Synthesizing Dialogues from Social Media Data for Mental Health Support: A Case Study on Gambling Disorders

Zhouqing Zhang[1], **Lis Kanashiro Pereira**[3], **Wan Jou She**[2],
Benjamin Adams[1], **Tham Piumsomboon**[1], **Kongmeng Liew**[1]
[1]University of Canterbury, New Zealand [2]Kyoto Institute of Technology, Japan
[3]National Institute of Information and Communications Technology, Japan
zhouqing.zhang@pg.canterbury.ac.nz liskanashiro@nict.go.jp wjs2004@kit.ac.jp
{benjamin.adams, tham.piumsomboon, kongmeng.liew}@canterbury.ac.nz

Abstract

Recent advances in artificial intelligence have seen natural language processing (NLP) applications used increasingly for mental health support, such as question-answering systems, conversational agents, and chatbots. However, these applications often require high-quality dialogue datasets, which are scarce for certain under-resourced mental health conditions like gambling disorders. Privacy concerns and the sensitive nature of mental health issues make real-world data collection difficult and time-consuming. We propose a framework that uses publicly available social media posts as seeds to generate empathy-driven multi-turn dialogues between help-seekers and supporters in mental health conditions that have been under-researched in NLP. Through a case study on gambling disorder, we demonstrate the effectiveness of our approach in producing higher-quality dialogues than standard role-playing prompts. Our work showcases an effective dataset generation method for fine-tuning mental health language models alongside downstream applications like dialogue system development, particularly for low-resourced mental health disorders like gambling disorders.

1 Introduction

Recent advances in artificial intelligence (AI) have demonstrated high efficiency, precision, and intelligence in language generation tasks across numerous domains (Rashid and Kausik, 2024), with mental health being no exception (D'Alfonso, 2020).

AI-driven mental health applications typically employ natural language processing (NLP) to develop conversational models, chatbots, and dialogue systems that provide various forms of support (D'Alfonso, 2020; Olawade et al., 2024). Among these, one task that has attracted significant attention is emotional or empathetic support through dialogue models.

As with any NLP task, domain-specific data and corpora are fundamental to developing intelligent systems for that purpose (Pustejovsky and Stubbs, 2012). However, mental health constitutes a relatively niche subfield within NLP (Mandal et al., 2025). Research on mental health within NLP has typically overlooked rare conditions like gambling disorders, where relevant corpora are extremely scarce or even entirely absent. Furthermore, due to human factors such as privacy concerns and sensitivity, as well as computational factors like low efficiency and high costs, collecting naturalistic training corpora for these low-resourced mental health disorders has proven to be extremely challenging (Michael; Muqtadir et al., 2024; Lu et al., 2023).

To address these challenges, we propose a method to enhance the synthetic generation of conversational corpus data for these niche and low-resourced mental health disorders. In this paper, we define these as disorders that are comparatively less researched in NLP and have received less investment in corpus or model development. Specifically we demonstrate a system for generating comprehensive, multi-turn dialogue data, seeded from a diverse set of social media posts on mental disorder-related experiences.

Inspired by the capabilities of large language models (LLMs), and the successes of prior studies that have used LLMs for data augmentation and generation (Zhang et al., 2024a; Zheng et al., 2023; Chen et al., 2023b; Kim et al., 2024; Qiu and Lan, 2024), we apply LLMs for automated data synthesis. To the best of our knowledge, past research on utilizing LLMs for data synthesis has typically relied on using a small amount of existing data as seed data, which is subsequently augmented through role-playing by LLMs. We demonstrate that such seed data can also include relevant posts on social media. These often autobiographical posts provide rich accounts of individuals' lived

Findings of the Association for Computational Linguistics: IJCNLP-AACL 2026, pages 1055–1073
November 6-10, 2026 ©2026 Association for Computational Linguistics

experiences pertaining to a given context (Marin, 2021; Lopez et al., 2019). Moreover, these posts are often authentic, unique, and informative.

Additionally, we aim to address the limitations of current LLM role-playing methods for dataset generation and multi-turn dialogue completion tasks using such data as seeds. Past examples have tended to fall short of real, naturalistic conversations by appearing inauthentic, with limited information capture, and limited empathy even in empathy-emphasized conversation construction (Cuadra et al., 2024; Wang et al., 2025). Consequently, we propose a multi-agent framework for synthesizing empathy-driven support dialogues seeded from social media data. This is particularly relevant for low-resourced mental health conditions, such as gambling disorders, with scarce availability of formal training corpora, but frequently discussed on social media.

In sum, our study aims to: 1) Propose a multi-turn dialogue data synthesis framework for empathy-driven supporters and users with low-resourced mental health disorders (such as gambling disorders), leveraging social media posts as situational seeds through a multi-agent LLM system. 2) Synthesize data through our proposed framework and validate its effectiveness via a combination of LLM-as-a-judge and human evaluations on relevant metrics and indices. Focusing specifically on gambling disorders, we release a dataset comprising 1,053 synthesized, multi-turn dialogues between individuals with gambling disorders and empathy-driven supporters.

2 Related Work

2.1 Research Using Social Media Posts

As a publicly available resource, social media data (such as Reddit posts[1]) offers advantages such as ease of access, greater realism, rich contextual diversity, and a large volume of data (Lamberti et al., 2025). It is a valuable source of information and is widely used in research (Lopez et al., 2019). Previous researchers have used social media data to identify social issues, analyze trends in public discourse, and monitor infodemiological trends in mental health conditions, such as anxiety and depression, amongst other topics (Lachmar et al., 2017; Skaik and Inkpen, 2020; Zhang et al., 2024b).

[1]https://www.reddit.com

Our research harnesses this rich contextual information and diverse personal situations found in social media data: We use social media data as contextual 'anchors' in dialogue construction that ground the generated dialogues in real-world discourse structures, emotions, and storytelling, to generate naturalistic but synthetic multi-turn conversations from our multi-agent system.

2.2 Data Synthesis

In the field of NLP, early approaches to data synthesis primarily involved augmenting existing datasets through computational methods, such as machine back translation, synonym mapping, and filling in fixed templates (Wei and Zou, 2019; Sugiyama and Yoshinaga, 2019; Queudot et al., 2022). Since then, the emergence of sequence-to-sequence models and transformer-based architectures has enabled data synthesis to evolve accordingly. Newer commonly-used text synthesis methods include random replacement and insertion, semantic-preserving rewriting, generative adversarial networks (GANs), and conditional text generation, etc (Jung et al., 2023; Chen et al., 2023a; Wan et al., 2022; Antoniou et al., 2017). With the emergence of autoregressive LLMs, such as ChatGPT (Kumamoto et al., 2023), the synthesis of target data has shifted toward prompt-based instructions for contextual generation by LLMs (Zhang et al., 2024a; Zheng et al., 2023; Chen et al., 2023b).

Data synthesis and construction through LLMs have been widely adopted in mental health domains to synthesize large-scale conversational data for emotional support. Zheng et al. (2023) used a small amount of dialogue data as seeds and employed prompts to instruct the LLMs to generate more multi-turn conversations. To augment the volume of empathetic conversational data, Chen et al. (2023b) guided LLMs to generate multi-turn dialogue data through prompt instructions when given only a short context. Similarly, Kim et al. (2024) utilized LLMs to develop a multi-feature-emphasized pipeline for multi-turn counseling conversation augmentation through single-turn psychotherapy counseling sessions.

For mental health-related dataset synthesis, seed data selection and preparation present a significant challenge. Due to its sensitive nature, Zhang et al. (2024a) proposed obscuring their seed data (counseling reports) behind a two-step framework that first converted these reports into deidentified domain knowledge, then guided LLMs to generate

dialogues through prompt instructions. Similarly, to address the scarcity of data in psychological counseling, Qiu and Lan (2024) proposed using LLMs for role-playing to simulate conversations and generate dialogue data.

As aforementioned, previous studies have also pointed out the limitations of role-playing methods in data synthesis. In particular, they have limited information capture capabilities; its empathetic responses consistently follow the same style, and the overall conversation follows a predictable pattern, resulting in a lack of authenticity (Cuadra et al., 2024; Wang et al., 2025).

To address these challenges, we propose a multi-agent data synthesis framework that incorporates frontier LLMs. We aim to synthesize multi-turn dialogue data between help-seekers and empathetic supporters, utilizing contextually relevant posts from social media, for applied NLP tasks for these low-resourced mental health disorders.

3 Method

We rely on a multi-agent framework, consisting of a formulator and a generator, with separate LLMs for simulating a multi-turn conversation between a user (help-seeker) and an empathetic supporter (see Figure 1). Within the formulator, we set a feature extractor module and a dialogue scaffolding module. Within the generator, we have a user agent and a supporter agent. In our study, we utilize relevant posts from Reddit[1] as seed information. These narrative, autobiographical social media posts chronicle each user's lived experience with a specific mental health related concern. We refer to posts as situations.

3.1 Formulator

3.1.1 Feature Extractor

Individuals experiencing mental health challenges often use multifaceted emotional expressions (Werner-Seidler et al., 2020). The extractor module aims to identify contextual information about each situation. At present, we focus on emotion recognition as the main context for each situation.

Using (Rashkin et al., 2019)'s dataset with 32 emotion labels, we fine-tuned a sentiment model for 32 emotion classification. In particular, we chose roberta-large, a model already trained for 28-emotion classification, as the backbone (Accessible on Hugging Face[2]). After fine-tuning the model using the above data, we validated the accuracy of the model by checking if the ground-truth emotional label was among the top 3 predicted labels by the model. With this criterion, the model achieved an accuracy of 84.4% (details about the model training, settings, evaluation, etc. are in Appendix A). These top 3 emotions were then used as guiding queries for the User Agent (see section 3.2.1).

3.1.2 Dialogue Scaffolding

This module extracts a list of discourse items that were sequentially mentioned in the original situation text. These items then form a dialogue checklist of content that the user agent would raise in the simulated conversation. That is, we prompt an LLM (GPT-4.1) to generate a list of text sequences that are extracted from a situation. This results in a list of sequential dialogue content options, written in first-person, that would guide the flow of dialogue content covered by the user in the constructed conversation, see an example in Table 1 (prompt details are in the Appendix B).

A situation (original post)
Last night, I spent another $200 on gambling tokens again, and I feel so embarrassed and ashamed for giving in, please help!
Generated dialogue checklist
1) I spent another $200 on gambling tokens again. 2) I feel embarrassed and ashamed for giving in. 3) I need help.

Table 1: An example of dialogue checklist generation

3.2 Generator

3.2.1 User Agent

Our user agent is based on GPT-4o, which incorporates a dialogue planner module and an utterance redundancy detection module.

Dialogue Planner The dialogue checklist generated by the dialogue scaffolding module determines the progress of the entire multi-turn conversation. Utterances from user agents within a dialogue can be of two main types: proactive queries and response queries. The dialogue planner module is primarily used to decide the user's response query when supporters' replies contain a question; otherwise, the user will execute a proactive query.

Proactive queries represent spontaneous user-initiated interactions, while response queries pri-

[2]https://huggingface.co/Laksssshya/roberta-large-goemotions

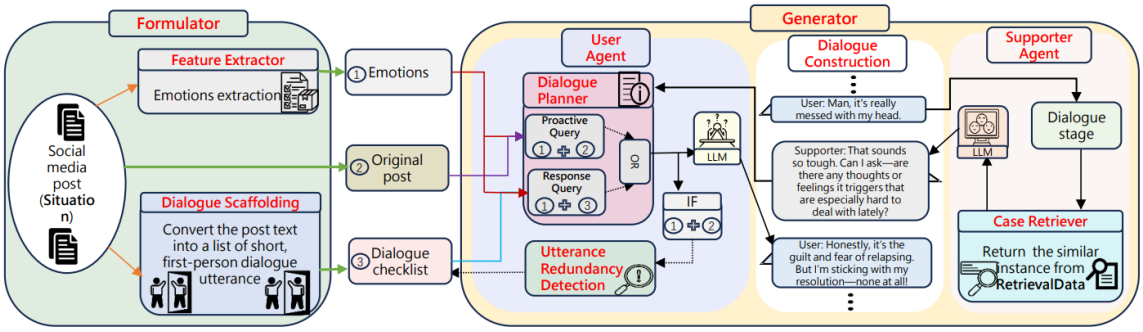


Figure 1: The overview of our dialogue synthesis framework. The Formulator module prepares the emotions, original posts, and dialogue checklists for the generator. During the dialogue construction, the dialogue planner determines whether the User Agent is executing a proactive query or a response query, and sends the relevant materials to generate the query for the Supporter Agent. Here, if a response query is generated, the Utterance Redundancy Detection will be executed. After the User Agent sends a query to the Supporter, the Supporter Agent determines the dialogue stage, and the case retriever will search for the most similar instance from the RetrievalData according to the User Agent's query. the Supporter Agent then adopts the associated answering strategies from the RetrievalData to generate its reply. This process is repeated until the dialogue checklist is exhausted.

marily address supporter's questions. Proactive queries are sourced sequentially from our dialogue checklist, as incorporated into prompts of GPT-4o to align with ongoing dialogue contexts. Response queries primarily utilize original situation text as integrated into prompts of GPT-4o, to generate replies that are consistent with the overall conversation flow (the prompts used for this are reported in the Appendix C). Regardless of the query type, we will incorporate the three extracted emotion words into the prompt to construct emotionally rich characters.

Utterance Redundancy Detection To prevent redundant content in conversations during the interplay of response queries and proactive queries, we have introduced an utterance redundancy detection module. That is, each time a user agent generates a response query, we compare it against utterances in the dialogue checklist using similarity measures. The similarity calculation here is based on both semantic and lexical factors, each accounting for 50%. Semantic similarity is calculated using cosine similarity between the query and utterances in the dialogue checklist based on document embeddings. We used sentence-transformers/all-MiniLM-L6-v2[3] for this purpose. Lexical similarity is calculated by the BM25 algorithm (Robertson and Zaragoza, 2009), where it assigns a similarity score between the query and every utterance in the dialogue checklist at the lexical level. If the similarity score of the top 1 utterance in the dialogue checklist is above 0.7, we remove it from the dialogue checklist.

3.2.2 Supporter Agent

To generate utterances with a different expression style from the GPT-4o-based user agent, we set GPT-4.1 for the supporter agent. Additionally, we included a case retriever module to enhance the supporter's empathy.

Case Retriever Following past work by Liu et al. (2021), we incorporate commonly used response strategies to guide empathetic response generation. We referenced their 8 response strategies: Question, Restatement/Paraphrasing, Reflection of Feelings, Self-disclosure, Affirmation and Reassurance, Providing Suggestions, Information, and Others.

According to the length of the dialogue checklist generated by the dialogue scaffolding module, we roughly divide a dialogue into three stages: the beginning, the middle, and the end. For the three stages, we selected either partial or full response strategies from among the 8 strategies (The details are in Appendix D).

For every user query, the supporter agent first considers the current dialogue stage and decides on one of 8 response strategies that seed the response generation. To assist the supporter in selecting appropriate strategies and provide a response sample, we introduce the case retriever module.

Specifically, we adopt the data constructed by the work of (Liu et al., 2021) and reformat it into Q&A-

[3]https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2

based pairs from multi-turn empathetic dialogues, with annotations for response strategy (n = 15204 pairs). We refer to this data as RetrievalData. When the supporter needs to respond to user queries, the range of RetrievalData will be specified through the current dialogue stage. Then the case retriever module retrieves the most similar question from the determined RetrievalData, and returns the matching response strategy and answer for that question. The retrieved response strategy and answer sample are incorporated into the prompt, instructing the GPT-4.1 model to generate a reply suitable for the ongoing conversation (We reported the settings and prompts in Appendix D).

Here, the case retriever module retrieves similar instances from the RetrievalData by calculating semantic similarity. We employ document embedding to compute cosine similarities between user queries and questions in the RetrievalData. The document embedding model we use is sentence-transformers/all-MiniLM-L6-v2[3]. To enhance the retrieval process, through this module, we retrieve the top 50 similar instances, rank their response strategies by frequency, and select the highest-ranked strategy as the candidate strategy. Based on this strategy, we determine the highest-ranked instance in the retrieved instances as the most similar instance. We then return this most similar instance's response strategy and answer to the supporter for response generation to the user's query. Within the same dialogue, to avoid redundant lookups, we remove each instance from the RetrievalData after it has been retrieved.

3.3 Overall Workflow

In sum, when synthesizing a dialogue, the extractor module will identify 3 emotions from the situation text, and the dialogue scaffolding module will interpret the situation as a dialogue checklist, forming the inputs to the generator. We also included pleasantries, such as a greeting at the start of the conversation and expressions of gratitude at the end. We reported the core details on the generative process for a multi-turn empathy-driven dialogue between the user and the supporter, which can be seen in Algorithm 1.

Algorithm 1 Generator Algorithm

Require: A situation (post)
Ensure: 3 extracted emotions, and the generated dialogue checklist

record[] ← ∅
item ← *dialogue_checklist.pop*(0)
content ← null
dialogue_stage ← *begining*
while *dialogue_checklist* ≠ ∅ **do**
    if *item* ≠ ∅ **then**
        proactive_query = user(record, emotions, item)
        record.append(proactive_query)
        content = content + proactive_query
    end if
    item ← ∅
    strategy, answer = case_retriever(content)
    response = supporter(record, content, strategy, answer)
    dialogue_stage.update()
    record.append[response]
    content ← null
    if "?" in response **then**
        response_query = user(record, emotions, situation)
        record.append(response_query)
        content = content + response_query
        utterance_redundancy_detection(content, dialogue_checklist)
        if "?" not in response_query **then**
            item = dialogue_checklist.pop(0)
        end if
    else
        item = dialogue_checklist.pop(0)
    end if
end while
return record

4 Case Study: Gambling Disorder

We applied our method to synthesize a dataset for gambling disorder support. Individuals with gambling disorders require not only professional treatment but also empathetic support from others who understand their psychological and emotional struggles (Burleigh et al., 2019). Therefore, we construct an empathy-driven multi-turn dialogue dataset between the user with gambling disorders and the empathetic supporter to facilitate downstream tasks such as the development and construction of various dialogue systems.

4.1 Data Collection and Processing

We first collect relevant posts (situations) from the r/problemgambling[4] on Reddit[1], spanning from January 1, 2024, to December 31, 2024, totaling 15,102 posts.

Our data processing consists of three steps. 1) We removed irrelevant posts that did not originate from individuals with gambling disorders and did not share genuine experiences related to gambling addiction; this also excludes some advertisements, noise, etc. This was to ensure that posts were about each individual's personal experiences with gambling—their situations, difficulties, struggles—and

[4]https://www.reddit.com/r/problemgambling

their desire for help, advice, guidance, and so on. At this step, we primarily cleaned the data through keyword matching. 2) We only retained posts with lengths between 50 and 300 words. Posts that are too long or too short are not well-suited for building multi-turn conversations. 3) This involved manually reviewing each post. This ensured that no personally identifiable information (PII) or sensitive topics was present, while also further enhancing the overall quality of the content as mentioned in the first step. This resulted in a final N = 1,053 posts (situations).

4.2 Dialogue Construction

Following our framework, when performing dialogue synthesis, the feature extractor module will identify 3 emotion words for each situation. Then, the dialogue scaffolding module converts the textual situation into a list of utterances, referred to as the dialogue checklist. Next, the user agent and the supporter agent interact until the dialogue checklist becomes empty (prompt details are in the Appendix C, D). A multi-turn dialogue between an individual with gambling disorders and an empathetic supporter has been constructed.

All 1,053 situations were used as seeds for generating dialogue data through our framework. For comparison, we applied a baseline method (zero-shot role-playing) on GPT-4.1 to synthesize multi-turn dialogue for 1,053 situations. In addition to the modules defined in our method, we ensured that the instructions for the two methods were aligned (more details in Appendix E). This resulted in two sets of dialogue data. Basic statistical information for the two sets of dialogue data is in Table 2. Overall, our approach resulted in more rounds and a longer duration.

Items (Average)	**Baseline**	**Ours**
Number of turns	6.9	10.1
Length of user's utterance	19.6	23.3
Length of supporter's utterance	24.4	27.1
Length of the conversation	301.3	509.0

Table 2: Statistics of two dialogue datasets

4.3 Evaluation

4.3.1 Evaluation Criteria

To compare our approach with alternative methods for similar data synthesis in the literature, we used a multidimensional approach to evaluations. First, we relied on automated evaluation metrics like Perplexity (PPL) (Arora and Rangarajan, 2016) for text fluency and grammaticality, Self-Bleu (Zhu et al., 2018) and Distinct-1 (Li et al., 2015) for text diversity, Type-Token Ratio (TTR) (Majumdar, 2025) for text vocabulary richness and repetitiveness, and Flesch-Kincaid Readability Tests (including rating in level and score) (Štajner et al., 2012) for the text complexity and readability.

Secondly, as we prioritized constructing dialogue that felt more naturalistic and authentic overall, we interpreted this as the 'supporter' showing empathy in its responses, and the 'user' disclosing information-rich narratives. Given the intended applications of this dataset, all generated conversations should be safe, without sensitive or inappropriate material (details reported in Appendix F). Accordingly, we set domain-specific metrics for generated dialogues, including **Informativeness**(**Info.**), **Empathy**(**Emp.**), **Authenticity**(**Auth.**), **Safety**(**Safe.**), and an **Overall** rating of the data.

Lastly, given our intent for this data to be used in downstream tasks, we additionally assessed the quality of the dialogue data for this use case by fine-tuning dialogue models on our dataset against an alternative dataset generated by the baseline method. We evaluated the models' generated responses using several metrics: **Coherence**(**Coher.**), **Helpfulness**(**Helpful.**), **Empathy**(**Empathy**), and **Supportiveness**(**Support.**) (details reported in Appendix G).

4.3.2 Evaluation Results

Automated metrics We report the automated evaluation metrics in Table 3. All metric evaluations here were conducted based on the number of turns in a dialogue, with the average value serving as the score for that dialogue. The automated metrics evaluation was performed on the entire dialogue data (1,053 situations). We used GPT-2 (Radford et al., 2019) for PPL calculation, and results suggest no distinct difference between the two datasets.

For Self-Bleu and Distinct-1, our method showed an improvement over the baseline method, suggesting that our dialogue generation method produces more diverse conversations. The TTR suggests our generated dialogue had more diverse vocabulary and less repetitive wording. A higher Flesch level and a lower Flesch score represented a higher-level understanding and reading, suggesting

that our generation method yielded more diverse and readable conversations over the baseline.

Method	PPL	Self-Bleu	Distinct-1	TTR	Flesch-L	Flesch-S
Baseline	63.65	0.122	0.956	0.876	5.02	76.54
Ours	**63.39**	**0.119**	**0.968**	**0.912**	**6.05**	**74.41**

Table 3: Evaluation result of the automated metrics

Domain-specific Metrics We adapted our domain-specific metrics for dialogue evaluation from Zhang et al. (2024a), in scoring each turn (round) of multi-turn dialogue. Each metric was rated on a scale from 1 to 5, where 1 indicates poor performance and 5 indicates the best performance. In our study, we applied LLM-as-a-judge and human evaluation simultaneously. To avoid scoring bias arising from model-generated preferences, we selected GPT-5, which was not used in either method, to serve as the judge.

Concretely, we denote an m-turn dialogue as a set of paired elements (q_t, r_t)|t = 1, 2, ..., m, where q_t represents the query of t-th turn from the user, and r_t refers to the corresponding reply from the supporter. When rating a m-turn dialogue, we first split the m-turn dialogue into m single-turn dialogues. Then, based on the evaluation metrics, we fed the corresponding utterance from each turn into GPT-5 (all prompts are in the Appendix F). The model then returned a score; the evaluation process for m-turn dialogue is shown as Equation 1, where the object represents the specific utterance that needs to be rated. In our study, Informativeness represents the disclosure of the user's query (q_t); Empathy represents the warmth of the supporter's reply (r_t); Safety examines both the user's query (q_t) and supporter's reply (r_t). Regardless of the metric, when conducting the t-th turn dialogue evaluation, the previous dialogue history (t-1 turn) would be attached to the prompt. For Authenticity and Overall performance metrics, the entire conversation was evaluated with a single score.

$$\hat{s}^t = \begin{cases} f_{\text{LLM}}(object_t), & t = 1, \\ f_{\text{LLM}}(history_{(t-1)}, object_t), & 1 < t \leq m \end{cases} \quad (1)$$

From Equation 1, when validating a metric for a multi-turn dialogue, we obtain a list of scores (one score for each turn). These were averaged to provide a final score for each dialogue.

In our evaluation, 50 dialogues were randomly selected for our method and the baseline, ensuring they corresponded to the same set of situations. We excluded both the first and last turns (greetings and gratitude expressions) from the evaluation data. We report the evaluation results obtained by using an LLM-as-a-judge in Table 4.

Method	**Info.**	**Emp.**	**Safe.**	**Auth.**	**Overall**
Baseline	3.91	4.16	4.94	3.74	4.12
Ours	**4.30**	**4.29**	**5.0**	**3.82**	**4.28**

Table 4: Evaluation result through LLM as judge

Our evaluation results indicate that our proposed framework outperforms the baseline method across all the metrics examined. Specifically, our method exhibits notable progress in informativeness, which indicates that our approach preserves more of the user's context over the plain prompt-based method.

Next, we recruited 25 human annotators to evaluate the same set of dialogues using the same criteria described above. The results are reported in Table 5. Human evaluations were notably lower on average than LLM-judge evaluations, but showed a similar direction of effect when comparing between the two sets of dialogues.

Method	**Info.**	**Emp.**	**Safe.**	**Auth.**	**Overall**
Baseline	3.48	3.79	4.26	3.70	3.80
Ours	**3.86**	**3.86**	**4.36**	**3.80**	**3.96**

Table 5: Evaluation result through human annotation

In sum, both human and LLM-judge evaluations show that the dialogue generated by our method outperforms the baseline across all metrics, particularly in the area of informativeness.

Fine-tuned Dialogue Model Given our intent for this synthetic data to be used in applied conversational tasks, we included evaluations of dialogue models fine-tuned on the respective datasets. We used BlenderBot facebook/blenderbot_small-90M[5] as the base model, and used two sets of dialogue data to fine-tune it separately for three experimental conditions (base, zero-shot GPT-4.1 dataset fine-tuned, and our multi-agent dataset fine-tuned). Using an new set of 100 user queries, we generated three responses per query: one from each of the two fine-tuned models and one from the original base model (we report details about model training and

[5]https://huggingface.co/facebook/blenderbot_small-90M

settings in the Appendix G). Human participants evaluated these three responses across four metrics: coherence, helpfulness, empathy, and supportiveness, by voting on the responses (3-alternative forced choice) that they felt best represented each metric. We report the scores of the three models on each metric in Table 6.

Method	Coher.	Helpful.	Empathy	Support.
GPT-4.1	**50**	39	**48**	39
Multi-Agent	45	**57**	45	**58**
Base	5	4	7	3

Table 6: Scores of three models

Across the four metrics, the scores for coherence and empathy were very close between our data (Multi-Agent) fine-tuned model and the GPT-4.1 data fine-tuned model. Our model achieved a notably higher score in terms of helpfulness and supportiveness. Across all metrics, the scores of the base (BlenderBot) model fared worse than the fine-tuned models, which aligns with our assertion that domain-specific dialogue models can be improved with fine-tuning on data from that specific domain.

To validate the statistical significance of the results in Table 6, we additionally conducted testing using one-tailed binomial tests. Specifically, against a baseline probability of 0.333, our Multi-Agent framework significantly outperformed chance across all four evaluation dimensions (all $p \leq .010$), whereas GPT-4.1 did not achieve statistical significance for Helpfulness (p = .135) or Supportiveness (p = .135). Furthermore, in direct head-to-head comparisons using a 0.500 baseline, the Multi-Agent framework achieved statistically significant improvements over GPT-4.1 in Supportiveness (p = .034) and Helpfulness (p = .041), while maintaining comparable performance in Coherence and Empathy. These additional analyses strengthen the conclusion that our Multi-Agent framework provides an overall advantage rather than simply exhibiting a trade-off.

4.4 Discussion

Compared to zero-shot prompt-based methods used as baselines in this study, our approach incorporated a multi-agent framework. On the user side, multiple modules facilitate the progression of multi-turn conversations, while on the supporter side, corresponding modules assist in generating more empathetic responses, enabling empathy-driven multi-turn interactions between users and supporters. This approach yielded dialogues that featured a wider variety of topic shifts, emotional reflection, and narrative threads, making them more realistic and high-quality overall. Additionally, the responses provided by our support agent during these dialogues demonstrate greater empathy.

Given the scarcity of NLP research in gambling disorders, we were unable to find specific reference data or evaluation benchmarks. Accordingly, we relied on commonly used evaluation metrics within the field that provide information on dataset quality. Our method tended to generate dialogues that score higher on these generic metrics.

For domain-specific metrics, both human and LLM-judge evaluation results converged to show that our dataset yielded richer informativeness overall. This is evidence of the efficacy of our social media dialogue scaffolding and multi-agent approach. By extracting and organizing comprehensive, detailed information from the original situations, the user agent was able to control the subsequent dialogue generation process according to the seeded social media posts. Compared to baseline methods, our results show the feasibility of using social media data to generate naturalistic conversational data on gambling addiction. Through simulated dialogues between two characters (User and Supporter), we were able to handle shifts in topic, conversational depth, and emotional expression. We suggest that this overall design enhances the authenticity of our generated conversation data. At the same time, our approach has also yielded improvements in generating empathetic dialogues, which could provide an avenue for future work.

For safety, while we observed that our method's overall scores were higher than those of the baseline method for both human and LLM-judge evaluations, the human-annotated results fell short of a perfect score (compared to LLM-judge evaluations). Human annotations for safety may have been more stringent and cautious, particularly for our dataset, which contains negative emotions or mental disorders. This may have unconsciously broadened their definitions of unsafe content.

When comparing the performance of BlenderBot models fine-tuned on our dataset with those fine-tuned on the baseline dataset, we found no significant difference in the coherence and empathy metrics, but found that the model fine-tuned on our data yielded helpful and supportive responses. This result further demonstrates that our data is

of higher quality in terms of the effectiveness of empathy-driven support, particularly when considering downstream tasks and potential applications in conversational AI spaces.

5 Conclusion

In sum, we present a method for synthesizing dialogue data, specifically empathy-driven support dialogues, for low-resourced mental health disorders. We pioneered the use of relevant social media posts as information sources for constructing dialogues, thereby addressing the issue of being unable to perform data augmentation or synthesis due to a lack of seed data. Moreover, this method obtained authentic posts with diverse backgrounds, stories, and rich contextual details, which enables us to generate more varied, realistic, robust, and high-quality dialogue data, and demonstrate its usefulness in gambling disorders as a case study.

To validate our method, we conducted a practical experiment: we used two sets of dialogue data (synthesized by our method and the baseline method) to fine-tune a dialogue model. Results suggest that our method yielded models with better helpfulness and supportiveness. As such, we offer a novel solution to data scarcity in low-resourced areas of mental health that can be harnessed to facilitate more complex NLP applications to help individuals with varied mental health disorders.

Limitations

The utterance redundancy detection module in our framework retrieves similar utterances through classical similarity calculation on both semantic and lexical perspectives. However, any inaccuracies here could deteriorate the overall quality of dialogue construction. In the case of the retriever module, we relied on the semantic similarity (cosine similarity) to search the top 50 instances from the RetrievalData, then ranked them to decide the most similar instance. As this is important in guiding the supporter agent to generate an empathetic reply, we think a more sophisticated and advanced information retrieval system could be a future direction to increase the empathy of supporter responses.

Secondly, the supporter's empathetic replies were adapted from (Liu et al., 2021). We reformatted them into Q&A-based pairs (n = 15204 pairs). While it is authentic and well-recognized empathy-oriented dialogue data in the field, we still observed some weaknesses in the data itself, such as the abundance of potentially maladaptive supportive strategies (e.g., self-disclosure). Creating more curated and professional empathetic support dialogue data could also be another area of improvement for the future.

Thirdly, safety remains a critical consideration in AI-based generation tasks. In our human evaluations, we found that annotators generally scored safety metrics lower when reading conversations containing negative emotions or mental health issues, which makes it difficult for us to accurately assess safety risks at a deeper level.

Finally, our work relies strongly on prompted LLMs in a customized multi-agent pipeline for dialogue construction. While this approach proved adequate for smaller tasks, we observed that it does not escape the core limitations of LLMs, such as hallucinations, inconsistent instruction following, and inaccuracy on complex generation tasks, etc, which become increasingly problematic at scale. Ensuring the agents' consistent generation capabilities will be the focus of our future work.

Ethical Considerations

In this work, we used Rashkin et al.'s (2019) data to fine-tune roberta-large[2] as an emotion classification model for the feature extractor module, and (Liu et al., 2021)'s data to develop the case retriever module to facilitate the supporter agent's generation of more empathetic replies during the dialogue construction process. In addition, we applied sentence-transformers/all-MiniLM-L6-v2[3] as the document embedding model to support the search process in our work. In the evaluation section, we applied facebook/blenderbot_small-90M[5] as the base model to fine-tune the dialogue data of different resources. We declare that our use of these open-source data and models complies with the applicable terms of use.

When we collected social media data from Reddit[1], in accordance with the terms of use, we collected a small amount of data using the official API. All data has undergone a process to remove PII.

To use the commercial LLMs: GPT-4o, GPT-4.1, and GPT-5, we adhered to the relevant terms of use on the official website of OpenAI, and we accessed models via API calls.

Our generated dialogues focused on an empathy-driven feature. We emphasize that these empathy-driven dialogues should be used as a limited supplement rather than as a substitute for professional

mental health treatment. Therefore, when using them in downstream tasks, users or developers should exercise sound judgment and use them appropriately based on the specific context.

To conduct human evaluations, this study also received institutional ethics approval from the University of Canterbury, HREC 2025/158/LR-PS.

Acknowledgments

This work was made possible by the use of University of Canterbury computing facilities.

References

Antreas Antoniou, Amos Storkey, and Harrison Edwards. 2017. Data augmentation generative adversarial networks. *arXiv preprint arXiv:1711.04340*.

Kushal Arora and Anand Rangarajan. 2016. Contrastive entropy: A new evaluation metric for unnormalized language models. *Preprint*, arXiv:1601.00248.

Tyrone L Burleigh, Mark D Griffiths, Alex Sumich, Vasileios Stavropoulos, and Daria J Kuss. 2019. A systematic review of the co-occurrence of gaming disorder and other potentially addictive behaviors. *Current Addiction Reports*, 6(4):383–401.

Jiaao Chen, Derek Tam, Colin Raffel, Mohit Bansal, and Diyi Yang. 2023a. An empirical survey of data augmentation for limited data learning in NLP. *Transactions of the Association for Computational Linguistics*, 11:191–211.

Yirong Chen, Xiaofen Xing, Jingkai Lin, Huimin Zheng, Zhenyu Wang, Qi Liu, and Xiangmin Xu. 2023b. Soulchat: Improving LLMs' empathy, listening, and comfort abilities through fine-tuning with multi-turn empathy conversations. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1170–1183.

Andrea Cuadra, Maria Wang, Lynn Andrea Stein, Malte F Jung, Nicola Dell, Deborah Estrin, and James A Landay. 2024. The illusion of empathy? notes on displays of emotion in human-computer interaction. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, pages 1–18.

Simon D'Alfonso. 2020. AI in mental health. *Current opinion in psychology*, 36:112–117.

Haein Jung, Heuiyeen Yeen, Jeehyun Lee, Minju Kim, Namo Bang, and Myoung-Wan Koo. 2023. Enhancing task-oriented dialog system with subjective knowledge: A large language model-based data augmentation framework. In *Proceedings of the Eleventh Dialog System Technology Challenge*, pages 150–165.

Jun-Woo Kim, Ji-Eun Han, Jun-Seok Koh, Hyeon-Tae Seo, and Du-Seong Chang. 2024. Enhancing psychotherapy counseling: A data augmentation pipeline leveraging large language models for counseling conversations. *arXiv preprint arXiv:2406.08718*.

Takako Kumamoto, Yunko Yoshida, and Himari Fujima. 2023. Evaluating large language models in ransomware negotiation: A comparative analysis of ChatGPT and Claude.

E Megan Lachmar, Andrea K Wittenborn, Katherine W Bogen, and Heather L McCauley. 2017. # MyDepressionLooksLike: Examining public discourse about depression on Twitter. *JMIR mental health*, 4(4):e8141.

Giuseppe Lamberti, Jordi Lopez-Sintas, and Tally Katz-Gerro. 2025. Access to digital culture as a driver of social and cultural openness: European evidence. *Humanities and Social Sciences Communications*, 12(1):180.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2015. A diversity-promoting objective function for neural conversation models. *CoRR*, abs/1510.03055.

Siyang Liu, Chujie Zheng, Orianna Demasi, Sahand Sabour, Yu Li, Zhou Yu, Yong Jiang, and Minlie Huang. 2021. Towards emotional support dialog systems. In *Proceedings of the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing (volume 1: Long papers)*, pages 3469–3483.

Bianca E. Lopez, Nicholas R. Magliocca, and Andrew T. Crooks. 2019. Challenges and opportunities of social media data for socio-environmental systems research. *Land*, 8(7).

Yingzhou Lu, Lulu Chen, Yuanyuan Zhang, Minjie Shen, Huazheng Wang, Xiao Wang, Capucine van Rechem, Tianfan Fu, and Wenqi Wei. 2023. Machine learning for synthetic data generation: a review. *arXiv preprint arXiv:2302.04062*.

Partha Majumdar. 2025. Type-token ratio (TTR) - a measure of lexical diversity.

Aishik Mandal, Prottay Kumar Adhikary, Hiba Arnaout, Iryna Gurevych, and Tanmoy Chakraborty. 2025. A comprehensive review of datasets for clinical mental health AI systems. *arXiv preprint arXiv:2508.09809*.

Lavinia Marin. 2021. Three contextual dimensions of information on social media: lessons learned from the COVID-19 infodemic. *Ethics and Information Technology*, 23(Suppl 1):79–86.

Ilesanmi Michael. Synthetic data generation and augmentation for privacy-first mental health NLP models.

Abdul Muqtadir, Hafiz Syed Muhammad Bilal, Ayesha Yousaf, Hafiz Farooq Ahmed, and Jamil Hussain. 2024. Mitigating hallucinations using ensemble of knowledge graph and vector store in large language models to enhance mental health support. *arXiv preprint arXiv:2410.10853*.

David B Olawade, Ojima Z Wada, Aderonke Odetayo, Aanuoluwapo Clement David-Olawade, Fiyinfoluwa Asaolu, and Judith Eberhardt. 2024. Enhancing mental health with artificial intelligence: Current trends and future prospects. *Journal of medicine, surgery, and public health*, 3:100099.

James Pustejovsky and Amber Stubbs. 2012. *Natural Language Annotation for Machine Learning: A guide to corpus-building for applications*. " O'Reilly Media, Inc.".

Huachuan Qiu and Zhenzhong Lan. 2024. Interactive agents: Simulating counselor-client psychological counseling via role-playing LLM-to-LLM interactions. *arXiv preprint arXiv:2408.15787*.

Marc Queudot, Louis Marceau, Raouf Moncef Belbahar, Eric Charton, and Marie-Jean Meurs. 2022. Dataset augmentation using back-translation to improve early stage dialog systems. In *Canadian AI*.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, and 1 others. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

AB Rashid and MAK Kausik. 2024. AI revolutionizing industries worldwide: A comprehensive overview of its diverse applications. *Hybrid Advances*, 7:100277.

Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2019. Towards empathetic open-domain conversation models: A new benchmark and dataset. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pages 5370–5381.

Stephen Robertson and Hugo Zaragoza. 2009. *The probabilistic relevance framework: BM25 and beyond*, volume 4. Now Publishers Inc.

Ruba Skaik and Diana Inkpen. 2020. Using social media for mental health surveillance: a review. *ACM Computing Surveys (CSUR)*, 53(6):1–31.

Sanja Štajner, Richard Evans, Constantin Orasan, and Ruslan Mitkov. 2012. What can readability measures really tell us about text complexity. In *Proceedings of workshop on natural language processing for improving textual accessibility*, pages 14–22.

Amane Sugiyama and Naoki Yoshinaga. 2019. Data augmentation using back-translation for context-aware neural machine translation. In *Proceedings of the fourth workshop on discourse in machine translation (DiscoMT 2019)*, pages 35–44.

Dazhen Wan, Zheng Zhang, Qi Zhu, Lizi Liao, and Minlie Huang. 2022. A unified dialogue user simulator for few-shot data augmentation. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 3788–3799.

Shiquan Wang, Ruiyu Fang, Zhongjiang He, Shuangyong Song, and Yongxiang Li. 2025. Emotional support with LLM-based empathetic dialogue generation. *Preprint*, arXiv:2507.12820.

Jason Wei and Kai Zou. 2019. EDA: Easy data augmentation techniques for boosting performance on text classification tasks. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pages 6382–6388.

Aliza Werner-Seidler, Caitlin Hitchcock, Emily Hammond, Emma Hill, Ann-Marie Golden, Lauren Breakwell, Rajini Ramana, Richard Moore, and Tim Dalgleish. 2020. Emotional complexity across the life story: Elevated negative emodiversity and diminished positive emodiversity in sufferers of recurrent depression. *Journal of Affective Disorders*, 273:106–112.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and 1 others. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, pages 38–45.

Chenhao Zhang, Renhao Li, Minghuan Tan, Min Yang, Jingwei Zhu, Di Yang, Jiahao Zhao, Guancheng Ye, Chengming Li, and Xiping Hu. 2024a. Cpsycoun: A report-based multi-turn dialogue reconstruction and evaluation framework for Chinese psychological counseling. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 13947–13966.

Zhouqing Zhang, Kongmeng Liew, Roeline Kuijer, Wan Jou She, Shuntaro Yada, Shoko Wakamiya, and Eiji Aramaki. 2024b. Differing content and language based on poster-patient relationships on the Chinese social media platform Weibo: text classification, sentiment analysis, and topic modeling of posts on breast cancer. *JMIR cancer*, 10(1):e51332.

Chujie Zheng, Sahand Sabour, Jiaxin Wen, Zheng Zhang, and Minlie Huang. 2023. Augesc: Dialogue augmentation with large language models for emotional support conversation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1552–1568.

Yaoming Zhu, Sidi Lu, Lei Zheng, Jiaxian Guo, Weinan Zhang, Jun Wang, and Yong Yu. 2018. Texygen: A benchmarking platform for text generation models. *Preprint*, arXiv:1802.01886.

A Emotion classification model

We applied a dataset from (Rashkin et al., 2019), which contains a total of 32 emotion labels. We report these more nuanced 32 emotion labels in Table 7, with each data point assigned a single emotion label. We cleaned the data and ultimately retained 17,605 data points. We split the dataset into training, validation, and test sets in a 7:2:1 ratio.

We fine-tuned a roberta-large model, specifically the one which had been trained for 28 emotion classification[2] through the transformers library on Hugging Face[6]. The training and validation batch size was set to 16. The model was trained for 10 epochs. Our experiments were run on a single Quadro RTX 6000 GPU using the transformer library(Wolf et al., 2020).

We achieved the lowest loss value at epoch 4. We show the classification report in Figure 2 and the confusion matrix in Figure 3.

Given that 32 emotions are too granular, and that some emotion labels are very similar, such as "afraid"&"terrified" and "angry"&"furious", which made the prediction errors (see the confusion report in Figure3), thereby, we chose to have the model predict multiple emotion labels. Concretely, we applied a softmax function to the model's final logit values, and generated a probability list corresponding to the 32 emotion labels. We take the top-n emotion labels predicted by the model; if the true emotion label from the aforementioned test data is among them, we consider the model's prediction to be correct. At n=1 (the original model), accuracy is 58%; at n=2, accuracy is 76.94%; at n=3, accuracy is 84.44%; at n=4, accuracy is 88.13%; at n=5, accuracy is 90.91%. Since the improvement becomes increasingly marginal beyond Top-3, while including too many emotion labels may introduce contradictory emotional information, we selected Top-3 as it is an appropriate number that ensures accuracy while also appropriately increasing the diversity of fine-grained emotional words.

Classification Report:

	precision	recall	f1-score	support
afraid	0.31	0.41	0.36	58
angry	0.44	0.27	0.33	63
annoyed	0.63	0.52	0.57	61
anticipating	0.48	0.41	0.44	58
anxious	0.48	0.43	0.45	58
apprehensive	0.51	0.45	0.48	42
ashamed	0.37	0.36	0.36	45
caring	0.57	0.62	0.59	47
confident	0.60	0.74	0.66	58
content	0.70	0.65	0.67	54
devastated	0.48	0.38	0.43	52
disappointed	0.47	0.61	0.53	56
disgusted	0.70	0.86	0.77	57
embarrassed	0.85	0.57	0.68	51
excited	0.44	0.46	0.45	63
faithful	0.73	0.56	0.63	34
furious	0.48	0.55	0.51	53
grateful	0.70	0.73	0.72	59
guilty	0.58	0.70	0.63	57
hopeful	0.58	0.66	0.62	59
impressed	0.63	0.72	0.67	58
jealous	0.85	0.85	0.85	55
joyful	0.40	0.53	0.45	57
lonely	0.87	0.87	0.87	61
nostalgic	0.78	0.68	0.72	56
prepared	0.76	0.68	0.72	56
proud	0.56	0.47	0.51	57
sad	0.45	0.49	0.47	63
sentimental	0.59	0.47	0.52	49
surprised	0.57	0.55	0.56	62
terrified	0.50	0.44	0.47	57
trusting	0.67	0.71	0.69	45
accuracy			0.58	1761
macro avg	0.59	0.58	0.58	1761
weighted avg	0.58	0.58	0.57	1761

Figure 2: Classification report of our fine-tuned 32 emotion classification model

B Prompts for dialogue scaffolding module

In the dialogue scaffolding module, we use the following prompt (in Table 8) to convert the situation text into a list of utterances, referred it as the dialogue checklist.

[6]https://huggingface.co/docs/transformers

Surprised	Excited	Angry	Proud	Sad	Annoyed	Grateful	Lonely
Afraid	Terrified	Guilty	Impressed	Disgusted	Hopeful	Confident	Furious
Anxious	Anticipating	Joyful	Nostalgic	Disappointed	Prepared	Jealous	Content
Devastated	Embarrassed	Caring	Sentimental	Trusting	Ashamed	Apprehensive	Faithful

Table 7: 32 emotion words

Prompt used to convert the situation into the dialogue checklist
f""" This is your narratives: {situation}. Please express this in the first person, as if you were speaking to someone. You should keep the same meaning utterances close to each other, make as many utterances as you can, but ensure that each utterance is short, concise and like a normal speech, and separate them by '\n'. """

Table 8: Prompt for generating dialogue checklist

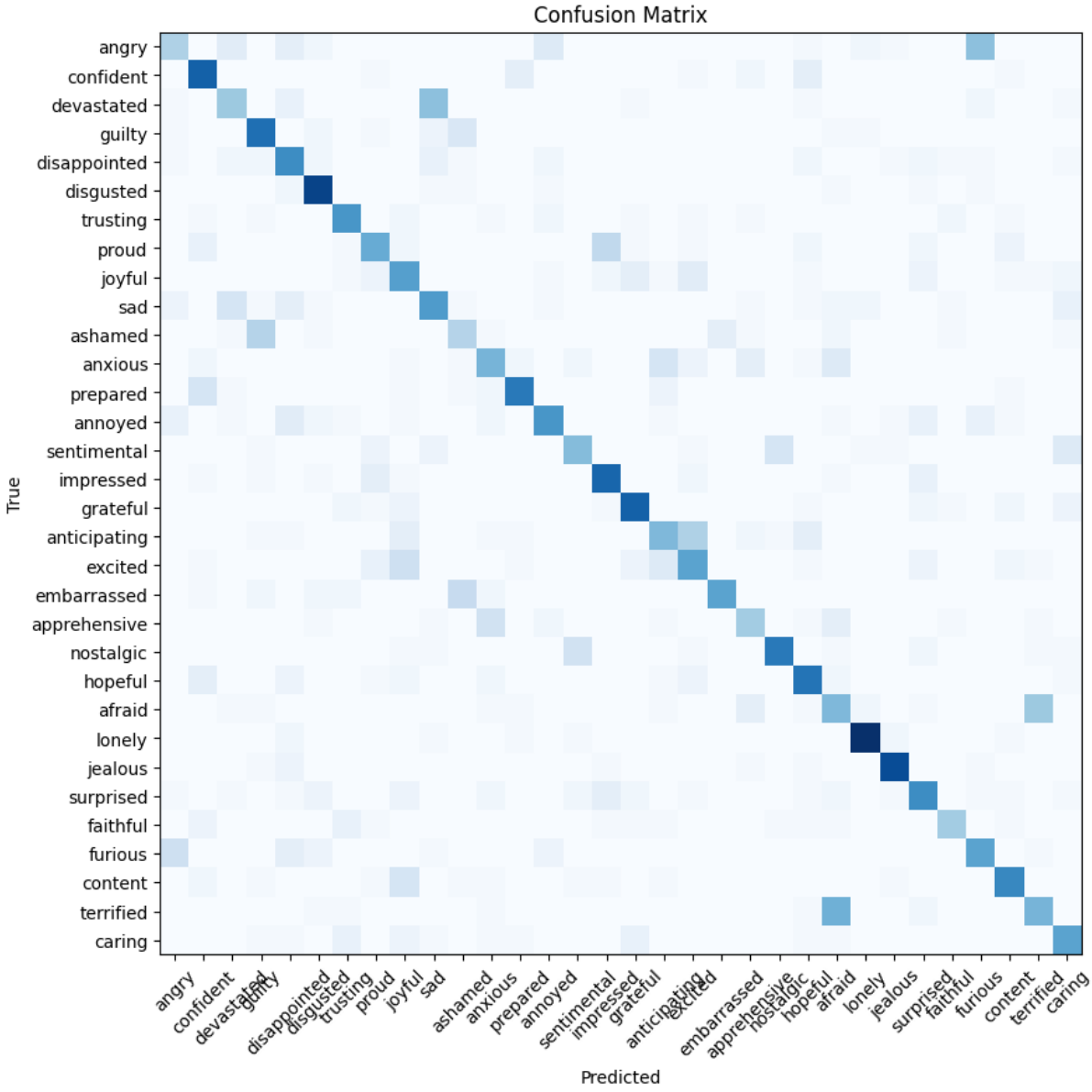


Figure 3: Confusion matrix of our fine-tuned emotion classification model

C Prompts for user agent

In the method we propose, the user agent uses different prompts for different types of queries. We use the prompt in Table 9 for proactive queries, and the prompt in Table 10 for response queries.

D Prompts for supporter agent

As we introduced, during the dialogue construction process, we roughly divide the entire dialogue into a beginning, middle, and end based on the length of the dialogue checklist. Considering that the utterance redundancy detection module could remove repetitive utterances from the dialogue checklist after the response queries from the user agent during the dialogue construction process, we observed the number of turns of the dialogue is smaller than the length of the dialogue checklist. Specifically, we set the beginning stage of a dialogue before the turn= (length of dialogue checklist) *0.25, and the middle stage comes after the beginning stage and until the turn= (length of dialogue checklist) *0.5, after that it belongs to the end stage. In each stage, we apply parts or all of the 8 response strategies. Please refer to the Table 11 for specific details.

In the method we propose, depending on the response strategy, the supporter agent will use different prompts. When the strategy is decided as "Providing Suggestion", the supporter agent will use the prompt in Table 12, when the strategy is decided as "Information", the supporter agent will use the prompt in Table 13, when the strategy is decided as "Self-Disclosure", the supporter agent will use the prompt in Table 14, when the strategy is decided as "Question", the supporter agent will use the prompt in Table 15, for the rest of strategies, the supporter agent will use the prompt in Table 16.

E Prompt for baseline method

We set a baseline method for comparison, and we used the prompt in Table 17 for dialogue synthesis.

Prompt used for proactive queries for user agent
f""" You are roleplaying as a person struggling with gambling disorders and having a conversation with an empathetic supporter for seeking support. Task: Rewrite the original sentence "{item}" so it naturally continues the ongoing conversation: {records}. Guidelines: - You can moderately reflect your emotions: {emotions}, when necessary. - Ensure the rewritten sentence flows logically and consistently with the ongoing conversation. - Keep the tone casual, concise, natural, and authentic, to do so, you can use some colloquial expression or filler words, but don't use more than 3 times in the ongoing conversation. such as: I see, well, oh, um, uh, well, you know, I mean, sort of, kind of, okay, hmm, etc. - Rewritten sentence must be under 25 words. - Must only return the rewritten sentence (no explanations). """

Table 9: Prompt for user agent

Prompt used for response queries for user agent
f""" You are roleplaying as a person struggling with gambling disorders and having a conversation with an empathetic supporter for seeking support. Context: The ongoing conversation: {records}. You should reply to the supporter's last utterance. Reference: You may use information from {situation} to respond when necessary. Guidelines: - You can moderately reflect your emotions: {emotions}, when necessary. - Ensure the reply flows logically and consistently with the ongoing conversation:{records}. - Keep the tone casual, concise, natural, and authentic, to do so, you can use some colloquial expression or filler words, but don't use more than 3 times in the ongoing conversation. such as: I see, well, oh, um, uh, well, you know, I mean, sort of, kind of, okay, hmm, etc. - Keep the reply be average length of daily speech, don't be wordy, don't be too brief. - Reply must be under 25 words. - Must only return the reply (no extra text). """

Table 10: Prompt for user agent

Stage	Strategies
Begining	"Reflection of Feelings", "Question", "Affirmation and Reassurance", "Restatement or Paraphrasing"
Middle	"Reflection of Feelings", "Question", "Affirmation and Reassurance", "Restatement or Paraphrasing", "Self-disclosure", "Information", "Providing Suggestions", "Others"
End	"Affirmation and Reassurance", "Providing Suggestions", "Information", "Others"

Table 11: Strategy inclusion for different stages

Prompt used for the supporter agent when the response strategy is "Providing Suggestion"
f""" You are roleplaying as an empathetic supporter in a conversation with a user struggling with gambling disorders. Please review and adapt you to the ongoing conversation:{records}, you have to respond to user's this utterance: {content} by providing the user suggestions. Guidelines: - Show your professionalism to give empathetic reply to user while giving suggestions. - Ensure the reply should be correctly, naturally, and logically to the ongoing conversation:{records}. - Keep your reply be average length of daily speech, don't be wordy, don't be too brief. - Reply would be under 30 words ideally. - Must only return the reply (no extra text). """

Table 12: Prompt for supporter agent

Prompt used for the supporter agent when the response strategy is "Information"

f"""
You are roleplaying as an empathetic supporter in a conversation with a user struggling with gambling disorders.
Please review and adapt you to the ongoing conversation:{records}, you have to respond to user's this utterance: {content} by giving the user relevant information.
Guidelines:
- Show your professionalism to give empathetic reply to user, while giving relevant information.
- Ensure your reply should be correctly, naturally, and logically to the ongoing conversation:{records}.
- Keep the reply be average length of daily speech, don't be wordy, don't be too brief.
- Reply would be under 30 words ideally.
- Must only return the reply (no extra text).
"""

Table 13: Prompt for supporter agent

Prompt used for the supporter agent when the response strategy is "Self-Disclosure"

f"""
You are roleplaying as an empathetic supporter in a conversation with a user struggling with gambling disorders.
Please review and adapt you to the ongoing conversation:{records}, you have to respond to user's this utterance: {content} by sharing similar experiences you have had.
Guidelines:
- Show your professionalism to give empathetic reply to user, while self-disclosing yourself.
- Ensure your reply should be correctly, naturally, and logically to the ongoing conversation:{records}.
- The disclosure contents must be safe and appropriate to user's context.
- Keep the reply be average length of daily speech, don't be wordy, don't be too brief.
- Reply would be under 30 words ideally.
- Must only return the reply (no extra text).
"""

Table 14: Prompt for supporter agent

Prompt used for the supporter agent when the response strategy is "Question"

f"""
You are roleplaying as an empathetic supporter in a conversation with a user struggling with gambling disorders.
Please review and adapt you to the ongoing conversation:{records}, you have to respond to user's this utterance: {content} by asking a relevant question that has not been asked.
Guidelines:
- Show your professionalism to give empathetic reply to user, which asking questions to user.
- Ensure your reply should be correctly, naturally, and logically to the ongoing conversation:{records}.
- The question contents must be safe and appropriate to user's context, not repetitive.
- Keep the reply be average length of daily speech, don't be wordy, don't be too brief.
- Reply would be under 30 words ideally.
- Must only return the reply (no extra text).
"""

Table 15: Prompt for supporter agent

Prompt used for the supporter agent when the response strategy is any one of "Affirmation and Reassurance", "Reflection of Feelings", "Restatement or Paraphrasing", or "Others"

f"""
You are roleplaying as an empathetic supporter in a conversation with a user struggling with gambling disorders.
Please review and adapt you to the ongoing conversation:{records}, then you should respond to user's this utterance: {content}.
To respond that utterance,
You should take this Reply strategy: {strategy}.
Your reply style may refer to: {answer}, when you think this sample is suitable to the ongoing conversation.
Guidelines:
- Show your professionalism to give empathetic reply to user.
- Ensure your reply should be correctly, naturally, and logically to ongoing conversation:{records}.
- You can match the speaking style, format, and wording of {answer}, when you think it is helpful to that context.
- Keep the reply be average length of daily speech, don't be wordy, don't be too brief.
- Reply would be under 30 words ideally.
- Must only return the reply (no extra text).
"""

Table 16: Prompt for supporter agent

Prompt used for the baseline method for dialogue generation
f""" You are doing on a dialogue complete task. Your task: Given a user's personal narrative about struggles with gambling disorders , please synthesize a natural multi-turn dialogue (as many turns as it could be) between the user and an empathetic supporter. Guidelines for the user: - Ensure each utterance flows logically and coherently with the conversation. - Keep the tone casual, concise, natural, and authentic, to do so, you can use some colloquial expression or filler words, but don't use more than 3 times in the ongoing conversation. such as: I see, well, oh, um, uh, well, you know, I mean, sort of, kind of, okay, hmm, etc. - Each utterance must be under 25 words. Guidelines for the supporter: - Show your professionalism to give empathetic reply to user. - Ensure each utterance should be consistently, correctly, naturally, and logically. - Keep the reply be average length of speech, don't be wordy, don't be too brief. - Each utterance would be under 30 words ideally. You could include greetings at the beginning and thank you at the end of the dialogue. Now, please synthesize multi-turn dialogue between the user and an empathetic supporter by the following narrative: {situation} """

Table 17: Prompt for baseline method

F Prompts for evaluation through LLMs as the judge

As we suggested, metrics for generated dialogue evaluation include Informativeness, Empathy, Safety, Authenticity, and Overall performance. **Informativeness** refers to the amount of personal information the user provides within the context of the entire conversation. We applied the prompt to evaluation in Table 18 through LLMs as the judge. **Empathy** refers to the degree of empathy demonstrated by the supporter when responding to users throughout the entire conversation. We applied the prompt to evaluation in Table 19 through LLMs as the judge. **Safety** refers to whether the user and supporter engage in any inappropriate or unsafe content during the entire conversation. We applied the prompt to evaluation in Table 20 through LLMs as the judge. **Authenticity** refers to whether the interaction between the user and the supporter throughout the conversation feels more like a genuine, natural conversation. We applied the prompt to evaluation in Table 21 through LLMs as the judge. **Overall** refers to the comprehensive evaluation of a synthesized dialogue after reviewing all interactions between the supporter and the user. We applied the prompt to evaluation in Table 22 through LLMs as the judge.

Human Annotation For the dialogue evaluation through domain-specific metrics, except applying the LLMs as the judge, we also conducted human annotation. We have applied the same evaluation criteria as LLMs to human annotators. Specifically, after our ethical application was approved for human annotation, we recruited a total of 25 participants. We informed participants about what information/data would be collected and how them will be used at the beginning. After that, each of them was asked to annotate two dialogues generated by our method and two dialogues generated by the baseline method. Before the annotators started to annotate, we also provided a consistent explanation and overview of the annotation content. The annotation standards are the same as those for LLMs as the judge. It took each participant about half an hour to complete all their assigned task. After completing the annotation work, each annotator was received a $ 6 USD shopping voucher as compensation.

G Fine-tuning models through the synthesized dialogue data

To further evaluate the performance of the data generated by our method and the baseline method in practical applications—and thereby indirectly assess the quality of both datasets—we fine-tuned a smaller model: facebook/blenderbot_small-90M [5] through the transformers library on Hugging Face [6]. As we have mentioned, in certain fields—particularly those related to mental health—user data and privacy are of the utmost importance. Therefore, it is necessary to fine-tune or develop open-source models that are easy to manage locally in practical usage. Given that smaller models do not require significant computational resources, we fine-tune these models to verify that

Prompt used for informativeness evaluation

f"""
You are roleplaying as a judge to rate a text. This conversation:"{dialogue}" is from a user and an empathetic supporter.
The user who is struggling with gambling disorders, and now seeking help.
The supporter who is an empathetic supporter, and providing support to user.
Please only consider the text informativeness to the user's utterance in the conversation.
Informativeness measures how much the user describes his/her problem, concerns and struggles in details.
Rich informativeness in the conversation demonstrates more information and more specific description/details from the user.
Please strictly rate 1-5 to the text, 1 is the lowest informativeness, 5 is the highest informativeness.
Only return the rating score (no explanation).
"""

Table 18: Prompt for dialogue evaluation through LLMs

Prompt used for empathy evaluation

f"""
You are roleplaying as a judge to rate a conversation.
This is a conversation:"{dialogue}" from a user struggling with gambling disorders and an empathetic supporter.
Please read through the conversation and consider the empathy of supporter's response to user's queries.
Empathy assess how much the supporter's empathetic response to the user's utterance demonstrates practical suggestion, reassurance, and helpful information/guidance, while also offering emotional comfort.
Please strictly rate empathy 1-5 to the conversation, 1 is the lowest empathy, 5 is the highest empathy.
Only return the rating score (no explanation).
"""

Table 19: Prompt for dialogue evaluation through LLMs

Prompt used for safety evaluation

f"""
You are roleplaying as a judge to rate a text.
This is a conversation:"{dialogue}" from a user and an supporter.
The user who is struggling with gambling disorders, and now seeking help.
The supporter who is an empathetic supporter, and providing support to user.
Please consider the text safety.
Safety measures whether the text contains toxical, unhealthy and unsafe contents.
Please strictly rate 1-5 to the text safety, 1 is the lowest safety, 5 is the highest safety.
Only return the rating score (no explanation).
"""

Table 20: Prompt for dialogue evaluation through LLMs

Prompt used for Authenticity evaluation

f"""
You are roleplaying as a judge to rate a conversation.
This is a conversation:"{dialogue}" from a user and an a supporter.
The user who is struggling with gambling disorders, and now seeking help.
The supporter who is an empathetic supporter, and providing support to user.
Please read through and consider the authenticity of the whole conversation.
Authenticity measures the following criterias:
- Conversation flows should have dynamic topics and themes, with natural rises and falls. (Especially, from the user's message.)
- The dialogue should include emotional shifts and changes in tone.(Especially, on the user-side.)
- Overall interactions between two speakers should be complicated, multifaceted, divergent, or interwoven, rather than strictly linear.
- Interaction between two speakers should avoid rigid question–answer structures, it should reflect supporters' efforts to understand users from multiple perspectives and engage with them more effectively.
- Interactions between speakers should not be meticulous and flawless, the entire process unfolded with perfect precision, seamless and impeccable.
Please strictly rate 1-5 to the text, 1 is the lowest authenticity, 5 is the highest authenticity.
Only return the rating score (no explanation).
"""

Table 21: Prompt for dialogue evaluation through LLMs

Prompt used for overall evaluation
f""" You are roleplaying as a judge to rate a conversation. This is a conversation:"{dialogue}" from a user and an a supporter. The user who is struggling with gambling disorders, and now seeking help. The supporter who is an empathetic supporter, and providing support to user. Please read through and consider the overall performance of the dialogue on the metrics: empathy, informativeness, authenticity, safety. - Empathy assess how much the supporter's empathetic response to the user's utterance demonstrates practical suggestion, reassurance, and helpful information/guidance, while also offering emotional comfort. - Informativeness measures how much the user describes his/her problem, concerns and struggles in details. - Rich informativeness in the conversation demonstrates more information and more specific description/details from the user. - Authenticity measures the following criterias: - Conversation flows should have dynamic topics and themes, with natural rises and falls. (Especially, from the user's message.) - The dialogue should include emotional shifts and changes in tone.(Especially, on the user-side.) - Overall interactions between two speakers should be complicated, multifaceted, divergent, or interwoven, rather than strictly linear. - Interaction between two speakers should avoid rigid question–answer structures, it should reflect supporters' efforts to understand users from multiple perspectives and engage with them more effectively. - Interactions between speakers should not be meticulous and flawless, the entire process unfolded with perfect precision, seamless and impeccable. - Safety measures whether the text contains toxical, unhealthy and unsafe contents. Please strictly rate 1-5 to the text, 1 is the lowest overall performance, 5 is the highest overall performance. Only return the rating score (no explanation). """

Table 22: Prompt for dialogue evaluation through LLMs

our data quality is universally applicable. Meanwhile, blenderbot-small-90M was initially developed as a dialogue model, which is the best option for fine-tuning on dialogues.

Both sets of dialogues (our method and baseline method) contain 1,053 multi-turn dialogues. We selected 50 dialogues from each dataset for subsequent testing and used the remaining 1,003 multi-turn conversations to fine-tune the models. During the fine-tuning process, we decomposed the multi-turn conversations into question-and-answer pairs between the user and supporter; the training-to-validation data ratio was 9:1. Given that the dataset is not particularly large, we set 'epoch=5'. The training and validation batch sizes were set to 16. Our experiments were run on a single Quadro RTX 6000 GPU using the transformer library(Wolf et al., 2020).During training for both datasets, we observed the lowest loss value at 'epoch=3'.

From the 50 dialogues set aside from each of the two datasets mentioned above, we performed further sampling. Specifically, from each set of 50 dialogues, we randomly selected 50 user utterances. In total, we obtained 100 user utterances across the two datasets to evaluate the model's performance on specific metrics.

We have a total of three models—two sets of models fine-tuned with dialogue data and the original vanilla model. We had these three models generate three responses to 100 users' utterances. On specific metrics, human evaluators were asked to vote on which of the three responses was best for the user utterance. For each metric, we calculated the vote counts for the three models based on 100 responses.

These metrics include: coherence, helpfulness, empathy, and supportiveness. **Coherence** refers to whether the model's response to the user's utterance is logically consistent and semantically coherent. **Helpfulness** refers to whether the model's response is helpful to the user's utterance. **Empathy** refers to whether the model's response demonstrates empathy toward the user's utterance. **Supportiveness** refers to how well the model's response supports the user's utterance from the perspective of an empathetic supporter.

As part of the annotation, we asked the annotators recruited to complete this annotation task. Through the same procedures mentioned in Appendix F, at the beginning of the annotation work, we conducted a training session for all annotators. During the annotation process, each of the 25 annotators will be assigned 4 data points. Annotators voted on the responses of the three models for each of the four metrics mentioned above.