Unlocking LLMs: Addressing Scarce Data and Bias Challenges in Mental Health

Vivek Kumar and **Eirini Ntoutsi**
Research Institute CODE,
University of the Bundeswehr, Munich, Germany
{vivek.kumar,eirini.ntoutsi}@unibw.de

Pushpraj Singh Rajawat
Barkatullah University,
Bhopal, India
psrajawatindia@gmail.com

Giacomo Medda and **Diego Reforgiato Recupero**
University of Cagliari, Cagliari, Italy
{giacomo.media,diego.reforgiato}@unica.it

Abstract

Large language models (LLMs) have shown promising capabilities in healthcare analysis but face several challenges like hallucinations, parroting, and bias manifestation. These challenges are exacerbated in complex, sensitive, and low-resource domains. Therefore, in this work we introduce IC-AnnoMI, an expert-annotated motivational interviewing (MI) dataset built upon AnnoMI by generating in-context conversational dialogues leveraging LLMs, particularly ChatGPT. IC-AnnoMI employs targeted prompts accurately engineered through cues and tailored information, taking into account therapy style (empathy, reflection), contextual relevance, and false semantic change. Subsequently, the dialogues are annotated by experts, strictly adhering to the Motivational Interviewing Skills Code (MISC), focusing on both the psychological and linguistic dimensions of MI dialogues. We comprehensively evaluate the IC-AnnoMI dataset and ChatGPT's emotional reasoning ability and understanding of domain intricacies by modeling novel classification tasks employing several classical machine learning and current state-of-the-art transformer approaches. Finally, we discuss the effects of progressive prompting strategies and the impact of augmented data in mitigating the biases manifested in IC-AnnoM. Our contributions provide the MI community with not only a comprehensive dataset but also valuable insights for using LLMs in empathetic text generation for conversational therapy in supervised settings.

1 Introduction

Motivational Interviewing (MI) is a client-centered, directive method of conversational counselling that enhances an individual's motivation to achieve behavioural change (Miller and Rollnick, 2012). MI helps the clients resolve ambivalence and focus on intrinsic motivations by "strengthening client's belief in their capability" or "providing a supportive environment" to make positive changes (Moyers et al., 2009; Martins and McNeil, 2009; Alperstein and Sharpe, 2016). MI has gained wide attention from the clinical psychology community due to its proven efficacy in catalyzing significant improvements in health behaviours such as reducing alcohol consumption, smoking cessation, dietary modification, substance abuse, and increasing physical activity (Apodaca et al., 2014; Barnett et al., 2014; Catley et al., 2012; Lundahl et al., 2013). In particular, MI have been very effective in interventions where client adherence and commitment are critical to successful treatment outcomes (Hettema et al., 2005; Tavabi et al., 2021). In a nutshell, the core principles of MI, namely, "expressing empathy", "developing discrepancy", "rolling with resistance", and "supporting self-efficacy", are designed to promote a non-confrontational approach that respects client autonomy and facilitates self-directed change (Moyers and Rollnick, 2002). Since MI is an interactive and time-intensive process, it is accessible to only a small population group, and the reasons account for "*individual's awareness towards mental health*", "*cost of intervention*", "*lifestyle constraints*", and so on. According to World Health Organization report[1], one in every eight people in the world live with a mental disorder and over half (**54.7%**) of adults with a mental condition do not have access to effective treatment, summing up over 28 million individuals (Organization, 2022; Reinert et al., 2021).

Hence, to overcome these challenges and break the barriers in catering to essential and effective treatment, recent research has focused on artificial intelligence (AI) applications. In particular, Large

[1] https://www.who.int/news-room/fact-sheets/detail/mental-disorders

Language Models (LLMs) have been recognised as a potential solution to alleviate the burden on clinicians (Tripathi et al., 2024; Wang et al., 2023; Yu et al., 2023). Undoubtedly, LLMs can be instrumental in tackling a wide range of problems directly or by means of assisting roles (Stella et al., 2023; Shiffrin and Mitchell, 2023). However, due to its specialised nature, the mental health domain poses unique challenges of complex language understanding that question LLMs efficacy (Demszky et al., 2023; Abramski et al., 2023). Empirical studies have delineated that in such complex domains, LLMs are prone to severe performance concerns like hallucinations (Li et al., 2023a; Sarkar, 2023), stochastic parroting nature (Bender et al., 2021; Duan et al., 2023), and biases (Yeh et al., 2023).

Therefore, this study aims to bridge this gap by addressing the scarce data and bias challenges in low-resource domains, such as mental health, by generating plausible synthetic data. In this context, we leverage LLMs, particularly ChatGPT and novel prompting strategies, to generate in-context (Brown et al., 2020; Chen et al., 2022; Dong et al., 2022) MI dialogues, considering whole therapeutic conversations at once. Furthermore, we develop an evaluation scheme adhering to the Manual for the Motivational Interviewing Skill Code (MISC) (Miller et al.) to assess the quality of generated MI dialogues by comprehensively touching down the psychological and linguistic dimensions. Moreover, we model a novel classification task to identify high- and low-quality MI dialogues. This setting is used to evaluate ChatGPT in terms of domain intricacies understanding, emotional reasoning ability, and biases (contextual, sampling, class imbalance) originated from the experimental dataset. Finally, we discuss the risks of unsupervised employment of LLMs in healthcare, emphasizing the need for collaboration with domain experts and human supervision to ensure responsible LLM implementation across healthcare settings. To put in perspective, our contributions are summarised below:

- **Tailored prompting approach**: We propose progressive prompt-based augmentation techniques using LLMs to generate in-context MI dialogue.
- **Expert annotation**: We develop a rigorous annotation scheme covering psychological and linguistic aspects (e.g., language comprehension, MI structure, false semantics change, contextual reasoning) of generated data grounded on MISC to propose the **IC-AnnoMI** dataset.
- **Model performance evaluation**: We perform extensive experiments with CML and state-of-the-art (transformer) approaches on the **IC-AnnoMI** dataset to (i) provide a broad set of baselines for the adopted task, (ii) assess the quality of **IC-AnnoMI**, and (iii) discuss potential risks and dangers of unsupervised use of LLMs in sensitive domain.
- **Reproducibility**: We publicly[2] provide **IC-AnnoMI** and the source code used for our experiments to contribute to the low resource domain and facilitate further research.

The rest of the paper is organised as follows. Section 2 presents the existing research on LLMs in healthcare. Section 3 presents the data augmentation, MISC annotation, and the dataset creation. Section 4 provides the problem statement and experimental design. Section 5 outlines our experimental setting and results. Section 6 addresses the implications of our study and opens up future research directions. Finally, the limitations section discusses the limitations of our work.

2 Related work

In this section, we introduce the works focused on developing public datasets to assist research into psychology and highlight the biases affecting LLMs.

2.1 Data scarcity in mental health domain

Domains like psychology and its sub-domains suffer from the scarcity of publicly available resources (datasets) that could be instrumental in mitigating bias in ML approaches and enforcing responsible and ethical AI (Wu et al., 2021). This problem has gained traction, and researchers have periodically attempted to bridge this gap by developing publicly available datasets. Early efforts in this direction can be credited to (Pérez-Rosas et al., 2016), where they released a dataset annotated with ten counselor behavioural codes of 22,719 counselor utterances extracted from 277 MI sessions. Subsequently, (Wu et al., 2022, 2023) released AnnoMI, an expert-annotated GDPR-compliant dataset of 133 high- and low-quality MI sessions. While some of the existing works used AnnoMI to model different tasks (Kumar et al., 2023b) and produce synthetic data (Kumar et al., 2023a; Kumar. et al.,

[2] https://github.com/vsrana-ai/IC-AnnoMI

2023), some research used it to create further new datasets (Hoang et al., 2024). Another study (Welivita and Pu, 2022) released a useful, publicly available dataset of social forums annotated by experts at the therapist statement level with labels adapted from the MITI code (Moyers et al., 2014). (Yan et al., 2022) released ØurResources, a dataset containing 96,965 conversations between doctors and patients, covering 843 types of diseases, 5,228 medical entities, and 3 specialties of medical services across 40 domains. Other notable works in related subdomains contributed with datasets based on textual and conversational settings (Sosea and Caragea, 2020; Buechel and Hahn, 2017; Bostan and Klinger, 2018; Bostan et al., 2020; Demszky et al., 2020; Balloccu et al., 2024).

2.2 Large language models application and challenge

LLMs could aid healthcare not only in the workplace but also in enhancing AI systems employed in healthcare. Several studies leveraged LLMs to generate synthetic data to augment the information fed to another model during training (Li et al., 2023c; Cai et al., 2023; Wozniak and Kocon, 2023; Chowdhury and Chadha, 2024). A few clinical works explored this methodology and reported satisfying results (Yuan et al., 2023; Tang et al., 2023; Li et al., 2023b). For instance, (Tang et al., 2023) used LLMs to augment the data for patient-trial matching tasks, while (Li et al., 2023b) proved that LLM-generated data can improve the automatic detection of signs related to Alzheimer's disease from EHRs. Despite the positive aspects of LLMs, researchers have recently pointed out potential threats associated with using these powerful systems. One of the most concerning factors is the bias in the outcomes of LLMs and AI systems (Wan et al., 2023; Morales et al., 2023; Badyal et al., 2023), especially when such systems are employed in clinical contexts (Smith et al., 2024; Giovanola and Tiribelli, 2023; Kumar et al., 2023b). In addition to the prevalent biases such as gender and racial biases, which can lead to misclassifying dosing based on patient ethnicity (Syn et al., 2018) or favoring certain ethnic groups in determining patients-in-need priority scores (Giovanola and Tiribelli, 2023), selection and cultural biases are also critical issues (Navigli et al., 2023). These biases can lead to skewed predictions and recommendations, potentially marginalizing minority groups and exacerbating healthcare disparities.

3 Data Augmentation, MISC annotation and dataset creation

In this section, we describe (i) the data augmentation strategy, (ii) how the MISC annotation scheme is developed, and (iii) how the annotation scheme was used to create the dataset. For ease of understanding, Table 1 outlines the notation used throughout the paper and Figure 1 depicts the process for the development of the **IC-AnnoMI** dataset.

Table 1: Notations and descriptions/definitions

IC-AnnoMI	The dataset built upon AnnoMI by generating in-context MI dialogues using LLMs progressive prompting.
$Client_{utt.}$	The client utterances in MI dialouges.
$Therapist_{utt.}$	The therapist utterances in MI dialogues.
$MI_{org.}$	The original MI sessions from AnnoMI dataset.
$MI_{syn.}$	The generated MI dialogues in **IC-AnnoMI**.
MI_{psych}	The parameter representing the psychological aspect of the annotation scheme.
$MI_{linguist}$	The parameter representing the linguistic aspect of the annotation scheme.

3.1 Augmentation

The increased quantity of data does not necessarily result in a reliable machine learning (ML) system. Plausible synthetic data can help mitigate inherent biases of experimental datasets such as sampling, contextual, and class imbalance to address the scarce data challenges comprising ML models' reliability. Target augmentation not only provides a better distribution of underrepresented classes but also helps the ML model generalise well. In this research, our primary focus has been context-based augmentation through tailored prompting of ChatGPT variants (4.0 and 3.5 Turbo)[3]. The prompts are engineered through the progressive refinement feedback loop (Song et al., 2023; Reynolds and McDonell, 2021; Su et al., 2023) until the desired

[3] https://platform.openai.com/docs/models/overview

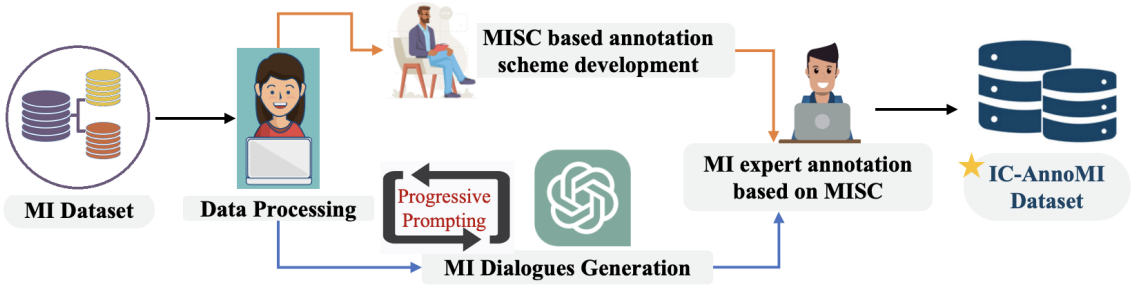


Figure 1: Development of the IC-AnnoMI dataset.

quality and predefined output format are met. In the first step, a prompt template is developed based on MI dialogues' context, plausibility, and quality for required outputs. Then, the generated output is manually evaluated for inconsistencies, and any deviation from the predefined output is used to tune the prompt further progressively. This process continues until the prompt output quality is comparable with $MI_{org.}$. For ease of understanding, an example of "initial" and "final" prompt is shown in Figure 2. Also, to give comparative insights, a sample of $MI_{org.}$ and $MI_{syn.}$ is provided in (Appendix A).

Initial Prompt

Consider you are a psychologist and linguist. You are given a conversational therapy session between the Therapist and the Client. Using paraphrasing and style transfer, generate augmented data with more context for the session provided in the input file "MI.csv" in the same client and therapist format with precisely the same number of client and therapist utterances

Progressive Construction of Prompt

Refined Prompt

Consider you are a psychologist and linguist. You are given a conversational therapy session between the Therapist and the Client in and input file "MI.csv". Use paraphrasing, professionalizing, style transfer and advance text augmentation methods to generate augmented data with more context for each client and therapist utterances. Parse the input file using the sep = ";"

Use the below-mentioned additional set of instructions :
1. The output text should not be the same as the original text utterances.
2. Do not translate the text to another language and translate it back to English.
3. Do not change the original conversational dialogue to an entire "question-answer" format.
4. Do not deviate from the original context and meaning of the conversational therapy.
Return a downloadable "MI.csv" file after generating the augmented data.

Figure 2: Progressive prompt refining.

3.2 MISC annotation

The annotation scheme is developed and executed by an expert from gold-standard institute in psychology by strictly adhering to the MISC 2.1[4] scheme. The developed annotation scheme is a combination of a two-stage annotation process. The first stage of annotation (MI_{psych}) covers the psychological dimension of the generated MI dialogues. The second stage ($MI_{linguist}$) covers the linguistic dimension of MI dialogues. The components of MI_{psych} are further explained as follows.

1. **Empathy**: It is one of the core components of MI and is essential for building rapport and understanding the client's perspective. MI emphasises the therapist's ability to demonstrate empathy through active listening, reflective statements, and genuine curiosity about the client's experiences and feelings (Miller et al.; Miller and Rollnick, 2012).

2. **Non-judgmental attitude**: MI encourages therapists to adopt a non-judgmental stance, accepting the client without criticism or a negative attitude. This attitude creates a safe and supportive environment where clients feel comfortable exploring their ambivalence and concerns, which are better captured by a five-point Likert scale.

3. **Competence of therapist**: Competence is the therapist's proficiency in applying MI techniques and principles effectively, and it is endorsed by the therapist's experience proven through academic certification and licences (Gaume et al., 2009).

4. **Ethical conduct**: In MI practice, ensuring that therapists prioritise the client's well-being, autonomy, and confidentiality is paramount. MI adheres to ethical guidelines established by professional organisations and regulatory bodies such as APA, RCI, etc. These guidelines give

[4] https://digitalcommons.montclair.edu/cgi/viewcontent.cgi?article=1026&context=psychology-facpubs

the clients autonomy and make sessions more comfortable. Ethical considerations are integral to building trust and maintaining the therapeutic alliance in MI. We follow APA, HIPPA, and other guidelines based on country/region.

5. **Reflectiveness**: It involves the therapist's ability to carefully consider and respond to the client's statements, exploring underlying motivations and values. MI encourages therapists to engage in reflective listening and evoke client self-awareness through strategic questioning, which may also include frequent summarisation. Reflective practice enhances the depth and effectiveness of MI interventions, facilitating the meaningful exploration of ambivalence and motivation for change in client sessions.

We have chosen the five-point Likert scale for MI_{psych} annotation because clients can express ambivalent differences in their perceptions, providing more detailed feedback than scales with fewer response options and rather more easily compared with more fine-grained ten-point Likert scale. Indeed, the five-point Likert scale minimises confusion and response errors, facilitating quantitative analysis in terms of mean, standard deviation, and other statistical measures for response summarisation. Compared with ten-pointer scales, converting subjective judgments into five categories enables a clearer alignment with the client's responses and provides sufficient scope to distinguish among different levels of empathy, non-judgmental attitude, competence, ethical conduct, and reflectiveness. MI_{psych} is a numeric value (0-4) averaged over the aforementioned 5 components of MI_{psych} assigned to each $MI_{syn.}$. The components of $MI_{linguist}$ are binary and can acquire either "Yes" or "No", and these components are briefly mentioned below.

1. **`Context`**: It represents the contextual coherence in $MI_{syn.}$ w.r.t. $MI_{org.}$.

2. **`Text Enrichment`**: It indicates if $MI_{syn.}$ is enriched due to style transfer, change in sentence structure, or if more context is added w.r.t. $MI_{org.}$.

3. **`MI Enhancement`**: It represents if text enrichment and contextual addition has overall enhanced the $MI_{syn.}$ w.r.t. $MI_{org.}$.

4. MI_{lang}: It measures if the diction and tone of $MI_{syn.}$ is preserved and language is refined but avoiding any deviation or false semantic change w.r.t. $MI_{org.}$.

3.3 Dataset creation

For data augmentation, we have used our `AnnoMI` (Wu et al., 2023), a publicly available expert-annotated dataset of 133 high- and low-therapeutic counselling dialogues to generate $MI_{syn.}$. First, we have filtered out a representative set of $MI_{org.}$ from `AnnoMI` considering the high- and low-quality and topic-based distribution of $MI_{org.}$, to develop a universal test set for all of our experiments avoiding data contamination. We note that the filtering is done at the MI dialogue level and not at the utterance level to align with our goal of in-context data augmentation, which requires the whole MI dialogue and not the fragments of multiple MI dialogues. This trade-off setup has resulted in 36 $MI_{org.}$ that constitute the representative test set for our experiments. The remaining 97 $MI_{org.}$ of `AnnoMI` constitute the training set and basis of augmentation and MISC annotation. To create **`IC-AnnoMI`** dataset, the 97 $MI_{org.}$ of the training set undergo an augmentation process followed by expert annotation using our developed MISC coding scheme. The annotation process overall results in 97 expert-annotated augmented MI dialogues ($MI_{syn.}$), containing 4,856 $Therapist_{utt.}$ and 4,792 $Client_{utt.}$ having a mix of high and low-quality MI dialogues.

4 Problem Statement and experimental design

This section presents the problem statement and the research questions we aimed to answer through this research, followed by the dataset description, the applied preprocessing strategies, and the evaluation setup to conduct the experiments.

4.1 Problem statement

In this work, we primarily focus on classifying high- and low-quality MI dialogues comprised of talk turns between client and therapist at the utterance level, making it a binary classification problem. Therefore, for given $Client_{utt.} \in (MI_{org.}, MI_{syn.})$ and $Therapist_{utt.} \in (MI_{org.}, MI_{syn.})$, the goal is to infer a classification function f_c so that $f_c\ (Client_{utt.}, Therapist_{utt.}) \rightarrow MI_{quality}$. Here, $MI_{quality}$ is the binary class that can only acquire values in $\{0, 1\}$. The task is designed to evaluate the quality of $MI_{syn.}$, the efficacy of LLMs in

in-context text generation, and address the below-mentioned research questions.

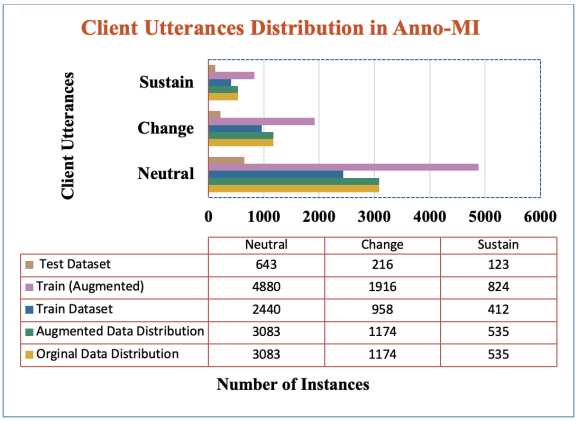
RQ(1): How and to what extent do contextual cues and domain-specific prompting strategies help generate real-like MI dialogues?

RQ(2): Can LLMs be used as a potential tool to generate plausible data, considering the whole therapeutic dialogue at once?

RQ(3): How effective is ChatGPT in understanding the complexity of MI dialogues and what are the risks associated with LLMs' employment in sensitive domains?

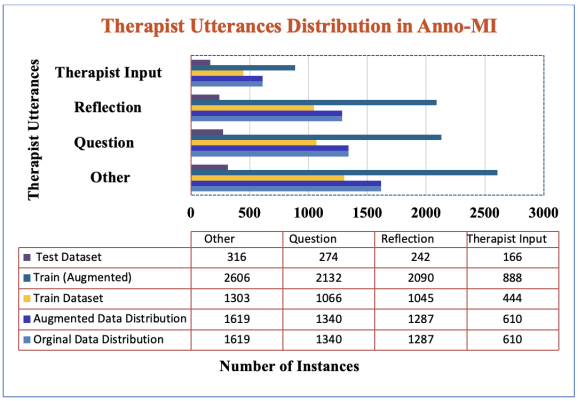
4.2 Dataset preprocessing

As it can be understood from Figure 3 and Figure 4, **IC-AnnoMI** has a skewed distribution over target class "high" and "low" quality MI. Also, several MI dialogues have short sentence length in $Client_{utt.}$, $Therapist_{utt.}$, which makes the task more challenging considering the complexity and the small number of MI dialogues.



	Neutral	Change	Sustain
Test Dataset	643	216	123
Train (Augmented)	4880	1916	824
Train Dataset	2440	958	412
Augmented Data Distribution	3083	1174	535
Orginal Data Distribution	3083	1174	535

Figure 3: The distribution of client utterances in training and test sets of IC-AnnoMI dataset.



	Other	Question	Reflection	Therapist Input
Test Dataset	316	274	242	166
Train (Augmented)	2606	2132	2090	888
Train Dataset	1303	1066	1045	444
Augmented Data Distribution	1619	1340	1287	610
Orginal Data Distribution	1619	1340	1287	610

Figure 4: The distribution of therapist utterances in training and test sets of IC-AnnoMI dataset.

Therefore, we have applied tailored preprocessing strategies to avoid semantic loss in $Client_{utt.}$, $Therapist_{utt.}$ and MI dialogue (Dessì et al., 2020; Kumar et al., 2021; Uysal and Gunal, 2014; Kumar et al., 2023c). The preprocessing steps include lowercasing the text for uniform representation (e.g., Psychology and psychology have a common representation → psychology). We have removed punctuation, whitespaces, newlines, and extra space,s but retained stopwords. This design choice relies on the fact that MI dialogues in **IC-AnnoMI** have several short $Client_{utt.}$, $Therapist_{utt.}$, up to 3 tokens length. Thus, removing stopwords (e.g., not) might change the whole course of the conversation, contributing to misclassification errors. We have also removed multilingual symbols, special characters, elements not part of the standard English language, and expanded contractions such as *it's* –> *it is*.

4.3 Experiments

We have employed various classification models for our experiments, including CML and transformer-based models, to provide a baseline and optimal experimental setup for such task in therapeutic settings. In CML, we have used Support Vector Machine, Naive Bayes, and Random Forest. In deep learning (DL), we used a BiLSTM-based deep neural network architecture with pre-trained word embeddings[5] for feature representation. For transformer-based models, we have employed $BERT_{base}$ (Devlin et al., 2019), and some of its variants, such as *DistilBERT* (Sanh et al., 2019), *RoBERTa* (Liu et al., 2019), *AlBERT* (Lan et al., 2019), *BART* (Lewis et al., 2020), and *XLnet* (Yang et al., 2019), using python libraries such as *Keras*[6], Tensorflow[7], and ML platforms like Hugging Face[8]. The metrics used to evaluate the performance of implemented ML models are accuracy, balanced accuracy, precision, recall and F1-Score and the formulas are provided in (Appendix B). The training, validation and test distribution for all the experiments are 63%, 10%, and 27% respectively, and the computational resource used to conduct the experiments is mentioned in (Appendix C).

[5] https://code.google.com/archive/p/word2vec/
[6] https://keras.io/
[7] https://tfhub.dev/google/collections/bert
[8] https://huggingface.co/docs/transformers/index

Emb.	Model	Acc.		Bal. Acc		Precision		Recall		F1-Macro	
		N-Aug.	Aug.	N-Aug.	Aug.	N-Aug.	Aug.	N-Aug.	Aug.	N-Aug.	Aug.
NA	Naive Bayes	.80	.83	.49	.50	.83	.83	.80	.83	.81	.83
	Random Forest	.89	.89	.51	.50	.84	.84	.89	.90	.86	.86
Static	BiLSTM (word2vec)	.87	.87	.50	.50	.83	.83	.87	.87	.85	.85
Contextual	$BERT_{base}$	.89	.90	.54	.56	.86	.87	.89	.90	.87	.88
	BART	.87	.89	.54	.57	.86	.86	.86	.89	.87	.87
	DistilBERT	.89	.89	.55	**.59**	.86	.87	.89	.89	.87	.88
	AlBERT	.89	.90	.52	.55	.85	.87	.89	.90	.87	.88
	RoBERTa	.88	.90	.54	.57	.86	.86	.88	.90	.87	.87
	XLNet	.88	.88	.54	.57	.85	.86	.88	.88	.86	.87

Table 2: The results of CML and DL approaches with the non-augmented (N-Aug) and augmented (Aug) dataset.

5 Result and discussion

In this section, we provide insights from our results and in-depth analyses based on our experimental outcomes. The classification results of the implemented ML models with the non-augmented and augmented **IC-AnnoMI** datasets are summed up in Table 2.

Note that the applied augmentation method is not centered on reducing the class imbalance in the experimental dataset by targeting the minority class, which is **low-quality** MI in our case, but on preserving the context of each dialogue. Therefore, this augmentation is not expected to contribute significantly to applied ML models' performance, but to have more of an impact on increasing the sample size of the training set. The main experimental observations are as follows:

- **Performance of CML models**: The CML models trained on 2,456 features have shown to be ineffective in accurately identifying the high- and low-quality MI, with a high misclassification rate towards the minority class, as evident from the confusion matrices shown in Figure 6 as expected. The reason is that the features selected in the bag-of-words approach are given weightage based on occurrence frequency, which in complex domains do not sufficiently capture the context of the entire MI dialogue.

- **Performance of DL (BiLSTM) model**: The DL model has also not shown much improvement over CML models due to the fact that the text length of utterances is small, the dataset is very imbalanced, and the number of training MI samples are far too less for a DNN based model to learn and generalise well for such complex domain.

- **Performance of $Bert_{base.}$ and its variants**: This is where the advantage of augmentation reflects. All the language models (LMs), namely $Bert_{base.}$, BART, DistilBERT, ALBERT, RoBERTa, and XLNet, have shown improvement in the performance. In particular, the increase in balanced accuracy is indicative of better generalisation and mitigation of inherent bias in **IC-AnnoMI**. Although all the models have comparable scores in terms of balanced accuracy, DistilBERT has scored the highest, which is **0.59**. A comparative insight through confusion matrices is presented in Figure 5. The observed improved performance in employed LMs verifies that the quality of $MI_{syn.}$ is in line with $MI_{org.}$.

- **Performance based on expert evaluation**: The statistics of expert annotated components of MI_{psych} and $MI_{linguist}$ of $MI_{syn.}$ are also in agreement with the above performance, which strengthens our results. For instance, MI_{psych} has received an average score of **3.31** for the 97 $MI_{syn.}$ averaged over its five attributes and then averaged over 97 $MI_{syn.}$. Also, for the $MI_{linguist}$ aspect of 97 $MI_{syn.}$, **95.88%** have preserved the **Context**, **83.51%** have contributed to **Text Enrichment**, **MI Enhancement** is observed in **88,66%** and overall MI_{lang} is **88,66%**.

- **Answer to the research questions**: These high scores of MI_{psych} and $MI_{linguist}$ are answers to research questions **RQ(1), RQ(2)** and **RQ(3)**. The experimental outcomes indi-

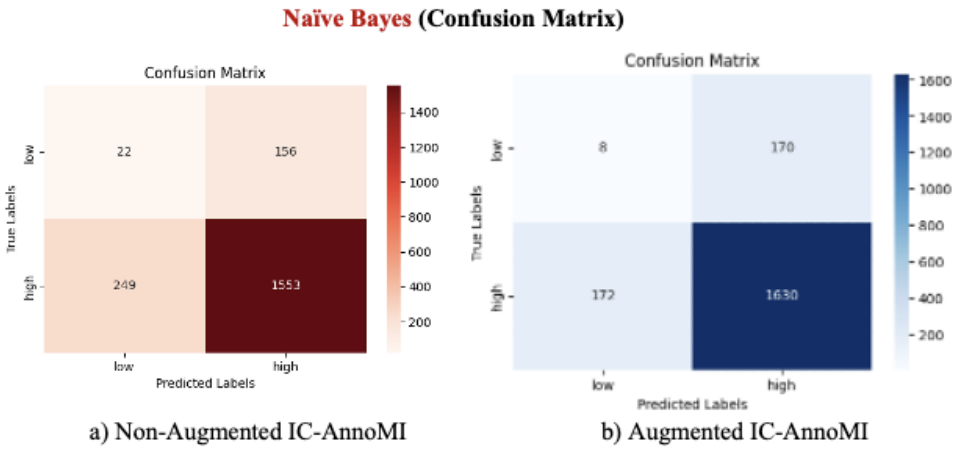


Figure 5: The confusion matrix of CML approaches for non-augmented and augmented experimental datasets.

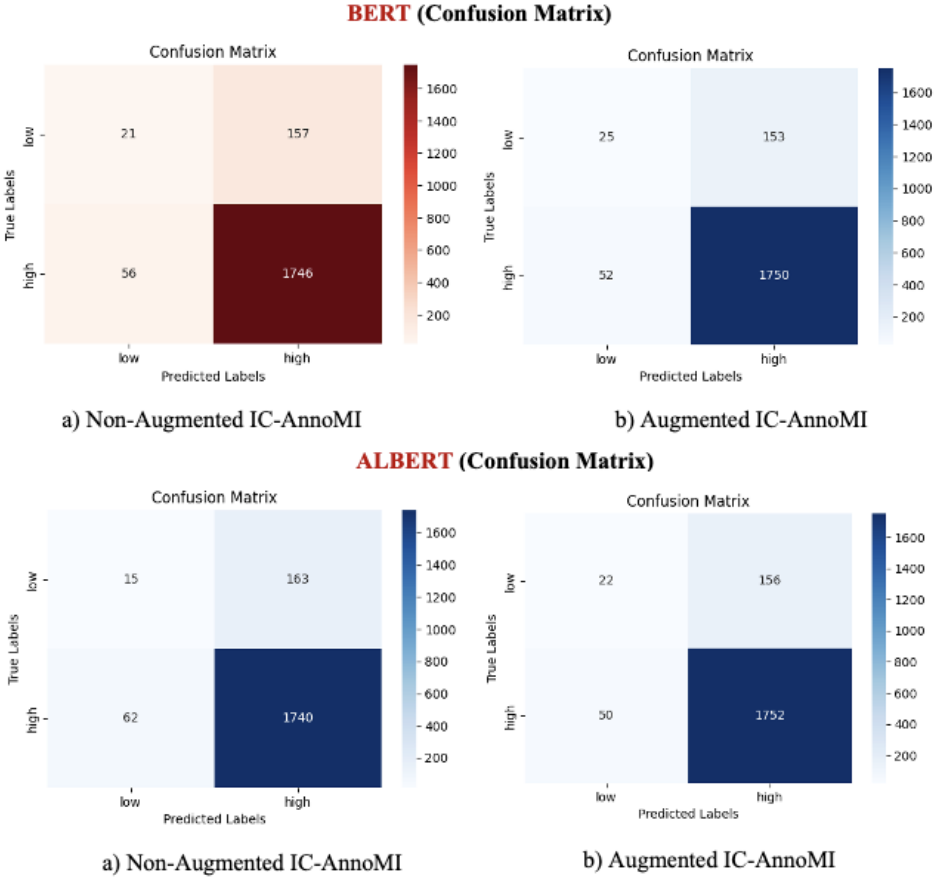


Figure 6: The confusion matrix of BERT model-based approaches for non-augmented and augmented experimental datasets.

cate that contextual cues and domain-specific prompting strategies can help generate dialogues qualitatively close to $MI_{org.}$. LLMs, in our case, ChatGPT, are considerably successful in understanding the fine-grain intricacies of MI and comprehending the flow, context, and nuances of therapeutic settings. However, we also observed inconsistencies in this experimental process at the stage of prompt designing, when hallucinations, absurd text generation, and stochastic parroting happened until they were humanly identified and elimi-

nated through rigorous prompt refining.

6 Conclusion and future work

This paper explores LLMs' capabilities, particularly ChatGPT, for data augmentation in mental health and therapeutic counselling scenarios. Through this research, we seek to study the operability of LLMs in solving the data scarcity issue in therapeutic counselling and verify that biases are not reinforced when models are trained on LLM-generated synthetic data. To this end, we employed a progressive prompt technique to generate in-context plausible MI dialogues and further expert annotated them by developing a comprehensive MISC coding scheme considering MI sessions' psychological and linguistic aspects. To evaluate the quality of generated MI dialogues and to understand to what extent the generated dataset is relevant to the annotation scheme, we employed several CML and transformer-based models to establish a baseline for the classification task of MI dialogues' quality at the utterance level. Our results highlight the efficacy of the augmentation and annotation scheme, given that the augmented dataset led to improvements in classification and mitigation of inherent biases. The findings demonstrate that the data generated through this rigorous quality control process is both plausible and substantially beneficial in enabling ML techniques to address the targeted biases, thereby supporting the use of LLMs for supervised, task-specific applications in sensitive domains like mental health. However, despite the favorable outcomes, risks and concerns are associated with the unsupervised application of LLMs in sensitive domains, and it is thus advised to use them with humans in the loop to promote responsible and ethical AI uses. The future research direction is set to explore other LLMs such as Mistral (Karamcheti et al., 2021), Falcon (Almazrouei et al., 2023), LLama (Touvron et al., 2023), etc., to understand their reliability in mental health domain and plausible data generation. We also aim to tackle MI dialogue-based classification instead of utterance-based and integrate domain knowledge (Kumar et al., 2022) in classification systems generated by LLMs to tackle domain adaptation problems.

Limitations

While our work provides a holistic novel annotation scheme adhering to MISC to create and annotate synthetic MI dialogues, covering both the psychological and linguistic dimensions, it has some limitations and room for improvement. The main limitation can be considered as the low number of MI sessions, which may lead to sub-optimal performance and biases in ML approaches. Another limitation is the computational resource that may have hampered the LMs from being used at their full potential. So we consider using larger resources to avoid this limitation. In this work our focus is in-context dialogue MI generation at the session level that necessarily reduces the class imbalance. Therefore, we aim to generate MI dialogues targeting underrepresented classes leveraging different LLMs to be in more contextual diversity.

Ethics statement

6.1 Expert Annotation

To maintain the integrity and quality of the data, qualified experts affiliated with the gold-standard organization in psychology have performed the annotations. The experts have significant experience and training in MI to ensure therapy's nuances and ethical considerations are appropriately enforced in the annotation process. The expert is also bound by confidentiality agreements to safeguard the privacy of the individuals in the MI recordings and transcripts.

6.2 Ethical Concerns

We acknowledge that our work has strictly followed the norms and protocols of ethical considerations throughout the research process. We also enforce adherence to ethical standards and guidelines for researchers who want to use our data to ensure ethical and responsible use of the resource.

Acknowledgements

This research work is funded by the European Union Horizon Europe Project STELAR, Grant Agreement ID: 101070122.

References

Katherine Abramski, Salvatore Citraro, Luigi Lombardi, Giulio Rossetti, and Massimo Stella. 2023. Cognitive network science reveals bias in gpt-3, gpt-3.5 turbo, and gpt-4 mirroring math anxiety in high-school students. *Big Data and Cognitive Computing*, 7(3).

Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, Alessandro Cappelli, Ruxandra Cojocaru,

Mérouane Debbah, Étienne Goffinet, Daniel Hesslow, Julien Launay, Quentin Malartic, et al. 2023. The falcon series of open language models. *arXiv preprint arXiv:2311.16867*.

Dion Alperstein and Louise Sharpe. 2016. The efficacy of motivational interviewing in adults with chronic pain: A meta-analysis and systematic review. *The Journal of Pain*, 17(4):393–403.

Timothy R Apodaca, Brian Borsari, Kristina M Jackson, Molly Magill, Richard Longabaugh, Nadine R Mastroleo, and Nancy P Barnett. 2014. Sustain talk predicts poorer outcomes among mandated college student drinkers receiving a brief motivational intervention. *Psychology of Addictive Behaviors*, 28(3):631.

Nicklaus Badyal, Derek Jacoby, and Yvonne Coady. 2023. Intentional biases in LLM responses. In *14th IEEE Annual Ubiquitous Computing, Electronics & Mobile Communication Conference, UEMCON 2023, New York, NY, USA, October 12-14, 2023*, pages 502–506. IEEE.

Simone Balloccu, Ehud Reiter, Karen Jia-Hui Li, Rafael Sargsyan, Vivek Kumar, Diego Reforgiato, Daniele Riboni, and Ondrej Dusek. 2024. Ask the experts: sourcing a high-quality nutrition counseling dataset through human-AI collaboration. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 11519–11545, Miami, Florida, USA. Association for Computational Linguistics.

Elizabeth Barnett, Theresa B. Moyers, Steve Sussman, Caitlin Smith, Louise A. Rohrbach, Ping Sun, and Donna Spruijt-Metz. 2014. From counselor skill to decreased marijuana use: Does change talk matter? *Journal of Substance Abuse Treatment*, 46(4):498–505.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 610–623, New York, NY, USA. Association for Computing Machinery.

Laura Ana Maria Bostan, Evgeny Kim, and Roman Klinger. 2020. GoodNewsEveryone: A corpus of news headlines annotated with emotions, semantic roles, and reader perception. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1554–1566, Marseille, France. European Language Resources Association.

Laura-Ana-Maria Bostan and Roman Klinger. 2018. An analysis of annotated corpora for emotion classification in text. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2104–2119, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D 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 Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Sven Buechel and Udo Hahn. 2017. Readers vs. writers vs. texts: Coping with different perspectives of text understanding in emotion annotation. In *Proceedings of the 11th Linguistic Annotation Workshop*, pages 1–12, Valencia, Spain. Association for Computational Linguistics.

Xunxin Cai, Meng Xiao, Zhiyuan Ning, and Yuanchun Zhou. 2023. Resolving the imbalance issue in hierarchical disciplinary topic inference via llm-based data augmentation. In *IEEE International Conference on Data Mining, ICDM 2023, Shanghai, China, December 1-4, 2023*, pages 956–961. IEEE.

Delwyn Catley, Kari Jo Harris, Kathy Goggin, Kimber Richter, Karen Williams, Christi Patten, Ken Resnicow, Edward Ellerbeck, Andrea Bradley-Ewing, Domonique Malomo, et al. 2012. Motivational interviewing for encouraging quit attempts among unmotivated smokers: study protocol of a randomized, controlled, efficacy trial. *BMC public health*, 12:1–8.

Mingda Chen, Jingfei Du, Ramakanth Pasunuru, Todor Mihaylov, Srini Iyer, Veselin Stoyanov, and Zornitsa Kozareva. 2022. Improving in-context few-shot learning via self-supervised training. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3558–3573, Seattle, United States. Association for Computational Linguistics.

Arijit Ghosh Chowdhury and Aman Chadha. 2024. Generative data augmentation using llms improves distributional robustness in question answering. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2024: Student Research Workshop, St. Julian's, Malta, March 21-22, 2024*, pages 258–265. Association for Computational Linguistics.

Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. 2020. GoEmotions: A dataset of fine-grained emotions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4040–4054, Online. Association for Computational Linguistics.

Dorottya Demszky, Diyi Yang, David S Yeager, Christopher J Bryan, Margarett Clapper, Susannah Chandhok, Johannes C Eichstaedt, Cameron Hecht, Jeremy

Jamieson, Meghann Johnson, et al. 2023. Using large language models in psychology. *Nature Reviews Psychology*, 2(11):688–701.

Danilo Dessì, Rim Helaoui, Vivek Kumar, Diego Reforgiato Recupero, and Daniele Riboni. 2020. TF-IDF vs word embeddings for morbidity identification in clinical notes: An initial study. In *Proceedings of the First Workshop on Smart Personal Health Interfaces co-located with 25th International Conference on Intelligent User Interfaces, SmartPhil@IUI 2020, Cagliari, Italy, March 17, 2020*, volume 2596 of *CEUR Workshop Proceedings*, pages 1–12. CEUR-WS.org.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Zhiyong Wu, Baobao Chang, Xu Sun, Jingjing Xu, and Zhifang Sui. 2022. A survey on in-context learning. *arXiv preprint arXiv:2301.00234*.

Haonan Duan, Adam Dziedzic, Nicolas Papernot, and Franziska Boenisch. 2023. Flocks of stochastic parrots: Differentially private prompt learning for large language models. In *Advances in Neural Information Processing Systems*, volume 36, pages 76852–76871. Curran Associates, Inc.

Jacques Gaume, Gerhard Gmel, Mohamed Faouzi, and Jean-Bernard Daeppen. 2009. Counselor skill influences outcomes of brief motivational interventions. *Journal of Substance Abuse Treatment*, 37(2):151–159.

Benedetta Giovanola and Simona Tiribelli. 2023. Beyond bias and discrimination: redefining the AI ethics principle of fairness in healthcare machine-learning algorithms. *AI Soc.*, 38(2):549–563.

Jennifer Hettema, Julie Steele, and William R Miller. 2005. Motivational interviewing. *Annu. Rev. Clin. Psychol.*, 1:91–111.

Van Hoang, Eoin Rogers, and Robert Ross. 2024. How can client motivational language inform psychotherapy agents? In *Proceedings of the 9th Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2024)*, pages 23–40, St. Julians, Malta. Association for Computational Linguistics.

Siddharth Karamcheti, Laurel Orr, Jason Bolton, Tianyi Zhang, Karan Goel, Avanika Narayan, Rishi Bommasani, Deepak Narayanan, Tatsunori Hashimoto, Dan Jurafsky, et al. 2021. Mistral–a journey towards reproducible language model training.

Vivek Kumar., Simone Balloccu., Zixiu Wu., Ehud Reiter., Rim Helaoui., Diego Recupero., and Daniele Riboni. 2023. Data augmentation for reliability and fairness in counselling quality classification. In *Proceedings of the 1st Workshop on Scarce Data in Artificial Intelligence for Healthcare - SDAIH,*, pages 23–28. INSTICC, SciTePress.

Vivek Kumar, Simone Balloccu, Zixiu Wu, Ehud Reiter, Rim Helaoui, Diego Reforgiato Recupero, and Daniele Riboni. 2023a. Data augmentation for reliability and fairness in counselling quality classification.

Vivek Kumar, Giacomo Medda, Diego Reforgiato Recupero, Daniele Riboni, Rim Helaoui, and Gianni Fenu. 2023b. How do you feel? information retrieval in psychotherapy and fair ranking assessment. In *Advances in Bias and Fairness in Information Retrieval*, pages 119–133, Cham. Springer Nature Switzerland.

Vivek Kumar, Diego Reforgiato Recupero, Daniele Riboni, and Rim Helaoui. 2021. Ensembling classical machine learning and deep learning approaches for morbidity identification from clinical notes. *IEEE Access*, 9:7107–7126.

Vivek Kumar, Diego Reforgiato Recupero, Rim Helaoui, and Daniele Riboni. 2022. K-lm: Knowledge augmenting in language models within the scholarly domain. *IEEE Access*, 10:91802–91815.

Vivek Kumar, Prayag Tiwari, and Sushmita Singh. 2023c. VISU at WASSA 2023 shared task: Detecting emotions in reaction to news stories using transformers and stacked embeddings. In *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 581–586, Toronto, Canada. Association for Computational Linguistics.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Junyi Li, Xiaoxue Cheng, Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. 2023a. HaluEval: A large-scale hallucination evaluation benchmark for large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6449–6464, Singapore. Association for Computational Linguistics.

Rumeng Li, Xun Wang, and Hong Yu. 2023b. Two directions for clinical data generation with large language models: Data-to-label and label-to-data. In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, pages 7129–7143. Association for Computational Linguistics.

Zhuoyan Li, Hangxiao Zhu, Zhuoran Lu, and Ming Yin. 2023c. Synthetic data generation with large language models for text classification: Potential and limitations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 10443–10461. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Brad Lundahl, Teena Moleni, Brian L. Burke, Robert Butters, Derrik Tollefson, Christopher Butler, and Stephen Rollnick. 2013. Motivational interviewing in medical care settings: A systematic review and meta-analysis of randomized controlled trials. *Patient Education and Counseling*, 93(2):157–168.

Renata K. Martins and Daniel W. McNeil. 2009. Review of motivational interviewing in promoting health behaviors. *Clinical Psychology Review*, 29(4):283–293.

William R Miller, Theresa B Moyers, Denise Ernst, and Paul Amrhein. Manual for the motivational interviewing skill code (misc).

William R Miller and Stephen Rollnick. 2012. *Motivational interviewing: Helping people change*. Guilford press.

Sergio Morales, Robert Clarisó, and Jordi Cabot. 2023. Automating bias testing of llms. In *38th IEEE/ACM International Conference on Automated Software Engineering, ASE 2023, Luxembourg, September 11-15, 2023*, pages 1705–1707. IEEE.

TB Moyers, JK Manuel, D Ernst, T Moyers, J Manuel, D Ernst, and C Fortini. 2014. Motivational interviewing treatment integrity coding manual 4.1 (miti 4.1). *Unpublished manual*.

Theresa B Moyers, Tim Martin, Jon M Houck, Paulette J Christopher, and J Scott Tonigan. 2009. From in-session behaviors to drinking outcomes: a causal chain for motivational interviewing. *Journal of consulting and clinical psychology*, 77(6):1113.

Theresa B Moyers and Stephen Rollnick. 2002. A motivational interviewing perspective on resistance in psychotherapy. *Journal of clinical psychology*, 58(2):185–193.

Roberto Navigli, Simone Conia, and Björn Ross. 2023. Biases in large language models: Origins, inventory, and discussion. *J. Data and Information Quality*, 15(2).

World Health Organization. 2022. World mental health report: Transforming mental health for all.

Verónica Pérez-Rosas, Rada Mihalcea, Kenneth Resnicow, Satinder Singh, and Lawrence An. 2016. Building a motivational interviewing dataset. In *Proceedings of the Third Workshop on Computational Linguistics and Clinical Psychology*, pages 42–51, San Diego, CA, USA. Association for Computational Linguistics.

Madeline Reinert, Danielle Fritze, and Theresa Nguyen. 2021. The state of mental health in america 2022.

Laria Reynolds and Kyle McDonell. 2021. Prompt programming for large language models: Beyond the few-shot paradigm. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI EA '21, New York, NY, USA. Association for Computing Machinery.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Advait Sarkar. 2023. Exploring perspectives on the impact of artificial intelligence on the creativity of knowledge work: Beyond mechanised plagiarism and stochastic parrots. In *Proceedings of the 2nd Annual Meeting of the Symposium on Human-Computer Interaction for Work*, CHIWORK '23, New York, NY, USA. Association for Computing Machinery.

Richard Shiffrin and Melanie Mitchell. 2023. Probing the psychology of ai models. *Proceedings of the National Academy of Sciences*, 120(10):e2300963120.

Benjamin Smith, Anahita Khojandi, and Rama K. Vasudevan. 2024. Bias in reinforcement learning: A review in healthcare applications. *ACM Comput. Surv.*, 56(2):52:1–52:17.

Yu Song, Santiago Miret, Huan Zhang, and Bang Liu. 2023. HoneyBee: Progressive instruction finetuning of large language models for materials science. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5724–5739, Singapore. Association for Computational Linguistics.

Tiberiu Sosea and Cornelia Caragea. 2020. Cancergemo: A dataset for fine-grained emotion detection. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8892–8904.

Massimo Stella, Thomas T. Hills, and Yoed N. Kenett. 2023. Using cognitive psychology to understand gpt-like models needs to extend beyond human biases. *Proceedings of the National Academy of Sciences*, 120(43):e2312911120.

Hongjin Su, Jungo Kasai, Chen Henry Wu, Weijia Shi, Tianlu Wang, Jiayi Xin, Rui Zhang, Mari Ostendorf, Luke Zettlemoyer, Noah A. Smith, and Tao Yu. 2023. Selective annotation makes language models better few-shot learners. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Nicholas L Syn, Andrea Li-Ann Wong, Soo-Chin Lee, Hock-Luen Teoh, James Wei Luen Yip, Raymond Cs Seet, Wee Tiong Yeo, William Kristanto, Ping-Chong Bee, L M Poon, Patrick Marban, Tuck Seng Wu, Michael D Winther, Liam R Brunham, Richie Soong, Bee-Choo Tai, and Boon-Cher Goh. 2018. Genotype-guided versus traditional clinical dosing of warfarin in patients of asian ancestry: a randomized controlled trial. *BMC Med.*, 16(1):104.

Ruixiang Tang, Xiaotian Han, Xiaoqian Jiang, and Xia Hu. 2023. Does synthetic data generation of llms help clinical text mining? *CoRR*, abs/2303.04360.

Leili Tavabi, Trang Tran, Kalin Stefanov, Brian Borsari, Joshua Woolley, Stefan Scherer, and Mohammad Soleymani. 2021. Analysis of behavior classification in motivational interviewing. In *Proceedings of the Seventh Workshop on Computational Linguistics and Clinical Psychology: Improving Access*, pages 110–115, Online. Association for Computational Linguistics.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Satvik Tripathi, Rithvik Sukumaran, and Tessa S Cook. 2024. Efficient healthcare with large language models: optimizing clinical workflow and enhancing patient care. *Journal of the American Medical Informatics Association*, page ocad258.

Alper Kursat Uysal and Serkan Gunal. 2014. The impact of preprocessing on text classification. *Information Processing & Management*, 50(1):104–112.

Yixin Wan, George Pu, Jiao Sun, Aparna Garimella, Kai-Wei Chang, and Nanyun Peng. 2023. "kelly is a warm person, joseph is a role model": Gender biases in llm-generated reference letters. In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, pages 3730–3748. Association for Computational Linguistics.

Yuqing Wang, Yun Zhao, and Linda Petzold. 2023. Are large language models ready for healthcare? a comparative study on clinical language understanding. In *Proceedings of the 8th Machine Learning for Healthcare Conference*, volume 219 of *Proceedings of Machine Learning Research*, pages 804–823. PMLR.

Anuradha Welivita and Pearl Pu. 2022. Curating a large-scale motivational interviewing dataset using peer support forums. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3315–3330, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Stanislaw Wozniak and Jan Kocon. 2023. From big to small without losing it all: Text augmentation with chatgpt for efficient sentiment analysis. In *IEEE International Conference on Data Mining, ICDM 2023 - Workshops, Shanghai, China, December 4, 2023*, pages 799–808. IEEE.

Zixiu Wu, Simone Balloccu, Vivek Kumar, Rim Helaoui, Diego Reforgiato Recupero, and Daniele Riboni. 2023. Creation, analysis and evaluation of annomi, a dataset of expert-annotated counselling dialogues. *Future Internet*, 15(3).

Zixiu Wu, Simone Balloccu, Vivek Kumar, Rim Helaoui, Ehud Reiter, Diego Reforgiato Recupero, and Daniele Riboni. 2022. Anno-mi: A dataset of expert-annotated counselling dialogues. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6177–6181.

Zixiu Wu, Rim Helaoui, Vivek Kumar, Diego Reforgiato Recupero, and Daniele Riboni. 2021. Towards detecting need for empathetic response in motivational interviewing. In *Companion Publication of the 2020 International Conference on Multimodal Interaction*, ICMI '20 Companion, page 497–502, New York, NY, USA. Association for Computing Machinery.

Guojun Yan, Jiahuan Pei, Pengjie Ren, Zhaochun Ren, Xin Xin, Huasheng Liang, Maarten de Rijke, and Zhumin Chen. 2022. Remedi: Resources for multi-domain, multi-service, medical dialogues. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 3013–3024.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Kai-Ching Yeh, Jou-An Chi, Da-Chen Lian, and Shu-Kai Hsieh. 2023. Evaluating interfaced llm bias. In *Proceedings of the 35th Conference on Computational Linguistics and Speech Processing (ROCLING 2023)*, pages 292–299.

Ping Yu, Hua Xu, Xia Hu, and Chao Deng. 2023. Leveraging generative AI and large language models: A comprehensive roadmap for healthcare integration. *Healthcare (Basel)*, 11(20):2776.

Jiayi Yuan, Ruixiang Tang, Xiaoqian Jiang, and Xia Hu. 2023. Large language models for healthcare data augmentation: An example on patient-trial matching. *AMIA Annu. Symp. Proc.*, 2023:1324–1333.

A Appendix

An excerpt from high-quality MI Counselling Session

Therapist:	Thanks for filling it out. We give this form to everyone once a year regardless of why they come in. It helps us provide better care. Is it okay if I take a look at what you put down?
Client:	Sure.
Therapist:	So, let's see. It looks that you put-- You drink alcohol at least four times a week on average?
Client:	Mm-hmm.:
Therapist:	Okay
Client:	Usually three drinks and glasses of wine.
Therapist:	Okay. That's at least 12 drinks a week.
Client:	Something like that.
Therapist:	Okay. Just so you know, my role, um, when we talk about alcohol use, is just to share information about risk and to help patients who want help. This is different than telling them what I think they should do. I don't do that.
Client:	Okay.
Therapist:	Uh, what else can you tell me about your drinking.
Client:	Well, I usually drink when I'm at home trying to unwind and I drink while I'm watching a movie. And sometimes, um, I take a bath but I also drink when I take a bath sometimes.
Therapist:	Okay. So, can I share with you some information on alcohol use?
Client:	Okay.
Therapist:	Okay. So, there has been a lot of research on alcohol use and the guidelines we use in this country says that having seven or more drinks per week can raise the risk of health problems for women.
Client:	Hmm. Seven?
Therapist:	Seven.
Client:	Wow. I knew my doctor didn't like me drinking the amount that I did but I didn't know that seven was the limit.
Therapist:	Yeah, you're surprised to hear that?
Client:	Yes. What-what kind of health problems?
Therapist:	Well things like heart disease, cancer, liver problems, uh, stomach pains, insomnia. Unfortunately, uh, people who drink at a risky level are more likely to be diagnosed with depression and alcohol can make depression worse or harder to treat.

Figure 7: The distribution of client utterances in training and test set of IC-AnnoMI dataset.

B Appendix

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

$$Bal.Acc. = \frac{TP(TN + FP) + TN(TP + FN)}{2 \times (TP + TN + FP + FN)} \quad (2)$$

$$Precision = \frac{TP}{TP + FP} \quad (3)$$

$$Recall = \frac{TP}{TP + FN} \quad (4)$$

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (5)$$

where TP stands for true positive, TN for true negative, FP for false positive, and FN for false negative.

An excerpt from LLMs generated MI Counselling Session

Therapist:	Thank you for taking the time to fill out the form. We provide this questionnaire annually to all our clients to ensure we're offering the best possible care. May I review what you've shared?
Client:	Of course.
Therapist:	Alright, let's have a look. It seems you indicated that you consume alcohol at least four times a week, on average.
Client:	Yes, usually about three drinks or glasses of wine.
Therapist:	I see. So that's roughly 12 drinks per week, correct?
Client:	Around that, yes.
Therapist:	Just to clarify, my role here is to provide information about the risks associated with alcohol use and to support those who seek help. I don't dictate what you should or shouldn't do.
Client:	Understood.
Therapist:	Is there anything else you'd like to share about your drinking habits?
Client:	Well, I tend to drink when I'm at home unwinding, watching movies, or taking a bath. It's become a bit of a routine.
Therapist:	I see. Would you be open to hearing some information about alcohol use?
Client:	Sure, go ahead.
Therapist:	According to the guidelines, consuming seven or more drinks per week can increase the risk of health issues for women.
Client:	Seven drinks? That's surprising.
Therapist:	Yes, it can lead to various health problems such as heart disease, cancer, liver issues, and even worsen conditions like depression.
Client:	I didn't realize it could impact my health that much.
Therapist:	How do you think your drinking habits might be connected to your experiences with depression?
Client:	Honestly, I sometimes turn to alcohol when I'm feeling low because it distracts me, but then I feel worse afterward.
Therapist:	I see. The form also mentions instances where you drink more than intended and feel guilty afterward
Client:	Yes, sometimes I realize I've had too much, especially when I finish a whole bottle.
Therapist:	It seems like you're not entirely comfortable with your drinking habits.

Figure 8: The distribution of therapist utterances in training and test set of IC-AnnoMI dataset.

C Appendix

Item	**Specification**
CPU	Intel Core i3-7100 (-HT-MCP-) CPU @ 3.90 GHz
GPU	NVIDIA GP102 [TITAN X], 12 GB memory
Graphic Driver	NVIDIA graphic driver version 440.33.01
CUDA	Version 10.2
OS	Ubuntu (17.10)
Python	Version 3.6.6

Table 3: Server specifications.