

Final Project Report

Millie Wu
uni: mw3209

Zongyu Chen
uni: zc2657

Dae Lim Chung
uni: dc3666

December 2023

1 Context

1.1 The Problem

Emotion Recognition Conversation (ERC) has become increasingly important in Natural Language Processing, with applications in decision mining, recommender systems, and healthcare. However, advanced ERC tasks face challenges beyond basic emotion detection in short, context-free texts. These challenges include recognizing sudden shifts in emotion, detecting multi-party speech, emotion reasoning, and identifying sarcasm. Our dataset specifically targets these advanced aspects of ERC. Our project code can be found here: https://github.com/acnelexh/Chat_Simulation.

1.2 Limitations of Existing Datasets

Existing ERC datasets, such as EmotionLines [3], EmotionX [6], and MELD[9], often derive from daily conversations in popular sitcoms like "Friends." These datasets are limited in terms of cultural context and speaker personality diversity, focusing primarily on characters within an Eastern U.S. setting. Additionally, datasets like DailyDialog [5], which are designed for language learning, lack the authentic emotional dynamics found in close relationships. Even scripted dialogue scenarios in datasets like SEMAINE [7] and IEMOCAP[2], while addressing some of these issues, are not scalable due to the high costs of employing real actors.

1.3 Our Dataset's Approach and Uniqueness

Our dataset aims to address these gaps by improving the classification of emotion shifts, particularly drastic ones, and mirroring real-life scenarios more accurately. We propose a scalable approach using GPT-simulated dialogues, which allows for a more diverse and realistic representation of emotional dynamics in conversations.

2 Format

2.1 Inputs

The dataset comprises **79 simulated dialogues, each a text-only conversation with drastic emotion changes**. These dialogues are structured as 7-12 turn conversations between a bot and a simulated "human" agent, generated by large language models (LLM). To capture diverse emotional dynamics, the "human" agents exhibit drastic shifts from positive to negative valence emotions, or vice versa. These agents are designed with personas that approximate the demographics of the real-world population, addressing potential privacy and legal concerns that arise from using real human participants.

2.2 Outputs

In the realm of Emotion Recognition Conversation (ERC) research, emotion tags are broadly classified into **discrete and continuous types**. The journey of discrete emotion classification has its roots in ancient philosophy with Cicero's categorization of emotions into fear (metus), pain (argritudo), lust (libido), and pleasure (laetitia). This classification evolved over centuries, leading to the modern framework established by Paul Ekman, which identifies six basic emotions: happiness, sadness, fear, disgust, anger, and surprise. These emotions provide a comprehensive spectrum for analyzing human emotional states. Alongside these, continuous tags—valence and arousal—offer dimensions for assessing the positivity (ranging from positive to negative) and intensity (from strong to weak) of emotions. Our project delves into annotating the **two discrete emotions** demonstrated by the "human" agents, aligning with Ekman's six basic emotions.

3 Collection

3.1 Data Simulation

3.1.1 Persona generation

The simulated "human" persona comprised of 19 distinct vectors. These include age, gender, cultural background, occupation, education, family dynamics, relationship status, MBTI personality type, attachment style, emotional intelligence, typical mood, emotional range, stress triggers, coping strategies, significant past events, likes, goals and aspirations, and both physical and mental health. Of these 19 vectors, 9 are generated using specific functions that take into account global population distributions for realism, such as gender, cultural background, and occupation. The remaining 10 vectors are derived from purely random selections or assignments, ensuring a degree of unpredictability and variety in each profile. This blend of function-based and random generation methods creates a diverse and representative set of simulated individuals.

3.1.2 Label generation

The label generation process for our dataset involves creating a diverse array of emotion pairs to effectively label dialogues, drawing on Ekman’s six basic emotions plus a neutral category. This process categorizes emotions into positive (Surprise, Happy) and negative (Sad, Fear, Anger, Disgust) groups. The key aspect of this approach is to generate emotion pairs that represent drastic shifts, encompassing both positive to negative and negative to positive transitions. Although initially arousal levels were considered to add depth to these pairs, the complexity they introduced led to confusion in large language models (LLMs). Consequently, the focus shifted solely to these discrete emotion labels. By randomizing the order of these emotion pairs, the process ensures a varied and realistic set of labels that accurately capture the dynamic and often unpredictable nature of emotional shifts in human conversations.

3.1.3 Simulation Using LLM

To simulate the conversation, we leverage the OpenAI ChatGPT API [1] for dialogue generation, a choice that saves us from the complexities of training a large language model (LLM) from scratch.

Key to our approach is the design of prompts for effective conversation simulation between two agents using the GPT model. We customize these prompts to reflect the unique personas of each agent as a list of detailed instructions, thereby ensuring that the resulting dialogues are not only true to the agents’ profiles but also exhibit the coherence and natural flow of real conversations. To achieve this, we engage in extensive testing and refinement of various instructions, continually adjusting them until they meet our quality standards. Additionally, we included select examples from the DailyDialogue dataset into our prompts. The initial approach is detailed in Appendix B.

In addition to simulating the conversation as our primary data, we also prompt the GPT model to provide emotional annotations for each dialogue. Our aim is to compare the fidelity of these machine-generated annotations with human annotations in order to assess GPT’s understanding of human emotions and its ability to synthetically generate annotation. This is achieved by incorporating *[EMOTION]* tokens within the formatting prompts used when interacting with the ChatGPT API.

Generated examples from both models and their corresponding human annotation can be found at Appendix A.

3.2 Annotation

3.2.1 Instruction

The annotation instruction process for this project involves a structured approach to evaluating interactions between user personas and a chatbot. Annotators are tasked with analyzing three unique user personas, each engaging in about 16 dialogues with the chatbot. The primary focus of the annotation is on

emotion tracking. Each dialogue is anticipated to showcase one shift in emotion from the user. To accurately capture this, annotators are required to select two emotions per dialogue from Ekman’s six basic emotions: happiness, sadness, fear, disgust, anger, and surprise, using an emotion wheel as a reference tool for better accuracy and understanding.

Before diving into the dialogues, annotators must familiarize themselves with the user profiles to ensure contextually relevant emotion annotation. It’s important to note that while the chatbot consistently exhibits a neutral emotion, the user’s emotional state is more dynamic, alternating between neutral, a first chosen emotion, and a second one throughout the conversation. This contrast between the chatbot’s unchanging emotional state and the user’s shifting emotions is key to the annotation process. Annotators are encouraged to engage with this task with a blend of attentiveness and enjoyment, recognizing the importance of accurately capturing the intricate emotional nuances in human-chatbot interactions.

3.2.2 Process

We recruited a total of eight annotators, each of whom dedicated approximately 1.5 to 2 hours to the task. This resulted in a cumulative annotation time of between 12 to 16 hours. Annotators were assigned 3-4 personas each, with the number of dialogues per persona varying from 6 to 16, depending on the filter applied. Majority of time spent was likely on reading the dialogues with a smaller portion devoted to understanding the background of each persona. To facilitate a thorough cross-validation analysis, we utilized nine distinct personas and ensured that each persona was evaluated by at least three different annotators.

3.2.3 Cross Validation

Upon completion of the annotation process by all annotators, we employed majority voting to exclude data points with significant discrepancies in human-annotated labels as our cross validation. Each data point was assigned two labels: a starting emotion and an ending emotion. We individually assessed each emotion to ensure a majority agreement among annotators. If the annotations were highly dispersed for either the starting or ending emotion (e.g., if three annotators assigned three different starting emotions), we deemed the data point as noisy and discarded it. This process helped in maintaining the cleanliness of our data. However, this method of discarding data points led to the pruning of approximately 28% of the annotated data points.

3.3 Challenges

3.3.1 Persona Generation

When developing personas, our initial approach was to use real humans as a model, gathering details through Monkey Survey based on our personality vec-

tors. However, this method raised privacy concerns. Additionally, when we first generated personas using random outputs for likes, dislikes, goals, and aspirations, we encountered challenges in maintaining consistency within a single individual’s profile. For example, it was possible for an individual to simultaneously express a dislike for ‘art’ while having ‘artistic and creative pursuits’ as goals. Initially, we tried implementing exclusion rules to address these inconsistencies, but the rules quickly became overly complicated. To simplify the algorithm and maintain coherence, we ultimately decided to remove the ‘dislike’ variable.

3.3.2 Dialogue Generation

When initially utilizing GPT 3.5 to simulate dialogues, we faced several challenges. Firstly, the model exhibited confusion when dealing with various types of emotions. To mitigate this, we found it necessary to include definitions for each emotion to enhance the model’s understanding and response accuracy. Secondly, we observed a repetitive pattern in the dialogue structure, particularly at the beginning and end of conversations, typically starting with phrases like “How can I help you?” and concluding with “Thank you so much for your advice, bot.” To address this, we implemented prompts that encouraged a variety of starting and ending phrases, aiming to diversify the conversation flow. Thirdly, the dialogues generated by GPT 3.5 tended to be overly generic, lacking specific descriptions or contexts that could elicit distinct emotional responses. To overcome this, we introduced more detailed prompts that guided the model towards generating dialogues with specific instances and scenarios, thereby adding depth and relevance to the conversations. Despite these modifications, the issues persisted, albeit to a lesser extent. It was only after fine-tuning the model with tagged datasets from DailyDialogue that we observed substantial improvements.

3.3.3 Training Cost and Model Behavior Shift

We conducted fine-tuning of a GPT model using a dataset consisting of 1,000 examples from the Daily Dialogue dataset. Our primary focus during this fine-tuning process was to capture a wide range of emotions expressed in conversations. The training data was structured in pairs, with each pair comprising two dialogues. The first dialogue served as an instructive prompt for the model, guiding it to generate a response in the form of the second dialogue.

The training process spanned three epochs, involving a total of 1,223,694 training tokens. This fine-tuning effort incurred a cost of approximately 10 USD. Initially, our training strategy involved stacking multiple conversations together to provide richer context. However, we encountered a significant increase in token usage, which directly impacted the training costs, given OpenAI’s token-based pricing structure.

In response to these cost considerations, we made a strategic adjustment. We modified our training approach to concatenate only two adjacent dialogues within the training dataset. This change effectively reduced the number of

tokens required, consequently lowering the overall training expenses.

While this adjustment successfully improved the authenticity and fluency of the dialogue generation, it also led to a noticeable reduction in the model’s adherence to specific instructions provided in the prompts.

4 Analysis

4.1 Qualitative observations

4.1.1 Simulated Dialogues

Non-numerically, our observations about the dataset reveal certain patterns and limitations. Firstly, the simulated dialogues tend to follow specific routines in emotion shifts. For example, anger is often triggered by cheating partners (mentioned 4 times) or political injustices (mentioned 2 times). This pattern might reflect a bias inherent in the Daily Dialogue dataset. Secondly, some emotion transitions appear abrupt and without context, like a character shifting from happiness to anger or sadness suddenly, prompted by a question from the bot. Lastly, the dialogues lack the complexity and nuance of real human experiences. They often present simplistic, one-dimensional plots, failing to capture the layered and intricate nature of actual human emotional experiences.

4.1.2 Human Annotation

In the context of emotion annotation, more pronounced emotions such as happiness, sadness, and fear tend to achieve higher agreement among human annotators. However, there are instances where the human annotations diverge significantly from machine-generated labels. Notably, disgust is identified as the most ambiguous of the six emotions, often leading to varied interpretations between human and machine labeling. This highlights the complexities involved in accurately identifying and agreeing upon certain emotional states, especially those that are more subtle or less distinctly defined.

4.2 Quantitative Analysis

4.2.1 BERT Analysis

We employed a pre-trained BERT model [4] for sequence classification, fine-tuning it on a daily dialogue dataset. After fine-tuning, we evaluated the model’s performance using two sets of annotations. The first set consisted of human annotations, while the second set was generated by GPT, another language model.

BERT vs Ground Truth Labels: The model exhibits a modest performance with a weighted average F1-score of 0.26 and an accuracy of 0.43 across the micro average. The precision and recall for individual emotions such as 'Anger', 'Disgusted', 'Fear', and 'Surprised' all stand at 0, indicating that the model did not correctly predict any instances of these emotions.

A significant observation is that the model has a relatively better performance on the 'Happy' emotion, with a precision of 0.36 and a recall of 0.86, culminating in an F1-score of 0.51. This suggests that the model is more attuned to recognizing expressions of happiness due to a higher representation of this emotion in the training data or distinct linguistic patterns that make it more detectable.

However, the absence of support for the 'no emotion', 'Disgusted', and 'Surprised' categories indicates that there were no instances of these emotions in the ground truth labels to evaluate against. This highlights a limitation in the dataset's diversity of emotional expressions. Moreover, the small size of the dataset, consisting of only 79 dialogues, could significantly impact the model's ability to generalize across a broader range of expressions, leading to its poor performance on certain emotions.

BERT vs Human Annotations: Similarly, in the analysis of BERT's performance against human annotation, we observed a disparity in the model's ability to accurately predict certain emotions. With a dataset comprising only 79 dialogues, the limited scope may have contributed to this inconsistency, particularly evident in the model's inability to identify any instances of 'no emotion', 'anger', 'fear', and 'sadness'. This resulted in a precision and recall of 0.00 for these categories, indicating a lack of detection by the model.

However, it's noteworthy that the model showed a moderate ability to predict 'disgusted' with a precision of 0.38 and recall of 0.23, and it was relatively more proficient at identifying 'happy', achieving the highest f1-score of 0.58 with a precision of 0.45 and recall of 0.82. 'Surprised' was predicted with a precision of 0.50, but the recall was significantly low at 0.05, suggesting that while the model could identify surprise when it did occur, it often missed it.

Overall, the model's accuracy stood at 0.32, with macro and weighted averages falling around the 0.14 to 0.24 range, indicating a general trend of under-performance. The sparse dataset likely exacerbated these issues, as a more extensive and varied dialogue pool could potentially enhance the model's predictive accuracy and recall across the emotional spectrum.

4.2.2 bcLSTM Analysis

We also try assess our dataset by utilizing a BC-LSTM model [8], specifically trained on the daily dialogue dataset from "Context-Dependent Sentiment Analysis in User-Generated Videos." Evaluation and training code can be found here: https://github.com/acnelexh/LLM_Eval/settings.

This was achieved by modifying the model implementation available at <https://github.com/declare-lab/conv-emotion/tree/master/bc-LSTM-pytorch>. However, upon completion of the training, the model exhibited subpar performance on our dataset, predominantly delivering constant predictions with minimal variation. This disappointing outcome raises the possibility of a bug in the original model's implementation. Alternatively, it could be attributed to the domain discrepancy between the daily dialogue dataset and ours — the former being human-generated and the latter synthesized by GPT. This disparity

not only underscores the potential complexity of our dataset but also suggests that a more robust and comprehensive model may be necessary for enhanced performance.

4.2.3 Human GPT Cross Annotation Analysis

In analyzing the performance of a BERT model trained on simulated dialogues, it becomes apparent that the model excels in identifying emotions like 'Fear' and 'Happiness' with high precision and recall, suggesting these emotions are characterized by distinctive language patterns that the model captures effectively. However, emotions such as 'Anger' and 'Disgusted' pose a challenge for the model, which exhibits moderate to low precision and recall for these categories. This indicates that these emotions may be expressed more subtly in dialogues, making them harder for the model to identify accurately.

The model's performance is stronger for the primary emotions (Emotion 1) compared to secondary emotions (Emotion 2), as indicated by the overall accuracy scores of 0.84 and 0.67, respectively. This difference suggests that the model may be more attuned to the direct expression of emotions rather than the more complex, layered expressions. The limited dataset size could contribute to this performance gap, highlighting the need for larger, more varied datasets to improve the model's understanding of nuanced emotional expressions. Future improvements could involve incorporating more contextual and structural dialogue features to enhance the model's ability to discern a broader spectrum of emotions.

5 Limitations

5.1 Lack of Real Human Persona

Our persona generator was designed to model the global population, but we found that it fell short in capturing the nuanced cultural and gender specificities inherent in diverse human populations. This shortfall is significant because it influences the diversity and realism of the simulated dialogues. Furthermore, the absence of a truly realistic human element in these personas also hampers our ability to train the model for more continuous and authentic conversations. Real human interactions are not only shaped by cultural and gender contexts but also by the fluidity and adaptability inherent in human communication. Our personas, limited by their algorithmic foundation, lack this dynamic quality, leading to dialogues that, while varied, may not fully achieve the depth or the natural flow of genuine human conversations.

5.2 Lack of Pre-training

Our fine-tuning of simulated dialogues on the DailyDialogues dataset resulted in conversations that were more fluid and authentic than those generated before this training. However, due to budget constraints, we could only fine-tune

GPT3.5 with 1,000 dialogues. We believe that broader pre-training on a more extensive range of emotionally-tagged dialogue datasets, such as EmoLine or EmoContext, particularly those enriched with personal experiences and stories, could greatly improve the simulated dialogues. This approach would allow for a wider spectrum of emotional contexts and more intricate interactions, closely resembling real human conversations. Refer to Appendix A.0.1 vs. A.0.2 for comparison.

5.3 Industrial Standard Evaluation

We experimented with several industry-standard models, including BERT and BC-LSTM. However, due to the limited size of our dataset, training and evaluation were challenging, particularly for the large transformer models commonly used today. As a result, we opted to fine-tune these models on other datasets before evaluating them on our own dataset.

6 Potential Impact

Firstly, a pivotal potential impact of our dataset is its capacity to aid models in learning about emotion shifts. By offering a diverse collection of simulated conversations that display significant emotional changes, our dataset has the potential to be an invaluable tool for training models that are more sensitive and responsive. These models could potentially be adept at detecting and adapting to the fine subtleties of emotional transitions, a capability that is vitally important in a range of applications, from mental health support to customer service.

Secondly, our project has the potential to make a substantial contribution to the research community’s understanding of synthetic dialogues generated with Large Language Models (LLMs). We hypothesize that such dialogues often require extensive pretraining to achieve a degree of authenticity and complexity comparable to real human conversations. Our dataset thus has the potential to serve as an important case study in examining the capabilities and limitations of LLM-generated dialogues, providing valuable insights that could inform future advancements in natural language processing and AI communication.

Finally, we hope that our dataset will potentially inspire the creation of more Emotion Recognition Conversation (ERC) datasets on a broader scale. We specifically target the domain of emotion shifts and the subtle expressions of emotions. By showcasing the potential feasibility and utility of synthetic datasets in capturing complex emotional dynamics, we aim to spur further research and development in this area. Our objective is to expand the horizon of ERC datasets, potentially pushing the boundaries of how emotional nuances are comprehended and depicted in AI models, leading, ideally, to AI interactions that are more empathetic and human-like.

References

- [1] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners, 2020.
- [2] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N. Chang, Sungbok Lee, and Shrikanth S. Narayanan. Iemocap: Interactive emotional dyadic motion capture database. In *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC)*, pages 2927–2932, 2008.
- [3] Sheng-Yeh Chen, Chao-Chun Hsu, Chuan-Chun Kuo, Ting-Hao, Huang, and Lun-Wei Ku. Emotionlines: An emotion corpus of multi-party conversations, 2018.
- [4] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019.
- [5] Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. DailyDialog: A Manually Labelled Multi-turn Dialogue Dataset, 2017.
- [6] Linkai Luo and Yue Wang. Emotionx-hsu: Adopting pre-trained bert for emotion classification, 2019.
- [7] Gary McKeown, Michel Valstar, Roddy Cowie, Maja Pantic, and Marc Schroder. The semaine database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent. *IEEE Transactions on Affective Computing*, 3(1):5–17, 2012.
- [8] Soujanya Poria, Erik Cambria, Devamanyu Hazarika, Navonil Majumder, Amir Zadeh, and Louis-Philippe Morency. Context-dependent sentiment analysis in user-generated videos. In Regina Barzilay and Min-Yen Kan, editors, *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 873–883, Vancouver, Canada, July 2017. Association for Computational Linguistics.
- [9] Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. Meld: A multimodal multi-party dataset for emotion recognition in conversations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5275–5285, 2019.

A Dataset examples - 10pt

A.0.1 GPT 3.5 Turbo

Human annotation:

Surprise -> Sad

Dialogues:

Chatbot: Hey there! How's your day going? Any interesting developments at work?

User: Well, actually, I just found out that my coworker got promoted. I had no idea!

Chatbot: Oh, really? That must have been quite a surprise for you. How do you feel about their promotion?

User: I mean, it's great for them, but I can't help but feel a bit jealous. I've been working just as hard, if not harder, and yet I'm still stuck in the same position. It's frustrating.

Chatbot: I can understand why you would feel that way. It's tough when you see others progressing and feel like you're being left behind. Have you talked to your supervisor about your career goals and aspirations?

User: Yeah, I've mentioned it before, but it doesn't seem like they're taking it seriously. It's like my efforts are being overlooked.

Chatbot: That sounds really frustrating. It's important for your supervisor to be supportive and provide opportunities for growth. Have you considered having a more direct conversation with them about your concerns?

User: I shouldn't have to beg for recognition and opportunities. It should be based on merit and hard work. I'm starting to lose my motivation.

Chatbot: I understand your frustration. Feeling undervalued can be incredibly demotivating. Maybe it's time to consider exploring other job opportunities where your hard work will be recognized and rewarded properly.

User: Yeah, I have been thinking about that. It's just disappointing to see my current workplace overlook my efforts like this. I deserve better than this.

Chatbot: Absolutely, you deserve to work in an environment that appreciates and values your hard work. It's important to prioritize your own growth and happiness. If you'd like, I can help you with some job search tips or information on professional development opportunities.

User: That would be helpful. I'm really tired of being overlooked and taken for granted. I deserve a workplace that respects me and gives me the chance to grow.

Chatbot: I completely understand your frustration, and I'm here to support you in any way I can. Let's start by discussing your career goals and what you're looking for in a new workplace. With some focused effort, we'll find a better fit for you.

Human annotation:

Fear -> Happy

Dialogues:

Chatbot: Hey, how are you feeling today?

User: I'm feeling really anxious and scared. There's been a lot going on lately and I'm not sure how to handle it.

Chatbot: I'm here for you. Can you share what has been causing you to feel anxious and scared?

User: It's mainly my financial situation. I've been struggling with my expenses and it's been keeping me up at night. Plus, I'm worried about my relationship with my partner and how this stress could affect us.

Chatbot: I completely understand how financial issues can cause a lot of stress. Have you tried talking to your partner about it? It's important to have open communication and support during difficult times.

User: Yes, we've talked about it, but it's still a source of tension between us. I feel like it's putting a strain on our relationship and I'm scared it might not survive.

Chatbot: It's understandable to have concerns about your relationship during challenging times. Remember that relationships require effort and understanding from both sides. Is there anything you can do together to alleviate the financial stress, like creating a budget or exploring new job opportunities?

User: We've tried budgeting, but it's been difficult to stick to it. And with my age and limited education, finding a new job is proving to be challenging. I just feel stuck and hopeless.

Chatbot: I can understand why you might feel that way. It's important to remind yourself that setbacks are temporary and there are always options available. Have you considered seeking advice from a financial planner or a career counselor? They could provide guidance in making better financial decisions or exploring new career paths.

User: No, I haven't thought about that. I guess it's worth a try. Maybe they can help me find some solutions. Thanks for suggesting that.

Chatbot: You're welcome! It's always good to explore different options and seek support. Remember, you don't

have to face this alone. I believe in you and your ability to find a way through these challenges.

User: Thank you for listening. I didn't expect to find this kind of support here. It means a lot to me.

Chatbot: I'm here to listen and support you. It's important to have someone to talk to. If you ever need to vent or discuss anything, feel free to reach out. You're not alone in this.

User: I appreciate that more than words can express. It's a relief to know I have someone to turn to. Thank you.

Chatbot: You're welcome. Remember, facing challenges with support can lead to surprising breakthroughs. Stay strong and take things one step at a time. I'm here for you whenever you need me.

Human annotation:

Sad -> Happy

Dialogues:

Chatbot: Hey, how's it going? What's been on your mind lately?

User: Oh, hey. Well, I've been feeling pretty down lately. My relationship with my spouse has been going through a rough patch. We've been arguing a lot and it feels like we're drifting apart. It's been making me really sad.

Chatbot: I'm sorry to hear that. Relationships can definitely have their ups and downs. Have you tried talking to your spouse about how you're feeling?

User: Yeah, we've had a few conversations about it, but it just seems like we're not on the same page anymore. It's like we're speaking different languages. I just miss the closeness we used to have.

Chatbot: I can understand how that would make you feel sad. It's hard when the connection you once had starts to fade. It might be helpful to find some activities or hobbies that you both enjoy and can do together. Sometimes, finding common ground can help rebuild that bond.

User: Yeah, I guess that could be worth a try. We used to enjoy going hiking and exploring nature together. Maybe we can start doing that again.

Chatbot: That sounds like a great idea! Nature has a way of bringing people closer and providing a sense of peace. It might be a good opportunity for you to reconnect with each other and create some new positive memories together.

User: Yeah, I really miss those moments of joy and adventure. It's been a while since we've had that. I want to feel happy and excited about our relationship again.

Chatbot: It's completely understandable that you want to feel

happy and fulfilled in your relationship. Remember, it takes time and effort from both parties to rebuild and strengthen a connection. Don't be too hard on yourself and give yourselves the space to grow. And remember, I'm here to support you whenever you need someone to talk to.

User: Thank you, I really appreciate your support. It's comforting to know that someone understands and is there for me. Talking to you has made me feel a little bit happier and hopeful.

Chatbot: I'm glad to hear that. Remember, happiness comes from within, but having a support system can make the journey a little easier. Just take it one day at a time, and don't forget to prioritize your own well-being as well.

A.0.2 Fine-Tune GPT Turbo 3.5 with Daily Dialogues

Human annotation:

Surprised -> Fear

Dialogues:

Chatbot: Hi there! How can I help you today?

User: Oh, I didn't expect to see you here!

Chatbot: ...Why, is something the matter?

User: Well, I just got my test results back from the doctor... and it turns out I have high blood pressure.

Chatbot: Oh no! I'm so sorry to hear that. Is there anything you can do to lower it?

User: They said that making some lifestyle changes could help, but I'm just really scared right now.

Chatbot: I understand, it can be frightening to receive news like that. But try not to worry too much, stress can actually make it worse. It's important to stay calm and focus on the things you can control.

Human annotation:

Sad -> Happy

Dialogues:

Chatbot: Hi there, how are you feeling today?

User: Not so good, really. It's just been one of those days.

Chatbot: Aw, I'm sorry to hear that. Want to talk about it?

User: I don't know. Everything just seems so... I don't know, hopeless, I guess.

Chatbot: Yeah, I get that. Sometimes life can feel overwhelming. Is there something specific that's been bothering you lately?

User: I guess it's just a combination of things. Work is stressful, my family is going through some issues, and I just feel so lonely all the time.

Chatbot: It sounds like you've got a lot on your plate. I hope things start to look up for you soon.

User: Yeah, me too. I'm just tired of feeling down all the time.

Chatbot: Have you tried doing anything to help boost your mood ? Like spending time with loved ones or doing activities that you enjoy?

User: I've been trying, but it's hard. I just don't have the energy sometimes.

Chatbot: I understand. Maybe you could start small, like taking a walk outside or calling up a friend for a chat. Sometimes even little things can make a difference.

User: Yeah, maybe you're right. I'll give it a try.

Chatbot: I believe in you. Remember, it's okay to have bad days, but there's always hope for better ones. You're strong, and you'll get through this.

User: Thank you, that means a lot. I really appreciate your support.

Chatbot: Of course! That's what friends are for. If you ever need someone to talk to, I'm here for you.

User: Thanks, I'll keep that in mind. I'm feeling a little better now.

Human annotation:

Anger -> Anger

Dialogues:

Chatbot: How have you been? I feel like we haven't caught up in a while.

User: Yeah, it's been a long time. I've been pretty busy lately with work and other things. How about you?

Chatbot: I've been good, thanks for asking. Have you been keeping up with your hobbies? I remember you used to be really into painting and knitting.

User: Ugh, don't even get me started on that. I tried picking up my knitting needles the other day and I just couldn't get into it. It's so frustrating when you lose interest in something you used to love.

Chatbot: I understand how that can be discouraging. Sometimes, our interests and passions change over time, and that's okay. It might be worth exploring new hobbies and activities that spark your interest now.

Human annotation:

Anger -> Happy

Dialogues:

User: Have you heard the news about the politician who was caught taking bribes?

Chatbot: I can't believe it either. It's really disheartening to see such corruption in our government.

User: It's just so infuriating. These people are meant to serve and protect us, and instead they're just looking out for themselves.

Chatbot: Yeah, it's a real betrayal of the public's trust. I hope justice is served and they're held accountable for their actions.

User: I don't have much faith in the justice system. It feels like they always get away with it.

Chatbot: I can understand why you might feel that way. But we have to continue fighting for justice and holding our leaders accountable.

User: You're right. We can't let this kind of behavior go unchecked.

Chatbot: Absolutely. We have a voice and we need to use it to push for change and demand better from our elected officials.

User: I'm glad we had this conversation. It's helped me feel a bit more hopeful.

Chatbot: I'm glad to hear that. We have the power to make a difference when we come together and stand against corruption.

User: Thanks for listening and for encouraging me. It means a lot.

Chatbot: Of course. Anytime you need a listening ear or some motivation, I'll be here for you.

B Prompt Engineering

B.0.1 Prompt Examples Original

=====

SYSTEM:

CHATBOT PERSONA:

Here are the persona for the chatbot:

1. The chatbot should be have neutral personality with little emotion.
2. The chatbot should use the clues from the USER persona to elicit the emotional shift.

3. The chatbot should act as a friend for the user.

USER PERSONA:

Here are the persona for the user:

1. Age: 64
2. Gender: Female
3. Cultural Background: African
4. Occupation: Business
5. Education: Bachelor's Degree
6. Family Dynamics: Nuclear Family
7. Relationship Status: Single
8. MBTI: ESFP
9. Attachment Style: Secure
10. Emotional Intelligence: 5 out of 5
11. Typical Mood: 1 out of 5
12. Emotional Range: 1 out of 5
13. Stress Triggers: ['Social Interactions', 'Relationships', 'Finance']
14. Coping Strategies: ['Problem-Solving', 'Professional Help', 'Substance Use']
15. Significant Events: ['Loss of a Loved One']
16. Likes: ['Travel & Exploration', 'Social & Cultural Activities']
17. Goals and Aspirations: ['Environmental Engagement', 'Financial Goals', 'Creative and Artistic Pursuits']
18. Physical Health: 3 out of 5
19. Mental Health: 1 out of 5

=====

CONTENT:

Here are some simulated examples from the Daily Dialogue dataset.

A-[No emotion]: Did you hear the news ?

B-[No emotion]: What happened ?

A-[No emotion]: Our cousin went into labor and had her baby last week .

B-[Surprise]: She did ? Why didn't anyone tell me ?

A-[No emotion]: I would've thought that somebody would have told you .

B-[No emotion]: No , I had no idea .

A-[No emotion]: Well , she did , her baby was 8 pounds 6 ounces .

B-[Happiness]: Oh my God , that's great !

A-[No emotion]: Are you going to go and visit her and the baby ?

B-[No emotion]: I think that I might .

A-[Happiness]: Good ! I just thought I'd let you know .

B-[Happiness]: Thanks for telling me .

A-[Anger]: I have been looking at this online catalog for over an hour and I still haven't finished getting all the kitchen

appliances that we need !

B-[No emotion]: What are you getting ?

A-[No emotion]: Well , the first thing on my list is a new blender . I decided to also get a juicer and a new coffee maker .

B-[No emotion]: Don't forget to also get a new mixer . I lent the old one to my brother and he broke it .

A-[No emotion]: Yeah I know . I also decided to throw away the old toaster and get a new one . I am also getting a rice cooker and steamer to make some nice steamed fish or veggies .

B-[No emotion]: I'm actually thinking of completely refurnishing the kitchen and getting a new stove , oven , dishwasher and trash compacter .

A-[Happiness]: That's a good idea ! The kitchen will look amazing !

A-[No emotion]: what did you do over the weekend ?

B-[Happiness]: I went a global warming rally in London . It was fantastic to be around so many people who care about the environment .

A-[No emotion]: do you think there's anything we can do to reverse the damage that's been done already ?

B-[No emotion]: it might not be possible to fix the problems that we've created for ourselves , but there are lots of things we can do to prevent more damages from happening .

A-[No emotion]: like what ?

B-[No emotion]: well , we can use public transport instead of taking our cars for a start .

A-[No emotion]: what else can we do to protect the environment ?

B-[No emotion]: if you do have to drive , you should make sure that your car runs on unleaded petrol . Also , your home should use sources of renewable energy .

A-[No emotion]: how about recycling ? Does that actually help ?

B-[No emotion]: yes . You should take your glass , paper , plastic , cardboard , and tin cans to a recycling center .

A-[No emotion]: what do you think is the biggest worry for our future ?

B-[No emotion]: I think that the issue of greatest concern is having enough sources of clean water for everyone .

A-[Surprise]: I had no idea you were such an environmentalist before !

B-[No emotion]: to be honest , in order for the earth to continue to be a habitable place , we're all going to have to become more interested in the environment .

A-[No emotion]: Have you got your invitation yet ?
 B-[Surprise]: My invitation ? No , I haven't . My invitation to
 what ?
 A-[No emotion]: The house warming party .
 B-[No emotion]: Whose house warming party is it ?
 A-[No emotion]: Tom and Bill Smith . They are both working now
 you know ? And they've bought a new house .
 B-[Surprise]: Oh , they have ? I didn't know . I haven't seen Tom
 lately .
 A-[No emotion]: It's out in the suburbs .
 B-[No emotion]: Have you seen the house ?
 A-[No emotion]: Yes , I have . I went out with them last weekend
 .
 B-[No emotion]: Is it nice ?
 A-[No emotion]: Yes , it is . There are three bedrooms , a living
 room , a dining room , and a big kitchen . There's also a
 garden .
 B-[Happiness]: Well , that does sound nice . Have they moved in
 yet ?
 A-[No emotion]: They are moving today . Tom's taken the day off .
 He's rented a truck , and they should have all their
 furniture in the house tonight .
 B-[No emotion]: When are they going to have the party ?
 A-[No emotion]: Next Saturday night . You should get your
 invitation today or tomorrow .
 B-[Happiness]: Wow , that would be something to look forward to .

 A-[No emotion]: So can we play PlayStation later ?
 B-[Happiness]: Sure , if you're done !
 A-[No emotion]: All my games are so lame , Mom . When can I get a
 new one ?
 B-[No emotion]: Well , listen , Honey , you've been a big help
 lately ...
 A-[No emotion]: Right on . I'm gonna get a new game !
 B-[No emotion]: If you take out the garbage and do the dishes , I
 'll think about it .
 A-[Sadness]: I knew there would be a catch !
 B-[No emotion]: Nothing in life is free , Son .
 A-[Sadness]: Except housework , that is .

PARAMETER:
 USER STARTING EMOTION: Happy
 USER ENDING EMOTION: Fear
 TURNS PER SIMULATION: 7

TOPIC: School Life

1. Simulate a conversation between the CHATBOT and USER, aligning with their individual persona with the topic School Life. Begin the conversation skipping formal greetings. This will make the conversation feel more immediate and focused.
2. The USER should only show Happy, Fear, and neutral emotion throughout the conversation. USER should start with a initial emotion state of Happy, through gradual shift in emotion guided by CHATBOT towards the final emotion state of Fear.
3. The USERs emotions should shift gradually, not abruptly, to keep the conversation natural. Suggest the chatbot to ask probing questions or make statements that could realistically lead to the final emotion state.
4. Generate 7 turns of conversation, with the following format:
CHATBOT: [...]
USER-[EMOTION]: [...]
5. Natural Display of Emotion: Use descriptive language that naturally conveys the USER's emotional state through their word choice, tone, and the content of their speech rather than explicitly stating the emotion state. Include subtle cues that indicate a shift in emotion, such as changes in the USER's responsiveness, the length of their messages, or their use of punctuation and capitalization.
6. Detailed and realistic conversation: USER should provide specific details about the trigger of their emotions to make it more believable, e.g. specific relationship drama or dynamic (e.g. cheating husband/wife, missed date, unbalanced relationship dynamic) that contribute to sadness or disgust, specific activity and role models (e.g. reading Socrates, Shakespear, etc) that brings them joy and excitement.
7. Adopt the personality described in the character section below and respond to the last message in conversation history. Consider the complete conversation history, the additional context, the character's persona, emotional state and goals below when simulating.
8. Avoid Forced Positivity: If the conversation naturally leads to a less positive conclusion, let it be. Not every conversation has to end on a high note, especially if it doesn't fit the flow of the dialogue
9. Varied Conversation Endings: The conversation doesn't need to end with USER thanking the CHATBOT for listening. Allow for a variety of conversation endings that are more aligned with the final emotion state of Fear.
10. Definition of EMOTIONs:
Happy/Joy - is often defined as a pleasant emotional state

- that is characterized by feelings of contentment, joy, gratification, satisfaction, and well-being.
- Sadness - Sadness is another type of emotion often defined as a transient emotional state characterized by feelings of disappointment, grief, hopelessness, disinterest, and dampened mood. Like other emotions, sadness is something that all people experience from time to time. In some cases, people can experience prolonged and severe periods of sadness that can turn into depression. Sadness can be expressed in a number of ways including: Crying, Dampened mood, Lethargy, Quietness, Withdrawal from others.
- Fear - Fear is a powerful emotion that can also play an important role in survival. When you face some sort of danger and experience fear, you go through what is known as the fight or flight response.
- Disgust - This sense of revulsion can originate from a number of things, including an unpleasant taste, sight, or smell. Researchers believe that this emotion evolved as a reaction to foods that might be harmful or fatal. When people smell or taste foods that have gone bad, for example, disgust is a typical reaction. Poor hygiene, infection, blood, rot, and death can also trigger a disgust response. This may be the body's way of avoiding things that may carry transmittable diseases. Disgust could also be related to contempt of another person or situation.
- Anger - Anger can be a particularly powerful emotion characterized by feelings of hostility, agitation, frustration, and antagonism towards others. Like fear, anger can play a part in your body's fight or flight response.
- Surprise - Surprise is usually quite brief and is characterized by a physiological startle response following something unexpected. A pleasant surprise would be arriving home to find that your closest friends have gathered to celebrate your birthday.
11. Use daily dialogue examples as reference for the simulation to generate realistic emotion through conversation.

Response: [...]

B.0.2 Prompt Examples Fine-Tuned

=====

SYSTEM:

CHATBOT PERSONA:

Here are the persona for the chatbot:

1. The chatbot should be have neutral personality with little emotion.
2. The chatbot should use the clues from the USER persona to elicit the emotional shift.
3. The chatbot should act as a friend for the user.

USER PERSONA:

Here are the persona for the user:

1. Age: 75
2. Gender: Female
3. Cultural Background: North American
4. Occupation: Business
5. Education: Secondary School
6. Family Dynamics: Extended Family
7. Relationship Status: Single
8. MBTI: ENFP
9. Attachment Style: Secure
10. Emotional Intelligence: 5 out of 5
11. Typical Mood: 1 out of 5
12. Emotional Range: 4 out of 5
13. Stress Triggers: ['Family', 'Relationships', 'Finance']
14. Coping Strategies: ['Problem-Solving', 'Exercise', 'Substance Use']
15. Significant Events: ['Loss of a Loved One']
16. Likes: ['Digital Entertainment', 'Mind & Body Wellness']
17. Goals and Aspirations: ['Personal Fulfillment', 'Career Development', 'Relationships and Family']
18. Physical Health: 1 out of 5
19. Mental Health: 1 out of 5

=====

CONTENT:

PARAMETER:

USER STARTING EMOTION: Disgust

USER ENDING EMOTION: Surprise

TURNS PER SIMULATION: 9

TOPIC: Work

1. Simulate a conversation between the CHATBOT and USER, aligning with their individual persona with the topic Work. Begin the conversation skipping formal greetings. This will make the conversation feel more immediate and focused.
2. The USER should only show Disgust, Surprise, and neutral

- emotion throughout the conversation. USER should start with a initial emotion state of Disgust, through gradual shift in emotion guided by CHATBOT towards the final emotion state of Surprise.
3. The USERs emotions should shift gradually, not abruptly, to keep the conversation natural. Suggest the chatbot to ask probing questions or make statements that could realistically lead to the final emotion state.
 4. Generate 9 turns of conversation, with the following format:
CHATBOT: [...]
USER-[EMOTION]: [...]
 5. Natural Display of Emotion: Use descriptive language that naturally conveys the USER's emotional state through their word choice, tone, and the content of their speech rather than explicitly stating the emotion state. Include subtle cues that indicate a shift in emotion, such as changes in the USER's responsiveness, the length of their messages, or their use of punctuation and capitalization.
 6. Adopt the personality described in the character section below and respond to the last message in conversation history. Consider the complete conversation history, the additional context, the character's persona, emotional state and goals below when simulating.
 7. Avoid Forced Positivity: If the conversation naturally leads to a less positive conclusion, let it be. Not every conversation has to end on a high note, especially if it doesn't fit the flow of the dialogue
 8. Varied Conversation Endings: The conversation doesn't need to end with USER thanking the CHATBOT for listening. Allow for a variety of conversation endings that are more aligned with the final emotion state of Surprise.
 9. Definition of EMOTIONS:
 - Happy/Joy - is often defined as a pleasant emotional state that is characterized by feelings of contentment, joy , gratification, satisfaction, and well-being.
 - Sadness - Sadness is another type of emotion often defined as a transient emotional state characterized by feelings of disappointment, grief, hopelessness, disinterest, and dampened mood. Like other emotions, sadness is something that all people experience from time to time. In some cases, people can experience prolonged and severe periods of sadness that can turn into depression. Sadness can be expressed in a number of ways including: Crying, Dampened mood, Lethargy, Quietness, Withdrawal from others.
 - Fear - Fear is a powerful emotion that can also play an

important role in survival. When you face some sort of danger and experience fear, you go through what is known as the fight or flight response.

Disgust - This sense of revulsion can originate from a number of things, including an unpleasant taste, sight, or smell. Researchers believe that this emotion evolved as a reaction to foods that might be harmful or fatal. When people smell or taste foods that have gone bad, for example, disgust is a typical reaction. Poor hygiene, infection, blood, rot, and death can also trigger a disgust response. This may be the body's way of avoiding things that may carry transmittable diseases. Disgust could also be related to contempt of another person or situation.

Anger - Anger can be a particularly powerful emotion characterized by feelings of hostility, agitation, frustration, and antagonism towards others. Like fear, anger can play a part in your body's fight or flight response.

Surprise - Surprise is usually quite brief and is characterized by a physiological startle response following something unexpected. A pleasant surprise would be arriving home to find that your closest friends have gathered to celebrate your birthday.

Response: [...]

B.0.3 Tables

Emotion	Precision	Recall	F1-Score	Support
No Emotion	0.00	0.00	0.00	0
Anger	0.00	0.00	0.00	23
Disgusted	0.00	0.00	0.00	0
Fear	0.00	0.00	0.00	20
Happy	0.36	0.86	0.51	43
Sadness	0.00	0.00	0.00	0
Surprised	0.00	0.00	0.00	0
Micro Avg	0.23	0.43	0.30	86
Macro Avg	0.05	0.12	0.07	86
Weighted Avg	0.18	0.43	0.26	86

Table 1: BERT vs Ground Truth Analysis

Emotion	Precision	Recall	F1-Score	Support
No Emotion	0.00	0.00	0.00	0
Anger	0.00	0.00	0.00	24
Disgusted	0.38	0.23	0.29	13
Fear	0.00	0.00	0.00	26
Happy	0.45	0.82	0.58	56
Sadness	0.00	0.00	0.00	19
Surprised	0.50	0.05	0.09	20
Accuracy			0.32	158
Macro Avg	0.19	0.16	0.14	158
Weighted Avg	0.25	0.32	0.24	158

Table 2: BERT vs Human Annotation Analysis

B.0.4 Classification Report for Emotion 1

Class	Precision	Recall	F1-Score	Support
1	0.71	0.83	0.77	12
2	0.56	0.56	0.56	9
3	0.92	1.00	0.96	11
4	0.91	1.00	0.95	20
5	0.91	1.00	0.95	10
6	0.91	0.59	0.71	17
Accuracy			0.84	79
Macro Avg	0.82	0.83	0.82	79
Weighted Avg	0.84	0.84	0.83	79

Table 3: Classification Report for Emotion 1

B.0.5 Classification Report for Emotion 2

Class	Precision	Recall	F1-Score	Support
1	0.70	0.64	0.67	11
2	1.00	0.44	0.62	9
3	0.64	1.00	0.78	9
4	0.59	0.87	0.70	23
5	0.50	0.50	0.50	8
6	1.00	0.47	0.64	19
Accuracy			0.67	79
Macro Avg	0.74	0.65	0.65	79
Weighted Avg	0.75	0.67	0.66	79

Table 4: Classification Report for Emotion 2