doesn't 'want' to talk to humans either.

Excessive conversation from either humans or voicebots only increases the likelihood of something eventually going wrong. The bot may not be able to understand increasingly new amounts of input or information, and the caller may end up giving too much or too little information to the voicebot. Either situation may result in unfavorable conversation conditions.

6.3 Proactive conversation design

Identifying potential flaws ahead of time to prevent them from occurring is preemptive and proactive. That holds true for chatbot conversation as well. Though a well-trained model should already be able to handle frequent topics users call or message about and in the many ways they can inquire about them, most mistakes that happen with technology are often a result of human error. A voicebot's error too is likely to be caused by human error - by poor design.

Giving callers important information first will reduce chances of them having to try to aurally dig around for it. Important information may not be what the chatbot owner deems important, but rather what is important to the person using the chatbot. For example, a bank might think it important for users to know card prices, but the user is looking for information about investments for their children.

Bouzid and Ma suggest one way to be proactive would be to include notifications to the human user if there is new information available. Notifications work in two parts, the notification itself, and the information content (2022, 140). This has been seen in voicemail systems, *"You have 1 new message. First message, ..."* This also applies to voicebots, if a user were to call about invoices, the voicebot could proactively tell them if they had an overdue balance.

Such information would be highly relevant and likely even the reason why they are calling. However, providing unsolicited, though still important information may produce an adverse effect. Notification urgency should be considered if notifications are to be given. If the

caller’s invoice was not yet overdue, it may not be the reason they are calling and informing them of it may lead their intended conversation off-course.

Each message from a caller has the potential to fail. By reducing the amount of things a caller has to say to a voicebot, the chance of failure is also reduced. Depending on the conversation's design, proactivity and also inactivity, a caller may not have to say much at all for the entire call. The less expectations and interactions required of the caller, the less chance for a failed conversation, but telling the user too much information may yield the same result.

Here it is important to leverage known information. Bouzid and Ma suggest that instead of waiting for the user to tell the voicebot what they want to do next, try to be proactive and tell them what they likely want to know and what they might ask about next (2022, 139). For example, when ordering food through a chatbot, assuming you have a personal account and have ordered from there before. When talking with the chatbot after logging in, it would have access to seven contexts to help navigate through conversation. The chatbot, or developers, could assume that you are hungry (emotional state), calling a restaurant (user category), have a preferred language and payment information set (preferences), what you’ve recently ordered (history), what you’re trying to do - order food, or what step of ordering you are in (activity), and your default delivery address and how long it will take to receive your order (location and orientation). By using context alone, the bot could proactively offer the human caller exactly what they wanted without them having to say much of anything.

6.4 Increased scope

A chatbot is primarily developed to do things, rather than compensate when it *cannot* do something. Though it is important to prepare for the latter, teaching a chatbot new things is a core part of training AI. By teaching chatbots more material for its intended purpose, the amount of mishaps can be reduced.

Figure 14 below shows that as the number of intents increased for a chatbot, the amount of unknown messages reduced. Though other factors may influence the reduction, the main cause was the chatbot simply “knowing more”.

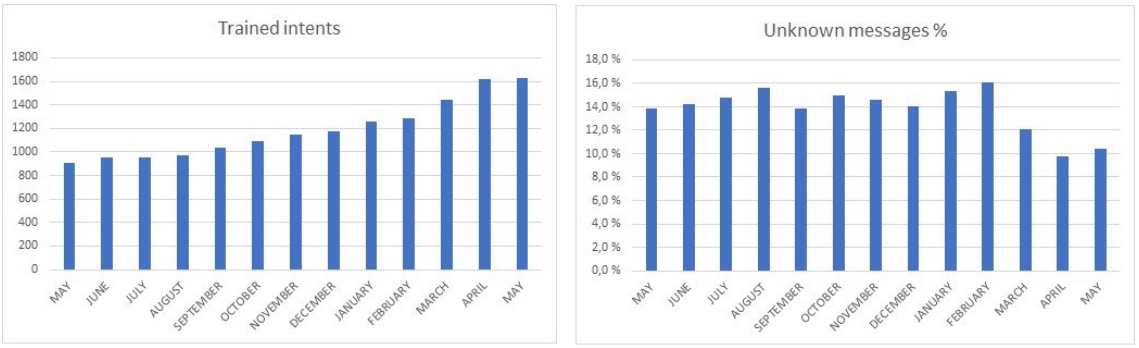

Figure 14: More intents contribute to reduced unknown messages.

Though it may not be viable to try to teach a chatbot everything about a brand, or to answer every identified question asked by users, there are methods to increase a chatbot's knowledge. One of those methods would be to teach it the opposite of what it already knows.

6.4.1 Over-education theories

Chatbots are helpful in explaining how to make something happen, but they could also be taught to assist in situations where something has happened that should not have. As mentioned earlier, usually when someone reaches out to a brand it is because they want to gain information, conduct an action, or because something has happened, and they want to troubleshoot a situation.

The author's *inverse action theory* states that for every transaction that a chatbot knows how to help perform, it should also know how to help with the opposite of it. A simple example, if the bot can explain how to apply for a card, it should also know how to help terminate a card.

The *double inverse action theory* states that for every transaction that a bot knows how to help perform, it should also know how to help with the opposite of it, and when performing those activities is not working out. For example, if it can help apply for a card, it should be able to help terminate a card, as well as with what to do when applying for a card or terminating a card is not working out.

Original transactions	1st tier transactions	2nd tier transactions
Open an account Get a card	Open an account Close an account Get a card Close a card	Open an account Close an account Unable to open account Unable to close account Get a card Close a card Unable to get a card Unable to close a card

Figure 15: Over-educating a chatbot with inverse action theories.

These theories suggest methods of over-educating a chatbot, though may not necessarily raise the utility of the chatbot to teach the inverses of every situation. By applying the inverse action theory, a chatbot's knowledge could theoretically be doubled. By applying the double inverse action theory, a chatbot's knowledge could theoretically be quadrupled.

In doing this, the bot would be prepared for all of the other outcomes of situations that it is already able to assist with, reducing the likeliness of unknown messages. Original transactions also serve to establish useful safety points that can be returned or referred to in the event of an error.

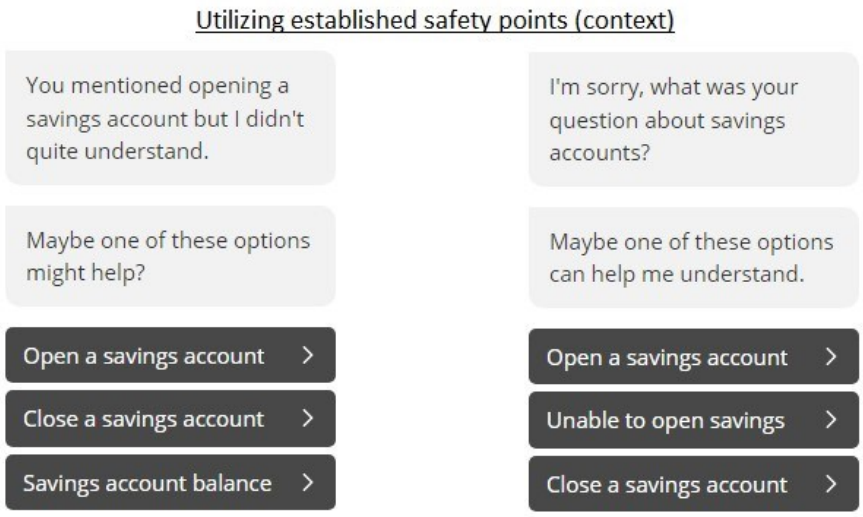

Figure 16: Utilizing context for leverage.

6.5 Human error

A very common type of failure revolves around the concept of chatbots not being able to interpret users' requests. Based on end-user feedback, the most frequent complaints are that the bot does not know anything and does not know how to help. From the chatbots perspective, most errors happen because the human does not know how to communicate with it. Booth claims that the problem of Human-Computer Interaction (HCI) is an issue of matching system functionality to the needs of the user in a specific context while presenting a response that can be easily understood (1991).

This can be resolved by trying to teach users how to communicate with bots successfully. This may be particularly useful if utilized in specific error types, or even as trained knowledge. Chatbots could offer users options for it to explain what it can help with or what kind of sentences or words it best understands.

Though it was noticed that some users were surprised to unknowingly be chatting with a robot, the ones that knew they were chatting with one perhaps had too high of expectations and lowered their input efforts preemptively. Users frequently sent chatbots single word messages as opening messages and seemed disgruntled when it was unable to understand what they wanted. If a person were to physically visit a service desk and engage with other humans using such minimal language, it would likely cause equal and awkward confusion – yet chatbots seem to be expected to understand.

6.5.1 Improper channel

Regardless of how well a voicebot is designed, its key players must always be considered. Though there are several stakeholders in the entire development process since conception, it is the end-users that the bot is built for - yet they are rarely included in any part of the development process. A chatbot could be designed very well, but forgetting the target audience might make it the least effective way to help users. A voicebot meant to help those hard of hearing, or not enough language options for a diverse target audience would likely have lots of problems, as would a text-based chatbot designed for people unable to read or even see.

A voicebot is not useful for long events such as cooking instructions, booking a flight, or checking movie times; all activities with a lot of steps or information the user would have to retain. Text-based chatbots have the advantage of previous messages being visible to the user, but with voicebots there is no visual user interface (VUI). They work better for quick information and short answers.

Again, preventing problems is key to handling them. Bouzid and Ma suggest chatbots should be clear in what they can assist with, or what is in its scope of abilities. The bot may offer this information explicitly, or upon the user asking about its abilities. The bot's message to users about its purpose should be specific. Rather than "I'm your kitchen mate", something more specific would be "I can help you find recipes" (2022, 120). Bouzid and Ma claim chatbots that explicitly state their scope have higher success rates.

7 Conclusion

One size does not fit all; this applies to the deployment of a chatbot, both voice and text based alike. How a chatbot performs during an erroneous situation is a matter that cannot easily be defined but is certainly a very important aspect to consider. With that, not every brand will share the same strategies as others, nor would they likely want their chatbots to look, sound, and feel like each other's. Different industries have different definitions of successful conversations and chatbot utility overall. Others may use them as a deflection method, while other brands find more utility in providing self-service features to their customers.

Out of the 375,100 conversations in this study, a high average of 15.9%, or 59,640 conversations, contained unknown messages. If 25% of those conversations resulted in the user calling the brand, that would be 14,910 telephone calls. Using the previously identified

€2.60-€6 cost per call industry standard, that would equate to €37,275 - €89,460 per year. With proper error handling mechanisms and deflection, if escalation were reduced by 10% to 15%, that would reduce the costs to €14,910 – €53,676 annually.

Seeing as unknown messages are a commonly identified problem, it should be a higher priority to manage these situations better. Even if not all conversations can be recovered after encountering an error, all potentially failing conversations would still at least be attempted to be recovered in an orderly fashion.

Though much of a chatbot's improvement methods rely on applied theory and should yield satisfactory results, actual development of software changes in a live environment with a team is a challenging task. In tailoring such custom solutions, it creates a lot of "moving parts" that ought to be properly documented for future developers, as well as following proper project management practices.

Newer chatbots such as OpenAI's ChatGPT utilize large language model (LLM) algorithms to produce human-like text. Paul and David (2023) explain that LLM's are trained on large amounts of text-based data, generally entire web page or brand content, to not produce results based on keywords, but from the actual meaning of users' prompts.

Though LLM based chatbots may be a strong contender in the industry, they too come with shortcomings and must be trained by humans to ensure accuracy.

This study was not fully conclusive due to several factors: mixed rating methods due to human difference and brand standards, implementations of processes at different times, and different implementations on different study targets. In theory, all objectives are attainable through the applied practices, however, better scheduling and narrowed criteria would yield more accurate results. Realistically, despite the potential cost savings from investing in proper error-handling practices, the initial investment cost to implement them may be off-putting to clients and customers.

- Each chatbot has unique strengths and weaknesses, though they all share a common issue – human error, from both chatbot users and developers. Because humans are the ones designing chatbots, the platforms they operate on, and the conversation flows; they are inevitably going to program human error into the chatbots. As humans are the root cause of chatbot problems, it could be speculated that a chatbot designed by an AI would be nearly impervious to human error.

References

Printed

Bouzid, A.& Ma, W. 2022. The Elements of Voice First Style: A Practical Guide to Voice User Interface Design. 1st edition. California: O'Reilly Media Inc.

Benyon, D. 2019. Designing user experience: A guide to HCI, UX and interaction design. Fourth edition. Harlow: Pearson Education Limited.

Marttinen, J. 2020. Robofobia: Mikä roboteissa ja tekoälyssä pelottaa? Viro: Tallinna Raamatutrukikoda.

Electronic

Boost.ai 2023. What is conversational AI, anyways? Boost.ai, 2 April. Accessed 1.5.2023. https://www.boost.ai/knowledge/what-is-conversational-ai

Booth, P. 1991. Errors and theory in human-computer interaction, ScienceDirect, December. Accessed 3.5.2023. https://www.sciencedirect.com/science/article/abs/pii/000169189190005K

C.David & J. Paul 2023. ChatGPT and large language models: what's the risk? National Cyber Security Centre, 14 March. Accessed 9.5.2023. https://www.ncsc.gov.uk/blog-post/chatgptand-large-language-models-whats-the-risk

George, T. 2023. What is Action Research? Definition & Examples, Scribbr, 21 April. Accessed: 19.5.2023. https://www.scribbr.com/methodology/action-research/#:~:text=Action%20research%20is%20a%20research,by%20MIT%20professor%20Kurt%20 Lewin.

Hsu, J. 2012. Why "Uncanny Valley" Human Looks-Alikes Put Us on Edge, 3 April. Accessed 18.12.2022. https://www.scientificamerican.com/article/why-uncanny-valley-human-lookalikes-put-us-on-edge/

Kendall, E. 2022. Uncanny valley. Encyclopedia Britannica, 8 Nov. Accessed 4.5.2023. https://www.britannica.com/topic/uncanny-valley.

LiveAgent. Cost per call. Accessed 2.5.2023. https://www.liveagent.com/customer-supportglossary/cost-per-call/

Mitchell, W., Ho, C., Patel, H. & MacDorman, K. 2011. Does social desirability bias favor humans? Explicit-implicit evaluations of synthesized speech support a new HCI model of impression management. USA: Computers in Human Behavior. Accessed 14.12.2022. http://www.macdorman.com/kfm/writings/pubs/Mitchell2010DoesSocialDesirabilityBiasFavor Humans.pdf

Oracle. What is Natural Language Processing? Accessed 12.1.2023. https://www.oracle.com/hk/artificial-intelligence/what-is-natural-language-processing/

Rouse, M. 2017. Error handling, 1 May. Technopedia. Accessed 9.5.2023. https://www.techopedia.com/definition/16626/error-handling

Zabój, D. 2022. All You Need to Know to Use Chatbots in Business. Complete Guide, 31 August. Accessed 3.12.2022. https://www.chatbot.com/blog/chatbot-guide/

Zhu, H. 2005. Software Design Methodology: From Principles to Architectural Styles, 22 March. Elsevier Science & Technology. ProQuest Ebook Central. Accessed 9.5.2023. https://ebookcentral.proquest.com/lib/laurea/detail.action?docID=269543

Zitek, N. ITIL Incident Management – How to separate roles at different support levels. Advisera. Accessed 21.4.2023. https://advisera.com/20000academy/knowledgebase/itilincident-management-separate-roles-different-support-levels/