

A Conversational Breakdown Detector for a Motivational Interviewing Conversational Agent

Zhen (Luther) Qin

Abstract

A conversational breakdown in human-chatbot interaction refers to a disruption or failure in the communicative flow between the human user and the chatbot. To recover a disrupted conversation, the first step is to detect the breakdown. Researchers have proposed methods using supervised learning and semi-supervised learning in dialogue systems to achieve the goal of detecting conversational breakdown. However, few studies have focused on detecting breakdowns in automated therapeutic conversations, especially conversations led by motivational interviewing chatbots. The presence of conversational breakdowns has negative impacts on the human-chatbot interaction, such as frustration, dissatisfaction, or loss of trust. This gap suggests a need to build a robust and efficient conversational breakdown detector that recognizes interruptions during the conversation. Conversational breakdown detection paves the way for further action to recover conversations. In this paper, I develop a novel, unifying framework called “CIMIC” for characterizing the conversational breakdowns of “MIBot,” a motivational interviewing conversational agent for smoking cessation. I collect 200 pieces of conversational data through Prolific and annotate them using the CIMIC framework with a group of four trained annotators. The annotated dataset is then applied as the training set to fine-tune GPT-3 models to build a conversational breakdown detector for the MIBot.

Keywords

natural language processing, human-computer interaction, fine-tuning large language models, conversational breakdown, artificial intelligence, conversational agent

1. Introduction

Clinician counselling is a common treatment to help people quit smoking. Motivational Interviewing (MI) is one of the therapeutic counselling methods proven to help with behavioural change, and it has been shown to be effective in smoking cessation (Heckman et al., 2010). Considering that the current availability of psychotherapists and psychiatrists cannot meet the high demand for smoking cessation support (Fricchione et al., 2012), designing an automated dialogue system that can apply the motivational interviewing method is a feasible way to help people with positive behavioural change and to expand the availability of therapy resources (Shingleton & Palfai, 2016).

The “MIBot,” a project led by researchers from the University of Toronto and the Centre for Addiction and Mental Health (CAMH), brings an accessible intervention to help with smoking cessation. I was a joint research student in this project. MIBot is a web-based chatbot that conducts real-time MI conversations with users. Its goal is to build users’ motivation to engage in positive behavioural changes that lead to smoking cessation. During the experiments with human participants in the MIBot MIV4.4 version, researchers found conversational breakdowns (see section 4) that resulted in negative impacts on human-chatbot interaction, such as frustration, dissatisfaction, and loss of trust. These results increase the burden of analyzing the conversational data and prevent the MIBot from reaching its goal. However, it is impractical to manually review breakdowns in all conversational data due to the large number of transcripts. To address such a challenge, this paper aims to redefine conversational breakdown analysis within the MIBot context by introducing a conceptual framework called CIMIC (Conversational breakdown In Motivational Interviewing Chatbot) to characterize breakdowns in the conversation between the MIBot and its users. This paper also describes the development of a conversational breakdown detector with a fine-tuned large language model (GPT-3). In the future, we will be able to use more advanced models, such as GPT-3.5 and GPT-4, to develop detectors with higher accuracy and apply them to more chatbots beyond the MIBot.

2. Background

2.1 Smoking Cessation

Smoking is a crucial public health issue that comes with significant health risks and is a leading cause of preventable death (Centers for Disease Control and Prevention, 2020). The number of smoking-related deaths has ris-

en to 7.2 million a year, which is more than the deaths caused by AIDS, malaria, and tuberculosis combined (World Health Organization, 2023). Moreover, smoking is a common contributing factor to various chronic conditions, such as emphysema or chronic bronchitis (Centers for Disease Control and Prevention, 2020). In 2016, the Canadian government reported multiple diseases and conditions that are smoking-related, including cancers, respiratory problems, and heart disease (Health Canada, 2016). Although people have realized the negative effects of smoking, it is challenging for smokers to quit. According to the Centers for Disease Control and Prevention, in 2018, over 25 million adult smokers attempted to quit, but only 7.5% of them were successful (Centers for Disease Control and Prevention, 2020).

Common treatments to help smokers quit include counselling that involves behavioural interventions and pharmacotherapies. Behavioural interventions involve strategies to change behaviour through techniques like counselling or lifestyle modifications, while pharmacotherapies refer to treatment using medications. There is evidence showing that the combination of counselling and pharmacotherapies can have the most significant positive effect on smoking cessation (US Department of Health and Human Services, 2020).

2.2 Motivational Interviewing

Motivational interviewing (MI) is a counselling method developed by clinical psychologists William R. Miller and Stephen Rollnick in the 1980s (Miller & Rose, 2009). MI is a common approach in clinical psychology that addresses the ambivalence and resistance that patients express when moving to a positive behavioural change. In their textbook, Miller and Rollnick (2013) state:

MI is a collaborative, goal-oriented style of communication with particular attention to the language of change. It is designed to strengthen personal motivation for and commitment to a specific goal by eliciting and exploring the person's own reasons for change within an atmosphere of acceptance and compassion. (p. 29)

Motivational interviewing has been validated as an effective approach for behaviour change in clinical trials (Burke et al., 2003). Research led by Carolyn J. Heckman, Brian L. Egleston, and Makary T. Hofmann (2010) has shown that MI approaches can be successful tools for smoking cessation. One of the primary objectives of MI is to help the patient overcome their ambivalence toward behaviour change since a state of ambivalence can lead to a lack of motivation. Instead of providing solutions and advice directly, MI practitioners engage in a well-structured conversation wherein patients can reflect on the

underlying causes of their behaviour and become empowered to resolve their ambivalence and work toward positive behavioural changes.

2.3 Chatbot for Motivational Interviewing

Recently, chatbots have gained public attention due to the popularization of OpenAI's products, such as GPT-3, GPT-4, and ChatGPT. A chatbot is a computer software application that is designed to simulate human conversation through textual or auditory interactions with human users, especially online. Chatbots are often artificial intelligence (AI) systems that can understand human language by utilizing natural language processing (NLP) technologies and can generate human-like conversations. Chatbots can be used in various fields, including education, business, and health care (Adamopoulou & Moussiades, 2020).

In recent years, with the development of AI and NLP, there is a growing body of research on motivational interviewing chatbots. Chatbots for MI are designed to simulate human MI practitioners, using motivational interviewing techniques to engage patients and produce positive transformations. Research has shown MI chatbots can be used to treat various conditions, including smoking, cannabis abuse (Becker et al., 2014), or severe stress (Park et al., 2019). These studies illustrate a promising future for MI chatbots to help patients achieve positive behavioural changes.

3. Related Research

3.1 MIBot and Previous Work

The majority of the project discussed in this paper is built on the MIBot project. I joined this project in 2021 as a research student. Prior to the MIBot project, Almusharraf et al. (2020) designed an MI-based chatbot to motivate smokers to quit. In their study, they improved the quality of motivational interviewing responses generated by the chatbot from the preexisting model. Then, they evaluated the impacts on participants by collecting feedback after a completed conversation. They recruited 121 participants through an online platform and received 6568 responses in total. Based on these responses, they generated 21 reasons for or against continuing smoking, with the frequency of each reason. The results show that the chatbot's accuracy in reflection and summary is increased after training it with the responses and smoking reasons.

The MIBot builds on Almusharraf et al.'s (2020) study. It is a cross-disciplinary project that involves researchers from the University of Toronto and the CAMH. It develops a therapeutic chatbot and uses the GPT-2XL model to

generate responses that follow motivational interviewing principles (Brown et al., 2023). These auto-generated responses aim to provide reflections on participants' answers to five questions regarding smoking habits. This conversation data was used to build a conversational breakdown detector, which is the focus of this paper.

3.2 Conversational Breakdown

Conversational breakdowns can have negative impacts on human-chatbot interaction. Conversational breakdown detection is therefore a crucial task in human-chatbot interaction, and several studies have proposed methods for detection—see subsection 3.2.2. This literature review summarizes the current state of research on conversational breakdown in human-chatbot interaction, focusing on the definition of conversational breakdown and detection methods.

3.2.1 Definition of Conversational Breakdown

Breakdowns during human-chatbot interaction prevent both sides from communicating effectively and have a negative impact on the user's perception of usability, transparency, and trust (Ashktorab et al., 2019). Most previous research literature considers conversational breakdowns to be caused by the chatbot, removing any blame from the user. However, by inspecting the conversational data collected from the MIBot project, we find that users can also be the source of conversational breakdowns.

Failure in a chatbot interaction can be annoying, and many users may give up the task when breakdowns occur. When a breakdown happens, it means the chatbot does not understand the user's intention in the conversation (Khurana et al., 2021). This lack of understanding will cause confusion for the users as they do not know what to do next. Khurana et al. (2021) developed a chatbot interface called Chatrex to reduce conversational breakdown.

Kawamoto et al. (2020) consider conversational breakdown to be a situation in which a conversation is not established successfully. To be more precise, it is a situation where the chatbot does not reply appropriately, causing users to repeat, give up, become silent, or be confused about the sentence. The authors applied fast Fourier transform analysis (FFT) to differentiate conversational breakdown from normal, successful conversation. They inspected users' body language and found their potential relationship with conversational breakdown.

Similarly, Ng et al. (2020) define conversational breakdown as a situation where the agent is prevented from completing the desired goal of the

dialogue due to either user exasperation or agent misunderstanding. In their work, they investigate various methods to improve conversational breakdown detection, such as continued pre-training and a manifold-based data augmentation method. They demonstrate the effectiveness of these methods on the Dialogue Breakdown Detection Challenge English shared task.

In all the literature mentioned above, the scope of conversational breakdown includes the chatbot's failure to recognize the user's intent, leading to inappropriate responses, confusion, and frustration. While the literature presented above tends to only discuss conversational breakdowns as being caused by the chatbot, in reality, users can cause breakdowns as well. In the transcript dataset collected from the MIBot project, we examined and analyzed each data entry to explore who and what causes the conversational breakdown, and we developed a conceptual framework for characterizing conversational breakdowns—see section 4.

3.2.2 Detection Methods

Several studies have proposed methods to detect conversational breakdown in human-chatbot interaction. Khurana et al. (2021) propose a design of an explainable Chatbot Interface, ChatrEx, which explains to the user the lower-level work of a chatbot when a conversational breakdown happens, enhancing the user's perception of usefulness, transparency, and trust in the chatbot. However, the work of Khurana et al. (2021) primarily focuses on interface design; it does not provide an algorithmic solution for detecting conversational breakdowns in the MIBot context. Ng et al. (2020) consider conversational breakdown detection as a text classification task and propose two semi-supervised learning methods to enhance the dialogue breakdown detector. In their work, they mention two stages of the dominant paradigm in NLP. The first is pre-training a large language model on a general corpus, such as Wikipedia or Common Crawl. The next step is adding a classification layer and fine-tuning the model on new data so this model can be used in downstream tasks. They choose BERT, a machine learning model for NLP, as their baseline, then apply the continued pre-training on the Reddit dataset and use Self-Supervised Manifold-Based Data Augmentation (SSMBA) methods to improve the baseline model. Their model achieved an accuracy of 73.92%, beat the baseline (58.66%), and won first place in the 2020 Dialogue Breakdown Detection Challenge 5 (Ng et al., 2020, p. 4). Their result shows that we can consider conversational breakdown detection as a binary text classification task (i.e., have a breakdown/no breakdown). For conversations between human users and chatbots, however, conversational breakdown caused by users is possi-

ble and should not be neglected. In this work, we introduce a computational method to (1) consider both user-caused and chatbot-caused breakdowns, and (2) detect breakdowns in text-based, asynchronous conversations.

4. Framework of Conversational Breakdown Definition

To understand conversational breakdown in textual, asynchronous conversations between users and the MIBot, the research team developed CIMIC (Conversational breakdown In Motivational Interviewing Chatbot), a novel conceptual framework for redefining conversational breakdown. Collaborating with experts from the CAMH, we adapt current conversational breakdown definitions (see 3.2.1) and extend them to fit the MIBot context. The CIMIC framework has three states presenting an intuitive outlook of conversational breakdown (CB) types: No CB, Type I CB, and Type II CB. In the previous section, related literature emphasizes that breakdowns are caused by the chatbot. However, in the conversational data gathered from MIV4.4 (described in 5.1), we find multiple user-caused breakdowns exist. We consider the breakdown caused by the chatbot as a Type I breakdown and the breakdown caused by the user as a Type II breakdown.

Table 1 contains a labelling scheme for conversational breakdowns, with example sentences. In the first row, the user gives an answer that does not match the intention of the question, which asks about a reason for enjoying smoking. In this situation, we consider it a conversational breakdown caused by the user. In the “Type” column, 0 represents a user-caused breakdown, while 1 represents a chatbot-caused breakdown. Thus, we label this example with 1 in column CB (there is a breakdown) and 0 in column Type (it is a user-caused breakdown). In the second row, the user’s answer matches the question, and the reflection generated by the MIBot addresses the user’s answer. We consider it to have no conversational breakdown, and label this paragraph with 0 in column CB and “NULL” in the Type column. In the third row, the MIBot gives a response that is semantically irrelevant to the user’s answer, which is considered a conversational breakdown caused by the MIBot. In this situation, we label the paragraph with 1 in the CB column and 1 in the Type column. In the fourth row, we can see intuitively that the conversational flow is perfect, and there is no breakdown. Therefore, we label it with 0 in the CB column and “NULL” in the Type column. Finally, in the fifth column, the MIBot’s response implies that the user does not consider there to be a difference between smoking and vaping. This response is unrelated to the user’s answer about not wanting to quit smoking at all. This example is a breakdown caused by the MIBot, so we

label 1 in the CB column and 1 in the Type column.

Table 1. Labelling Scheme for Conversational Breakdowns

Prompt	CB	Type
To start, what is the thing you like most about smoking? It stresses me out. Smoking makes you anxious.	1	0
Now, what is the thing you like least about smoking? The health implications. You don't like the consequences of your smoking.	0	NULL
Now, what is one thing about your smoking habit that you would like to change? Can't think of any to be honest. You are aware that you will need to make changes in order to stop your smoking habit.	1	1
What will it look like when you have made this change in your smoking habit? I would spend less money. You would be in control of how much money you spent.	0	NULL
Finally, what are the steps you need to take to make this change? I need to take zero steps as the change in question does not exist. You don't think that there is any difference between smoking and vaping.	1	1

5. Data Collection

First, I describe the transcript data collected by the MIBot team through Prolific, an online platform that helps researchers recruit and manage participants for online research. I then explore conversational breakdowns in the MIV4.4 transcripts and discuss the manual annotation process for determining the types of breakdowns based on CIMIC.

5.1 Transcript Data

The MIBot is a conversational agent based on motivational interviewing rules. The conversation between a human user and the MIBot is constructed in a certain format. At the beginning of the conversation, the MIBot will present a short introduction about itself and ask the user for consent:

MIBot: Hello! My name is MIBot, I am a computer chatbot and part of a research project at the University of Toronto. I was made by a group of researchers and doctors, with the goal of exploring your smoking behaviours. I will ask a few questions, listen to your answers, and then respond. However, I am learning how to have this conversation. So, I may ask you if the things I say make sense. Would you like to chat with me about your smoking?

If the user consents, the MIBot will start the conversation by asking five questions. Otherwise, the MIBot will thank the user for their participation and terminate the dialogue session. For a complete conversational flow with one user, the MIBot asks the following five questions in chronological order:

- Q1. To start, what is the thing you like most about smoking?
- Q2. Now, what is the thing you like least about smoking?
- Q3. Now, what is one thing about your smoking habit that you would like to change?
- Q4. What will it look like when you have made this change in your smoking habit?
- Q5. Finally, what are the steps you need to take to make this change?

After the user provides an answer to one question, the MIBot generates a reflection on the user's answer based on the Motivational Interviewing principles. Here is one real example of a complete conversational flow, structured in a Question-Answer-Reflection format, from the MIV4.4 version of MIBot:

MIBot (Question): To start, what is the thing you like most about smoking?

User (Answer): The heavy taste of smoking and makes me feel calmer.

MIBot (Reflection): Smoking makes you feel good.

MIBot: Now, what is the thing you like least about smoking?

User: Leaves scent on hair and clothes.

MIBot: You're concerned that people will notice your use of tobacco products.

MIBot: Now, what is one thing about your smoking habit that you would like to change?

User: When I smoke, I smoke too many cigars.

MIBot: You want to cut down on the amount of cigarettes that you smoke.

MIBot: What will it look like when you have made this change in your smoking habit?

User: I think I will prevent future health problems.

MIBot: You are worried about the effects on your health.

MIBot: Finally, what are the steps you need to take to make this change?

User: Learn self-control or seek that calm that gives me with something else.

MIBot: You want a way to be calmer when you're stressed out.

All the conversation is collected from the Prolific platform, and all results are completely anonymous. I collected 200 sets of conversations (i.e., 200 rows of text data in a .csv file) with about 40 participants from the MIV4.4 version of MIBot to train the conversational breakdown detector. For each transcript, I merge each Question-Answer-Reflection conversation set into one short paragraph (Figure 1), which serves as the entry to be labelled manually for conversational breakdowns.

Figure 1. Merge the Conversation.

*MIBot: To start, what is the thing you like most about smoking?
User: The heavy taste of smoking and makes me feel calmer.
MIBot: Smoking makes you feel good.*

Before

↓↓↓↓

After

To start, what is the thing you like most about smoking? The heavy taste of smoking and makes me feel calmer. Smoking makes you feel good.

5.2 Annotating CB

I recruited 4 researchers (Andrew Brown, Angus Wang, Leon Zhu, and Vidya Sujaya) as crowd workers and trained them on how to use the CIMIC framework to identify conversational breakdowns. In the annotation process, the trained crowd workers were given a transcript entry (one text paragraph synthesized of MIBot's question, user's answer, and MIBot's response) and asked to label the types of conversational breakdown as identified in CIMIC. For each entry, crowd workers annotated whether the transcript contained a Type I, Type II, or no breakdown. Each crowd worker annotated 50 transcript entries.

To focus on the conversational breakdown detection task rather than CB type identification, we simplify the annotated dataset by creating a column with only binary entries to identify conversational breakdown. We use 0 to represent no breakdown, and 1 to indicate that there is a conversational breakdown.

6. Model

I propose a fine-tuned model based on GPT-3 Davinci for conversational breakdown detection. The GPT-3 Davinci model is pre-trained by OpenAI, and I fine-tuned it with our simplified annotated dataset to build the conversational breakdown detector.

6.1 Experimental Setup

I split the dataset into training and validation sets (80:20). Then, I fine-tune the GPT-3 Davinci model and keep all hyperparameters as default since the default hyperparameters work well across a range of use cases.

6.2 Evaluation Metrics

I use the 200 annotated transcript data to train and test the fine-tuned GPT-3 Davinci model. To evaluate the performance of the CB detector, I calculate and compare the recall value and accuracy of each model. I chose these two metrics because accuracy measures the overall performance of a classification model, while recall measures the ability of the model to identify instances of a particular class correctly. In machine learning, accuracy and recall value are two important metrics used to evaluate the performance of a classification model (Barkved, 2022). Accuracy is the proportion of correctly classified instances out of the total number of instances in the dataset. In this case, true positive represents the result that the model predicts the positive class correctly, where “true” represents the instance flagged as a conversational breakdown by a crowd worker. Mathematically, it is expressed as:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN}$$

where TP (true positives) is the number of correctly predicted positive instances, TN (true negatives) is the number of correctly predicted negative instances, FP (false positives) is the number of incorrectly predicted positive instances, and FN (false negatives) is the number of incorrectly predicted negative instances.

Accuracy measures the overall performance of the model and is often used as the primary metric for evaluating classification models. In our case, it is the metric of how accurate each model is in recognizing the conversational breakdown. Recall value, on the other hand, is the proportion of true positive instances out of all positive instances in the dataset. Mathematically, it is ex-

pressed as:

$$Recall = \frac{TP}{TP+FN}$$

Recall is a metric that is particularly useful when the goal is to identify all instances of a particular class, even if that means sacrificing some accuracy. In the conversational breakdown detection task, it is more significant to correctly identify all instances of a breakdown (even if it means some false positives) than to achieve high overall accuracy.

7. Results

In this section, I analyze our model's performance in detecting conversational breakdowns compared to other models.

7.1 Overall Results

Here, I construct a new framework to complement the gap in the current conversational breakdown research by including the user-caused breakdown. I present how to identify different types of breakdowns in the framework. I then use the framework as the guideline to annotate conversational data to train a model to detect conversational breakdown under the MIBot context. Additionally, I present the performance attained through our suggested approach.

7.2 Baselines

Here are the three baselines used as reference points to compare the performance of GPT-3 Davinci:

- GPT-3 Ada
- GPT-3 Babbage
- GPT-3 Curie

Similar to GPT-3 Davinci, these baselines are base models that are pre-trained by OpenAI. I run the command from the Command Line Interface tool to train them with the same dataset and compute the classification metrics.

7.3 CB Detection Task

I fine-tuned these models by training them with the annotated data. Table 2 and Table 3 report the accuracy and recall values of each model. In Table 2, the accuracy of the fine-tuned Babbage model is slightly higher than

the fine-tuned Ada. The Babbage model has an accuracy of 65%, which means this model successfully predicts 26 data entries from the validation set. The fine-tuned Curie model has the same accuracy as the fine-tuned Ada model (62.5%), while the fine-tuned Davinci model has an accuracy of 67.5%, the highest of all four models at this stage.

Table 2. Accuracy for Each Model

Model	Recall Value
Fine-tuned Ada	0.625
Fine-tuned Babbage	0.65
Fine-tuned Curie	0.625
Fine-tuned Davinci	0.675

Table 3. Recall Values for Each Model

Model	Recall Value
Fine-tuned Ada	0.352941
Fine-tuned Babbage	0.411765
Fine-tuned Curie	0.294118
Fine-tuned Davinci	0.470588

As discussed in section 6.2, the recall value is another representative measurement of these fine-tuned models. Essentially, conversational breakdown detection is a binary classification task. Thus, a higher recall value indicates that the fine-tuned model better identifies both positive and negative examples. As shown in Table 3, the fine-tuned Davinci model has a recall value of 0.471, the highest among all four models. Meanwhile, the fine-tuned Curie model has the lowest recall value (0.294). Although the fine-tuned Curie model has the same accuracy as the fine-tuned Ada model, its recall value is lower than Ada's. Therefore, the fine-tuned Curie model does not perform well in identifying conversational breakdowns in our dataset. Among the baseline models, we obtain the best performance by fine-tuning the GPT-3 Davinci model.

8. Discussions

The result shows the fine-tuned Davinci model can be used as a conversational breakdown detector under the MIBot context. After fine-tuning the Davinci model, the accuracy reached 67.5% while the recall score reached 0.471. Both measurements beat the other models' results, and we consider the fine-tuned Davinci model to have an acceptable performance. Reflecting on the validation set, 27 of the 40 samples have been successfully predicted. However, although the Davinci model is the most powerful one, it does have a significant drawback—it is the most expensive of all four models. We have to pay \$0.02/1K tokens to fine-tune the model, while the cheapest and fastest Ada model only costs \$0.0004/1K tokens as of 2023 (OpenAI, 2023).

Due to data labelling being labour-intensive, we only prepared a limited number of transcript data. This has both positive and negative impacts. On the positive side, we save a lot of time in labelling these data and spend less money to train the models. On the negative side, the result is far from perfect. The fine-tuned Davinci model still has space to be optimized.

9. Conclusion

In this paper, I defined the concept of conversational breakdown under the MIBot context and discussed various types of breakdowns. Then, I described how we prepared and annotated the transcript data using crowd workers according to definitions of conversational breakdown and how we formatted the dataset for fine-tuning. We split the annotated dataset into the training set and validation set. In the fine-tuning process, we adapted four models (Ada, Babbage, Curie, and Davinci) that were pre-trained by OpenAI, fine-tuned them with the training set, and used them as conversational breakdown detectors. We evaluated these fine-tuned models on the validation set by comparing the accuracy and recall scores of each model. Then, we continued fine-tuning the four models with more epochs (epochs = 6) to improve the performance. Finally, the results indicated that the fine-tuned Davinci model had the highest accuracy and recall score among all four models, which means it had the best performance on detecting conversational breakdown in the MIBot transcript data.

9.1 Future Work

Future work can focus on adapting new pre-trained fine-tuning models to build better conversational breakdown detectors. New large language mod-

els, such as GPT-3.5 or GPT-4, may be able to generate better results in our classification task as more advanced models. GPT-3.5 and GPT-4 could have the potential to help us detect conversational breakdowns more efficiently. It could increase the effectiveness of the MIBot by helping in conversation repair and improving human-chatbot interaction so that the MIBot can better help users reconsider their smoking habits.

Future work could also focus on data augmentation. The transcript data we used in this research was limited, and increasing the training data could be a simple but effective method to optimize the fine-tuning models. One practical data augmentation technique is the back translation method (Beddiar et al., 2021): converting text data into a different language and subsequently translating it back to the original language. For example, we can translate the user's answers and MIBot's responses to French and translate them back to English, so we can generate new, distinct text data. Meanwhile, the context and meaning from the original transcript data remain unchanged. By using this method, we could generate more data for training and eventually optimize the performance of the fine-tuning models in detecting conversational breakdowns.

Last, recognizing different types of conversational breakdowns (either caused by the chatbot or the user) would be a valuable feature of future versions of MIBot. It could help us repair the conversational breakdown and guide the dialogue flow back to the correct track by selecting which repair technique to use under different breakdown situations. In this research, we discussed two types of conversational breakdowns: one caused by the user, and the other caused by the MIBot itself. There could be more breakdown types to identify in the future, and it would be challenging but important to design repair solutions that target different types of breakdowns.

9.2 Limitations

The primary limitation of this research was the scale of the dataset we used to fine-tune the models. Since our data are manually labelled, expanding the dataset would be labour-expensive and time-consuming. It takes time and effort to read through the conversation and use the definition of conversational breakdown to judge if the text contains a breakdown. Thus, it is almost unrealistic to produce a large-scale dataset to train the fine-tuning models. Meanwhile, since the MIBot team is still developing new features and adjusting the conversation structure, it would be challenging for us to collect textual data in the same format as MIV4.4 in the future. This could bring more difficulties in organizing and preparing the data.

9.3 Concluding Remarks

In conclusion, we are experiencing tremendous technological advancement in the field of AI, particularly in the natural language processing area. The emergence of large language models, such as GPT-3 and GPT-4, has revealed the potential of how powerful a language model could be. Software built on top of them has brought us a completely new level of human-chatbot interaction. The MIBot project is one of those successful cases of developing therapeutic chatbots to help people in need. The MIBot is a meaningful and valuable engineering innovation because it can help fill the gap caused by a shortage of well-trained human psychological therapists. Besides helping people who want to change their smoking behaviour, motivational interviewing chatbots can also be trained to help users with other addiction issues, such as alcohol or drug abuse. The fine-tuning functions of the large language model allow us to train chatbots for various tasks in the field of addiction and mental health: an exciting pathway for researchers to explore in the future.

References

- Adamopoulou, E., & Moussiades, L. (2020). An overview of chatbot technology. *IFIP Advances in Information and Communication Technology*, 373–383. https://doi.org/10.1007/978-3-030-49186-4_31
- Almusharraf, F., Rose, J., & Selby, P. (2020). Engaging unmotivated smokers to move toward quitting: Design of motivational interviewing–based chatbot through iterative interactions. *Journal of Medical Internet Research*, 22(11). <https://doi.org/10.2196/20251>
- Ashktorab, Z., Jain, M., Liao, Q. V., & Weisz, J. D. (2019). Resilient chatbots. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. <https://doi.org/10.1145/3290605.3300484>
- Barkved, K. (2022, March 9). *How to know if your machine learning model has good performance*. Obviously.ai. <https://www.obviously.ai/post/machine-learning-model-performance#:~:text=In%20fact%2C%20an%20accuracy%20measure,to%20understand%20what's%20going%20on.>
- Becker, J., Haug, S., Sullivan, R., & Schaub, M. P. (2014). Effectiveness of different Web-based interventions to prepare co-smokers of cigarettes and cannabis for double cessation: A three-arm randomized controlled trial. *Journal of medical Internet research*, 16(12). <https://doi.org/10.2196/jmir.3246>
- Beddiar, D. R., Jahan, M. S., & Oussalah, M. (2021). Data expansion us-

- ing back translation and paraphrasing for hate speech detection. *Online Social Networks and Media*, 24, 1–13. <https://doi.org/10.1016/j.osnem.2021.100153>
- Brown, A., Kumar, A. T., Melamed, O., Ahmed, I., Wang, A., Deza, A., Morcos, M., Zhu, L., Maslej, M., Minian, N., Sujaya, V., Wolff, J., Doggett, O., Iantorno, M., Ratto, M., Selby, P., & Rose, J. (2023). A motivational-interviewing chatbot with generative reflections for increasing readiness to quit among smokers: Iterative Development Study. *JMIR Mental Health*. <https://doi.org/10.2196/49132>
- Burke, B. L., Arkowitz, H., & Menchola, M. (2003). The efficacy of motivational interviewing: a meta-analysis of controlled clinical trials. *Journal of consulting and clinical psychology*, 71(5), 843–861. <https://doi.org/10.1037/0022-006X.71.5.843>
- Centers for Disease Control and Prevention. (2020). *Health effects*. Smoking & tobacco use. https://www.cdc.gov/tobacco/basic_information/health_effects/index.htm#:~:text=Smoking%20causes%20cancer%2C%20heart%20disease,includes%20emphysema%20and%20chronic%20bronchitis.
- OpenAI. (2023). *Deprecations*. Documentation. <https://platform.openai.com/docs/deprecations/>
- Fricchione, G. L., Borba, C. P. C., Alem, A., Shibre, T., Carney, J. R., & Henderson, D. C. (2012). Capacity building in global mental health: Professional training. *Harvard Review Of Psychiatry*, 20(1), 47–57. <https://doi.org/10.3109/10673229.2012.655211>
- Heckman, C. J., Egleston, B. L., & Hofmann, M. T. (2010). Efficacy of motivational interviewing for smoking cessation: A systematic review and meta-analysis. *Tobacco Control*, 19(5), 410–416. <https://doi.org/10.1136/tc.2009.033175>
- Kawamoto, A., Wada, K., & Saito, N. (2020). Development of method for detecting conversation breakdown using behavioral information. *2020 International Symposium on Community-Centric Systems (CcS)*, 1–6. <https://doi.org/10.1109/CcS49175.2020.9231508>
- Khurana, A., Alamzadeh, P., & Chilana, P. K. (2021). ChatrEx: Designing explainable chatbot interfaces for enhancing usefulness, transparency, and trust. *2021 IEEE Symposium on Visual Languages and Human-Centric Computing (VL/HCC)*. <https://doi.org/10.1109/VL/HCC51201.2021.9576440>
- Miller, W. R., & Rollnick, S. (2013). *Motivational interviewing: Helping people*

- change* (3rd ed.). Guilford Press.
- Miller, W. R., & Rose, G. S. (2009). Toward a theory of motivational interviewing. *American Psychologist*, 64(6), 527–537. <https://doi.org/10.1037/a0016830>
- Ng, N., Ghassemi, M., Thangarajan, N., Pan, J., & Guo, Q. (2020). *Improving dialogue breakdown detection with semi-supervised learning*. arXiv.org. Retrieved 30 August 2022, from <https://arxiv.org/abs/2011.00136>
- Park, S., Choi, J., Lee, S., Oh, C., Kim, C., La, S., Lee, J., & Suh, B. (2019). Designing a chatbot for a brief motivational interview on stress management: Qualitative case study. *Journal of Medical Internet Research*, 21(4). <https://doi.org/10.2196/12231>
- Shingleton, R. M., & Palfai, T. P. (2016). Technology-delivered adaptations of motivational interviewing for health-related behaviors: A systematic review of the current research. *Patient Education and Counseling*, 99(1), 17–35. <https://doi.org/10.1016/j.pec.2015.08.005>
- US Department of Health and Human Services. (2020). Interventions for smoking cessation and treatments for nicotine dependence. In *Smoking cessation: A report of the General Surgeon*. <https://www.ncbi.nlm.nih.gov/books/NBK555596/>
- World Health Organization. (2023). *Tobacco*. Health topics. https://www.who.int/europe/health-topics/tobacco#tab=tab_1

About the Author

Zhen (Luther) Qin, Master of Information Candidate
University of Toronto

Luther Qin is a second-year Master of Information student in the Human-Centered Data Science concentration. He has a Bachelor of Information degree from the University of Toronto. He is a joint research student in the Automated Medicine Lab at the Edward S. Rogers Sr. Department of Electrical & Computer Engineering. His research focuses on Natural Language Processing (NLP) applications and human-chatbot interaction, specifically using NLP techniques to solve problems in the interaction between human users and chatbots.

Email: z.qin@mail.utoronto.ca