Leveraging Large Language Models for In-context Data Generation to Address Bias and Scarce Data Challenges in Mental Healthcare

Anonymous ACL submission

Abstract

Motivational Interviewing (MI) is a widely used, evidence-based counseling approach, yet the development of robust NLP systems for MI remains constrained by severe data scarcity, high annotation costs, and the need for domain expertise. While recent work has explored large language models (LLMs) for clinical text generation and augmentation, existing studies largely focus on utterance-level transformations or evaluate a limited set of models, leaving the role of long-range conversational context underexplored. In this work, we present the first systematic, session-level benchmark of LLM-based data augmentation for MI dialogues. We compare prompt-based generative augmentation at the full-session level with utterance-level task-sensitive transformation methods across 13 state-of-the-art (SOTA) LLMs and three long-context classification models designed for extended clinical conversations. As a key outcome, we also present **ICAUGAnnoMI**, a novel dataset of 1,764 low- and high-quality MI dialogues, spanning nearly 81k talk turns between therapist and client alongside a fidelity-aware evaluation framework that assesses semantic drift, hallucination, and adherence to core MI principles. Our empirical results demonstrate that session-level augmentation consistently outperforms utterance-level approaches in improving MI session classification, particularly for minority and low-quality cases, highlighting the importance of preserving long-range conversational structure. Beyond performance gains, our analysis provides practical insights into the strengths and limitations of contemporary LLMs in sensitive mental health settings. We release our code and commit to publicly sharing the generated data to facilitate reproducible research and future benchmarking in MI-focused NLP. ICAUGAnnoMI data and source code is available at **repository**[1].

[1]https://anonymous.4open.science/r/ARR_Submission_Cycle_2026/README.md

1 Introduction

Anxiety and depression are among the most common psychological disorders, affecting social and economic functioning. According to the latest WHO report[2], mental disorders account for 1 in 6 years lived with disability, and individuals with severe conditions die 10–20 years earlier than the general population (Organization, 2022). AI-driven solutions have recently gained attention for alleviating this burden by supporting early screening, psycho-education, and self-help strategies (Shiffrin and Mitchell, 2023; Stella et al., 2023; Demszky et al., 2023; Sezgin, 2023; Balloccu et al., 2024a). Large Language Models (LLMs) are particularly promising due to their human-like dialogue capabilities (Han et al., 2024; Kumar et al., 2024, 2025) and potential to improve access in low-resource settings (Xu et al., 2024). However, integrating LLMs into mental healthcare faces challenges (Wang et al., 2023): i) *hallucination*, generating plausible but incorrect text (Li et al., 2023a); ii) *stochastic parroting*, producing fluent but shallow responses (Bender et al., 2021); iii) *bias*, reducing reliability for underrepresented groups due to representation and sampling issues (Balloccu et al., 2024b; Badyal et al., 2023; Morales et al., 2023; Harrer, 2023). These challenges converge on a single issue: the *lack of publicly available high-quality data for effective training*. To address this, we employ state-of-the-art (SOTA) LLMs for in-context human-like data generation at the MI session level (Brown et al., 2020; Chen et al., 2022; Dong et al., 2022) and task-sensitive NLP augmenters at the utterance level (Dhole et al., 2023).

Our experimental study is based on AnnoMI (Wu et al., 2022, 2023; Kumar et al., 2023), a publicly available expert-annotated dataset com-

[2]https://www.who.int/teams/mental-health-and-substance-use/world-mental-health-report

pliant with the General Data Protection Regulation (GDPR). Unlike prior work that evaluates a single model or augmentation strategy, our study serves as a controlled benchmark of session-level versus utterance-level data augmentation for MI, across diverse LLM families and long-context classifiers. To the best of our knowledge, ours is the first work that uses a diverse range of LLMs, including the latest **Deepseek** model, for in-context conversation generation at the level of full therapist-client sessions. Our research aims to answer the following research questions (RQs):

- **RQ1**: Are existing LLMs envisioned for mental health research sufficiently powerful to automatically identify the quality of MI sessions?
- **RQ2**: Are SOTA generative models applied at the session level the key solution for the generation of high-quality and contextually precise MI conversations?
- **RQ3**: Can task-sensitive text transformations surpass the quality of LLM-generated MI sessions by processing individual utterances at a time?

The key contributions of our study can be summed up as follows:

- **Session-level MI Augmentation**: We present the first systematic benchmark of session-level data augmentation for Motivational Interviewing, directly comparing prompt-based generative augmentation with utterance-level task-sensitive transformations while maintaining therapeutic fidelity..
- **Utterance-level MI Augmentation**: We also employ five task-sensitive augmentation strategies to assess whether transforming individual utterances could actually surpass prompt-based generative methods, despite lacking the ability to perform text manipulations on the basis of long and contextually rich session knowledge.
- **Model Evaluation in Long-range Text Settings**: We further evaluate long-context models to assess how augmented data affects training, particularly differences between low- and high-quality sessions and the impact of noisy content.
- **Curated MI Dataset**: We curate and release a high-quality dataset of 1,764 MI sessions with expert-informed quality labels, enabling long-context modeling and reproducible evaluation.
- **Augmentation Evaluation Scheme**: We develop a qualitative evaluation scheme comprised of six parametershighlighting the distinct features of low- and high-quality sessions and the potential for noisy content to adversely affect performance
- **Reproducibility**: We commit to provide public access to the source code publicly sharing generated data after acceptance.

2 Method

In this section, we outline the methods employed to develop and evaluate our approach to data augmentation in MI sessions. First, we define the task and challenges of generating high-quality MI dialogues. Second, we describe the dataset and the data preparation procedures. Finally, we delve into two data augmentation strategies for generating comprehensive datasets.

2.1 Problem definition

As previously emphasized, given the complex nature of psychotherapy, MI sessions are often difficult and expensive to capture at scale, and existing scarce data challenges in this domain make it increasingly difficult to develop efficient and robust systems. The main challenge arises from the specialized expertise required to assess the quality of MI sessions, which is crucial for building reliable datasets. Our study addresses this challenge and aims to advance MI research by creating enhanced classifiers that can automatically identify the key components of MI and, thus, the quality and context of therapy sessions, facilitating the development of a dataset and reliable systems. To this end, we explore various utterance and session-based data augmentation techniques to increase the variety and quality of MI dialogue examples derived from a few existing expert-annotated data.

This process starts from a dataset of L labeled MI sessions $\mathcal{D} = \{(S_i, q_i)\}_{i=1}^{L}$. A therapy session $S_i = \left(t_{i1}, t_{i2}, \ldots, t_{iM_i}\right) \in \mathcal{S}$ is defined as a sequence of M_i talk turns between a client and a therapist, while $q_i \in \{0, 1\}$ is a binary label representing the quality of the session S_i, where 0 is assigned to *low*-quality sessions, 1 to *high*-quality ones. We define a feature mapping $\Gamma : \mathcal{S} \rightarrow \mathbb{R}^d$, a classifier $f_\theta : \mathbb{R}^d \rightarrow [0, 1]$, and the following partition $\mathcal{D} = \mathcal{D}_{train} \cup \mathcal{D}_{test}$. Let ϕ be an augmentation function, drawn from some family of augmentation functions Φ. Formally:

$$\phi : \mathcal{D}_{train} \mapsto \mathcal{D}_{aug} \quad (1)$$

that is ϕ generates an augmented variation $\mathcal{D}_{aug}$ of the original training set. We seek to find the augmenter ϕ^* that solves the empirical risk minimization problem as shown in Appendix C.

2.2 Dataset

For this study, we used the GDPR-compliant dataset **AnnoMI**, which consists of 131 low- and high-quality MI dialogues, each containing talk turns between the client and the therapist (Wu et al., 2022, 2023; Kumar et al., 2023), for a total of 9,649 utterances. The dataset contains 23 low-quality and 108 high-quality MI dialogues. AnnoMI is among the few publicly available datasets that contain labeled MI sessions at both the talk-turn and entire conversational levels. This dataset orientation complements our experimental setup of synthetic MI generation and, thus, is the obvious choice for this study. For ease of understanding, Figure 1 in Appendix B depicts a sample of low- and high-quality MI sessions.

2.3 Data augmentation

We specifically focus on generating human-like and contextually sound MI sessions that maintain therapeutic fidelity and reflect the diverse range of patient responses and therapist strategies. Given the high cost and complexity of acquiring large-scale MI data, we explore two complementary strategies for data augmentation. First, we employ task-sensitive transformations, which alter transcripts through techniques specialized in text manipulation, e.g., paraphrasing. Second, we utilize prompt-based generation, leveraging LLMs to create new, contextually accurate MI dialogues.

2.3.1 Task-sensitive transformation

Task-sensitive transformations (TTs) refer to augmentation methods that adopt text-driven modifications on textual data with a specific expected output in mind, such as enhancing linguistic diversity, shifting stylistic properties, or rephrasing. These strategies are typically adopted for analyzing the impact of text manipulations on trained models, both from a negative (concerning text perturbation) and a positive (concerning text augmentation) viewpoint. Considering that their diffusion precedes the advent of LLMs, they are not typically devised for transforming long input sequences. In our scenario involving long MI sessions, it is counterintuitive to adopt TTs on several client-therapist talk-turns, which would cause the underlying language models (LMs) to drop excessive text and provide trimmed outputs. It follows that TTs should only be adopted at the utterance level to generate augmented data that preserves the original topic the most. Despite the downside of overlooking the contextual information of entire sessions, TTs represent solid augmentation baselines, ranging from rich paraphrasing variety to subtle word edits. Let $\Lambda : \mathcal{T} \rightarrow \mathbb{R}^e$ be a feature mapping for individual talks $t \in \mathcal{T}$, we can formally define an augmenter for TTs as mentioned in Appendix D.

This design choice was made to ensure a fair comparison across TTs, as some are not capable of generating diverse transformations from the same text input.

2.3.2 One-shot and few-shot based generation

Augmentation resulting from prompt-based generation (PG) refers to the adoption of generative models, specifically LLMs, to generate variations of original MI sessions while preserving contextually relevant information and keeping dialogues coherent with the overall session topic. LLMs represent the prevailing solution for most of language processing tasks, thanks to the versatile management of output diversity through specific parameters, e.g., temperature, top-p sampling, or prompt refinement. Despite these benefits, they can be inconsistent and difficult to steer their generation process toward the required results. Main weaknesses, such as hallucination and poor instruction following, mine the reliability of their replies and might require additional processing layers to filter out irrelevant information.

Unlike TTs, SOTA LLMs support textual inputs with large context windows, enabling our targeted task at the session level by interleaving individual turn talks and interlocutor delimiters. To guarantee high output fidelity and strong instruction following, we adopted several prompting techniques and iteratively refined the initial prompt until satisfied with PG augmentations. Figure 2 in appendix E depicts the final prompting setup, consisting of a system and a user prompt for accurate role assignment and a PG output example with augmented turn talks. The following formalizes the PG process as TT augmentation one:

$$\phi_{PG}(\mathcal{D}_{train}) = \left\{ \left(\phi^S_{PG}(S_i),\ q_i \right) \right\}_{i=1}^{L} \quad (2)$$

where $\phi^S_{PG} : \mathbb{R}^d \rightarrow \mathbb{R}^d$ denotes a PG sub-function that operates at the session level.

While TTs can guarantee that the resulting output includes an equal number of utterances as the original training set, PG reports higher variance in outcomes. Indeed, due to diverse model parameter sizes, configurations, and underlying functioning, PG may result in shorter or longer sessions compared with the original data.

3 Experimental Design

This section outlines the steps for preparing the experimental data and presents the classification architectures, refined promotion strategies, and the d experimental pipeline. The targeted objective is addressed by assessing the efficacy of TT and PG augmented data in improving classification performance under the experimental setup below:

- **Data Preparation:** Out of 131 sessions, we selected 97 sessions for the training set, of which 15 are low- and 82 are high-quality MI dialogues. The remaining 34 MI sessions constitute the test set, of which 8 are low- and 26 are high-quality MI sessions. 10 MI sessions (3 low- and 7 high-quality) are removed from the training set to create a validation set further. Such a splitting strategy guarantees that MI quality samples in each set remain representative of the original session distribution.

- **Classification Models:** We employed three classifiers, namely BigBird[3] (Zaheer et al., 2020), Longformer[4] (Beltagy et al., 2020) and ModernBERT[5] (Warner et al., 2024). The rationale behind using BigBird, Longformer, and ModernBERT over usual transformers is their capability to handle long text sequences. Preserving the context of the MI session is crucial to our experimental design, and given that MI sessions are as long as 8000 tokens, these models completely match the experiment's requirement. The inclusion of ModernBERT as the latest modern release of BERT models emphasizes our adoption of SOTA techniques for high-quality outputs. All the experiments adopt accuracy (**Acc**), balanced accuracy (**Bal. Acc.**), F1-score (**F1**), precision (**Prec**), and recall (**Rec**) as evaluation metrics for classification performance of the employed models. The experiment using only the original data $\mathcal{D}_{train}$, labeled as `AnnMI (Base Data)`, serves as the baseline to estimate the impact of augmented data in influencing classification performance.

- **Experiment with Domain-trained Model:** Since our experiments are session and not utterance-based, the classification task becomes complex and challenging, given the complex domain of psychology and a highly skewed and unbalanced distribution of the experimental dataset. Therefore, as a sanity check, we begin by evaluating the performance of the SOTA domain-trained LLM-MentaLLaMA[6] (Yang et al., 2024). This experiment is set to evaluate whether domain-trained models can outperform contemporary classifiers fine-tuned on experimental data, highlighting in favor or against the need for data augmentation.

- **Experiments with LMs:** To evaluate the performance of classification models, TT and PG augmenters, we rigorously conduct three sets of experiments:
i) First, we perform the experiments with the $\mathcal{D}_{train}$ with the three classifiers.
ii) Then $\mathcal{D}_{train}$ + $\mathcal{D}_{aug,PG}$ and $\mathcal{D}_{train}$ + $\mathcal{D}_{aug,TT}$ are used as training data with the three classifiers iteratively, where $\mathcal{D}_{aug,PG}$ and $\mathcal{D}_{aug,TT}$ are the augmented dataset generated by PG (i.e., LLMs) and TT augmenters respectively. Here $\mathcal{D}_{aug}$ can be only one augmented dataset for one experiment.
iii) Now, based on the performance of classification models, the best-performing $\mathcal{D}_{aug}$ datasets from each $\mathcal{D}_{aug,PG}$ and $\mathcal{D}_{aug,TT}$ are selected to make two separate training datasets namely $\mathcal{D}_{train+aug,PG}$ and $\mathcal{D}_{train+aug,TT}$ representing the combination of original data with data augmented by each method family.
iv) Finally, $\mathcal{D}_{train+aug,PG}$ and $\mathcal{D}_{train+aug,TT}$

[3]https://huggingface.co/docs/transformers/en/model_doc/big_bird
[4]https://huggingface.co/docs/transformers/en/model_doc/longformer
[5]https://huggingface.co/docs/transformers/en/model_doc/ModernBERT
[6]https://huggingface.co/klyang/MentaLLaMA-chat-7B

are used as training datasets for each of the three classifiers iteratively. This experiment aims to observe the generalization capability of classifiers over diverse datasets and explore the combined contributions of effective augmented data.

4 Results and Discussion

This section presents the experimental results of classification models and empirical analyses for each data augmentation setup, followed by an evaluation of LLMs using a novel tailored evaluation scheme.

4.1 Performance of domain-trained LLM MentalLLaMA

As mentioned in section 3, we begin our experiment with the SOTA domain-trained LLM MentalLLaMA (Yang et al., 2024) to observe if domain-specific models can outperform domain-agnostic models on specific task and dataset. Interestingly, MentalLLaMA performed poorly on the task, as observed over three iterations. The output explanations produced by MentalLLaMA are unrecognizable, and it mostly predicts all the MI sessions as "LOW," which is the minority class. This lapse in the performance of the MentalLLaMA being SOTA model highlights a clear research gap in low-resource domains such as mental health, and it raises concerns about the model's reliability and readiness for domain-specific NLP downstream tasks. It also indicates the need to enrich the low-resource domain to investigate optimal practices for generating high-quality data, enabling LMs and LLMs to better interpret the nuances and intricacies of the complex domain, leading to reliable performance.

4.2 Classifiers performance with PG-based augmented datasets

The performance of the three classifiers with each of the 13 LLM-augmented datasets is summarized in Table 1. For ease of interpretation, the best-performing augmented dataset and models are highlighted in bold. Table 1 shows that most of the LLM-augmented datasets have improved the employed LMs' accuracy compared to the baseline. For instance, BigBird has improved by **5.88%** with Llama 3.2, Longformer by **17.65%** with Smollm2, and ModernBERT by **2.94%** with Nemotron-Mini. It is also noteworthy that the baseline performance of ModernBERT is significantly higher and the highest among the three LMs, which emphasizes the efficient architecture of ModernBERT, designed to reduce bias and pay particular attention to efficiency. Also, peculiarly, Longformer has the lowest baseline performance, but augmentation shines and improves its performance by a noteworthy **17.65%**.

4.3 Classifiers Performance with TT-based augmented datasets

The performance of each of the three classifiers with each of the 5 NLP-augmented datasets is summed up in Table 3 in Appendix G. The best-performing data and models are highlighted in bold. As evident from Table 3, most of the NLP-augmented datasets have also improved the LMs' accuracy compared to the baseline. For instance, BigBird has improved by **5.88%** with Backtranslation, Longformer by **17.65%** with ProtAugment and Tweet style, and ModernBERT by **2.24%** with Backtranslation. Besides the best performance, the "Basic style" is consistently the lowest performer. Additionally, in retrospect, Longformer has the lowest baseline performance, and augmentation has been very useful, yielding an improvement of **17.65%**.

4.4 Performance of LMs with mixed augmented dataset

To evaluate the generalization capability of LMs over diversified data, i.e., the training dataset containing mixed-style synthetic data, we used a combination for LLM- and NLP-augmented datasets. From LLM, we have Nemotron-Mini(4B) and Phi4(14B), the best performers averaged over all experiments. Similarly, from NLP-augmented datasets, we select Backtranslation, ProtAugment and Tweet style for the best performance (averaged over all experiments). The performance of each of the three classifiers individually with the mixed LLM- and NLP-augmented dataset is presented in Table 2 and the best-performing data and models are highlighted in bold. The LMs BigBird and Longformer have improved their performance with the mixed, augmented dataset. To be specific, BigBird has observed an increase of **3**.64% in accuracy with combined LLM-augmented data and Longformer has scored a **17.65%** increase in accuracy with LLM-augmented data. As mentioned before, ModernBERT, being SOTA and a robust model, has set the highest baseline score among the LMs. Therefore, it is imperative to fine-tune ModernBERT on high-quality data to

Model	Augmented Data	Acc	Bal. Acc.	F1	Prec	Rec
BigBird	AnnoMI (Base Data)	0.8824	**0.9231**	0.8892	0.9216	0.8824
	ChatGPT (4.0)	0.7647	0.5000	0.6627	0.5848	0.7647
	DeepSeek (v3)	0.7059	0.5048	0.6674	0.6451	0.7059
	Gemini-Flash	0.8824	0.7500	0.8669	0.8980	0.8824
	Llama3.1 (70B)	0.7941	0.6923	0.7892	0.7859	0.7941
	Llama3.2 (3B)	**0.9412**	0.9183	**0.9412**	**0.9412**	**0.9412**
	Llama3.3 (70B)	0.8529	0.6875	0.8260	0.8767	0.8529
	Mistral-Nemo (12B)	0.8235	0.6250	0.7797	0.8566	0.8235
	Nemotron-Mini (4B)	0.8824	0.7933	0.8761	0.8789	0.8824
	Phi4 (14B)	0.9118	0.8125	0.9040	0.9209	0.9118
	Qwen2.5 (32B)	0.8235	0.6683	0.8004	0.8137	0.8235
	Qwen2.5 (72B)	0.7647	0.5000	0.6627	0.5848	0.7647
	Smollm2 (1.7B)	0.7941	0.5625	0.7263	0.8378	0.7941
Longformer	AnnoMI (Base Data)	0.7647	0.5000	0.6627	0.5848	0.7647
	ChatGPT (4.0)	0.7647	0.5000	0.6627	0.5848	0.7647
	DeepSeek (v3)	0.8529	0.7308	0.8400	0.8475	0.8529
	Gemini-Flash	0.7647	0.5000	0.6627	0.5848	0.7647
	Llama3.1 (70B)	0.7647	0.5000	0.6627	0.5848	0.7647
	Llama3.2 (3B)	0.7647	0.5000	0.6627	0.5848	0.7647
	Llama3.3 (70B)	0.7647	0.5000	0.6627	0.5848	0.7647
	Mistral-Nemo (12B)	0.8824	0.7500	0.8669	0.8980	0.8824
	Nemotron-Mini (4B)	0.7941	0.8221	0.8077	0.8550	0.7941
	Phi4 (14B)	0.8529	0.7308	0.8400	0.8475	0.8529
	Qwen2.5 (32B)	0.8235	0.6683	0.8004	0.8137	0.8235
	Qwen2.5 (72B)	0.8235	0.7548	0.8235	0.8235	0.8235
	Smollm2 (1.7B)	**0.9412**	**0.8750**	**0.9381**	**0.9454**	**0.9412**
ModernBERT	AnnoMI (Base Data)	0.9118	0.8125	0.9040	0.9209	0.9118
	ChatGPT (4.0)	0.8529	0.6875	0.8260	0.8767	0.8529
	DeepSeek (v3)	0.8824	0.8365	0.8824	0.8824	0.8824
	Gemini-Flash	0.7941	0.6490	0.7760	0.7740	0.7941
	Llama3.1 (70B)	0.8529	0.7740	0.8494	0.8478	0.8529
	Llama3.2 (3B)	0.7647	0.5865	0.7339	0.7294	0.7647
	Llama3.3 (70B)	0.7941	0.6058	0.7564	0.7736	0.7941
	Mistral-Nemo (12B)	0.8529	0.8173	0.8558	0.8604	0.8529
	Nemotron-Mini (4B)	**0.9412**	**0.8750**	**0.9381**	**0.9454**	**0.9412**
	Phi4 (14B)	0.9118	0.8125	0.9040	0.9209	0.9118
	Qwen2.5 (32B)	0.8235	0.6683	0.8004	0.8137	0.8235
	Qwen2.5 (72B)	0.8824	0.7500	0.8669	0.8980	0.8824
	Smollm2 (1.7B)	0.8529	0.7308	0.8400	0.8475	0.8529

Table 1: Classification Metrics for different models and LLM-generated augmented data.

further improve performance on mixed training data.

Across experiments, session-level PG consistently improves balanced accuracy over utterance-level TT, particularly for minority-class MI sessions, indicating that long-range discourse is critical for MI quality modeling.

5 Evaluation Scheme and LLM-generated Data Analysis

This section presents the evaluation scheme we developed, centered on the experimental setup and dataset, and how we implemented it, followed by a detailed explanation of each LLM's performance.

5.1 Evaluation scheme development

For the quantitative and qualitative evaluation of LLMs, we developed a novel evaluation scheme with the help of a domain expert. In the evaluation scheme, we parameterize the attributes of MI sessions to capture the full spectrum of their quality and content. The evaluation scheme is applied to the representative sample taken from $\mathcal{D}_{train}$ and each of the $\mathcal{D}_{aug_llm}$ containing original $MI_{orig.}$ and synthetic MI sessions $MI_{syn.}$. The outcome is the average score over the representative sample.

- **Semantic Drift:** It refers to the change in the meaning of words or phrases over time or across contexts. In our case, we use it to indicate how much semantic deviation happened in $MI_{syn.}$ with respect to (w.r.t) to $MI_{orig.}$. It is measured on a scale of High, Moderate, and Low.

- **Hallucination:** It refers to the false or unsupported information generation w.r.t to the input data. We use it to observe how often $MI_{syn.}$ has made the wrong interpretation of the therapeutic notion w.r.t $MI_{orig.}$. It is measured on a scale of High, Moderate, and Low.

- **MI Session Length:** We use this parameter to observe if extreme changes in the length of $MI_{syn.}$ w.r.t $MI_{orig.}$ are significantly contributing to the quality and false semantic deviation in $MI_{syn.}$. It is measured in Longer, Equal to Shorter w.r.t to $MI_{orig.}$.

- **Prompt Compliance:** We use this parameter to observe how strictly an LLM adheres to the instructions in the prompts. It signifies user-friendliness, input understanding, and desirable output format on the LLM level. It is measured on a scale of High, Moderate, and Low.

- **Fluency and Emotional Understanding:** We use it to refer to the naturalness, coherency, emotional stance, and smoothness of language in $MI_{syn.}$ w.r.t $MI_{orig.}$. The emotional stance here indicates if the emotion the client or therapist displays in $MI_{orig.}$ is presented more strongly or mildly in $MI_{syn.}$. For instance, if the client expresses mild tiredness in $MI_{orig.}$ while in $MI_{syn.}$ if perceived as extremely tired and sick, indicates **Low** fluency and emotional understanding.

5.2 LLM-generated data analysis

The performance of each LLM and the LLM-augmented data over the five parameters is presented in Table 4 placed in Appendix H, and we have consolidated the point-to-point discussion as follows.

- **Semantic Drift:** Interestingly, the LLMs have skipped extreme semantic drift (except Smollm2) and have only shown Moderate to Low semantic deviation.

- **Hallucination:** None of the LLMs have shown extreme hallucination. However, we noticed different hallucination variations in the augmented MI sessions. For instance, Gemini introduced a pseudo-name for the client to make the conversation flow more naturally. However, it does not change the context; what it does change is the client's gender, which might conflict with specific therapy courses.

- **MI Session Length:** In this case, LLMs have shown all variations on the scale, i.e., while Nemotron and Smollm2 have produced mostly extra-long MI sessions, the remaining LLMs have generated shorter sessions. Further, to give a fine-grain quantitative overview, we have presented the distribution of utterances per $MI_{orig.}$ vs. $MI_{syn.}$ for each LLM separately, as mentioned in Figure 3, in Appendix F.

- **Prompt Compliance:** None of the LLMs showed complete ignorance or inability to follow the prompt instruction. While some

Model	Augmented Data	Acc	Bal. Acc.	F1	Prec	Rec
BigBird	AnnoMI (Base Data)	0.8824	**0.9231**	0.8892	**0.9216**	0.8824
	Backtranslation + ProtAugment + Tweet Style	0.8824	0.8798	0.8865	0.8975	0.8824
	Nemotron-Mini (4B) + Phi-4 (14B)	**0.9118**	0.8558	**0.9097**	0.9097	**0.9118**
Longformer	AnnoMI (Base Data)	0.7647	0.5000	0.6627	0.5848	0.7647
	Backtranslation + ProtAugment + Tweet Style	**0.9412**	0.8750	**0.9381**	**0.9454**	**0.9412**
	Nemotron-Mini (4B) + Phi-4 (14B)	0.8824	**0.8798**	0.8865	0.8975	0.8824
ModernBERT	AnnoMI (Base Data)	**0.9118**	**0.8125**	**0.9040**	**0.9209**	**0.9118**
	Backtranslation + ProtAugment + Tweet Style	0.8235	0.6683	0.8004	0.8137	0.8235
	Nemotron-Mini (4B) + Phi-4 (14B)	0.8235	0.6683	0.8004	0.8137	0.8235

Table 2: Performance of each LM with combined LLM- and NLP-augmented datasets.

LLMs are seamless, a few exhibit moderate difficulties in interpreting or following the chain of instructions. The most common issue is the inability to produce the desired outcome in the prescribed format for the same prompts. And Gemini, Llama, Mistral Nemo, Phi4, Qwen, and Smollm2 were prone to it.

- **Fluency and Emotional Understanding:** While all the LLMs have shown great fluency, capturing the actual depth of emotions and context is somewhat more challenging. For instance, DeepSeek, Gemini and LLama have been observed to elicit stronger emotional responses. Table 5 in Appendix I provides a deeper look into each LLM's performance and reflects the extent of MI adherence in $MI_{syn.}$, summarizing the main trends and observations about each LLM.

6 Conclusion and Future Work

In this study, we investigate the performance of 13 LLM for in-context, human-like MI dialogue generation at the entire conversation level, of which **DeepSeek** is the latest release and SOTA. We also investigate TT-based augmentation techniques applied at the utterance level to assess whether the context and intricacies of complex domains can still be captured at the utterance level. Further, to analyze the quality of LLM and TT-generated augmented data, we used BigBird, Longformer, and ModernBERT classifiers designed to process long sequences, of which ModenrBERT is the latest release and SOTA transformer model. Our results show for such a difficult task to perform classification over considering an entire conversation on a severely unbalanced and small dataset belonging to a complex domain, the classifiers have demonstrated exceptional performance, with the highest balanced score being **98.08%**. Then, we employ our meticulously developed evaluation scheme parametrized to measure semantic drift, hallucination, MI adherence, prompt compliance, fluency, context, and emotional interpretation in LLMs and facilitate fine-grained qualitative and quantitative analysis. As a resource contribution, we provide the first-ever high-quality dataset containing 1764 low- and high-quality MI dialogues and $\tilde{8}$1k talk turns between therapist and client. Based on our empirical analysis, ChatGPT, DeepSeek, and Gemini have shown impressive capabilities in using open-ended questions to encourage exploration in MI, reflecting feelings, and affirming emotions to build rapport, encouraging small, manageable steps instead of overwhelming directives, and fostering collaboration and autonomy, allowing the client to choose changes and that may lead to change talk. For future work, we aim to explore additional domain-specific LMs, investigate possibilities for fine-tuning LLMs on our data for domain-specific NLP downstream tasks (automatic MI session annotation, summary generation, etc.), and further refine our evaluation framework to extend its applicability to related areas of mental health research.

Limitations

While this work presents a comprehensive benchmark for session-level LLM-based data augmentation in Motivational Interviewing, several limitations should be acknowledged. First, although the curated dataset is the largest session-level MI resource used for this type of study to date, it remains modest in scale compared to general-domain dialogue corpora. The dataset also exhibits class im-balance that reflects real-world MI practice, which may influence classifier performance and limit generalization to settings with different label distributions.

Second, the MI sessions used in this study are drawn from a specific clinical and cultural context. As MI practices can vary across populations, languages, and institutional settings, the findings may not directly generalize to other forms of counseling or to non-English MI interactions without further validation.

Third, our data augmentation strategies rely on contemporary large language models whose outputs are inherently sensitive to prompt design, decoding parameters, and model updates. While we employ standardized prompts and controlled generation settings, different prompt formulations or future model versions may yield varying results. In addition, despite incorporating qualitative checks for hallucination and semantic drift, LLM-generated data may still introduce subtle artifacts that are not fully captured by automated evaluation metrics.

Fourth, our evaluation focuses on downstream MI session classification rather than clinical outcomes or human-centered measures of counseling effectiveness. Consequently, improvements in classification performance should not be interpreted as direct evidence of improved therapeutic quality or clinical utility. Any real-world deployment of such systems would require rigorous human-in-the-loop validation and ethical oversight.

Finally, while we examine multiple long-context classification models, our study does not exhaustively explore architectural variations or fine-tuning strategies. Future work could investigate alternative model designs, multilingual settings, and tighter integration of domain expert feedback during both data generation and evaluation.

References

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. 2024a. 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.

Simone Balloccu, Patrícia Schmidtová, Mateusz Lango, and Ondrej Dusek. 2024b. Leak, cheat, repeat: Data contamination and evaluation malpractices in closed-source LLMs. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 67–93, St. Julian's, Malta. Association for Computational Linguistics.

Erkan Basar, Iris Hendrickx, Emiel Krahmer, Gert-Jan Bruijn, and Tibor Bosse. 2024. To what extent are large language models capable of generating substantial reflections for motivational interviewing counseling chatbots? a human evaluation. In *Proceedings of the 1st Human-Centered Large Language Modeling Workshop*, pages 41–52, TBD. ACL.

Erkan Basar, Xin Sun, Iris Hendrickx, Jan de Wit, Tibor Bosse, Gert-Jan De Bruijn, Jos A. Bosch, and Emiel Krahmer. 2025. How well can large language models reflect? a human evaluation of LLM-generated reflections for motivational interviewing dialogues. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 1964–1982, Abu Dhabi, UAE. Association for Computational Linguistics.

Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The long-document transformer.

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.

Andrew Brown, Ash Tanuj Kumar, Osnat Melamed, Imtihan Ahmed, Yu Hao Wang, Arnaud Deza, Marc Morcos, Leon Zhu, Marta Maslej, Nadia Minian, Vidya Sujaya, Jodi Wolff, Olivia Doggett, Mathew Iantorno, Matt Ratto, Peter Selby, and Jonathan Rose. 2023. A motivational interviewing chatbot with generative reflections for increasing readiness to quit smoking: Iterative development study. *JMIR Ment Health*, 10:e49132.

Andrew Brown, Jiading Zhu, Mohamed Abdelwahab, Alec Dong, Cindy Wang, and Jonathan Rose. 2024. Generation, distillation and evaluation of motivational interviewing-style reflections with a foundational language model. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1241–1252, St. Julian's, Malta. 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, and 12 others. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

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.

Dorottya Demszky, Diyi Yang, David S Yeager, Christopher J Bryan, Margarett Clapper, Susannah Chandhok, Johannes C Eichstaedt, Cameron Hecht, Jeremy Jamieson, Meghann Johnson, and 1 others. 2023. Using large language models in psychology. *Nature Reviews Psychology*, 2(11):688–701.

Kaustubh D. Dhole, Varun Gangal, Sebastian Gehrmann, Aadesh Gupta, Zhenhao Li, Saad Mahamood, Abinaya Mahendiran, Simon Mille, Ashish Srivastava, Samson Tan, Tongshuang Wu, Jascha Sohl-Dickstein, Jinho D. Choi, Eduard Hovy, Ondrej Dusek, Sebastian Ruder, Sajant Anand, Nagender Aneja, Rabin Banjade, and 106 others. 2023. Nl-augmenter: A framework for task-sensitive natural language augmentation.

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*.

Ji-Eun Han, Jun-Seok Koh, Hyeon-Tae Seo, Du-Seong Chang, and Kyung-Ah Sohn. 2024. PSYDIAL: Personality-based synthetic dialogue generation using large language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 13321–13331, Torino, Italia. ELRA and ICCL.

Stefan Harrer. 2023. Attention is not all you need: the complicated case of ethically using large language models in healthcare and medicine. *EBioMedicine*, 90:104512.

Vivek Kumar, Giacomo Medda, Diego Reforgiato Recupero, Daniele Riboni, Rim Helaoui, and Gianni Fenu. 2023. 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, Pushpraj Singh Rajawat, and Eirini Ntoutsi. 2025. Mitigating semantic drift: Evaluating llms' efficacy in psychotherapy through mi dialogue summarization. *Preprint*, arXiv:2511.22818.

Vivek Kumar, Pushpraj Singh Rajwat, Giacomo Medda, Eirini Ntoutsi, and Diego Reforgiato Recupero. 2024. Unlocking LLMs: Addressing scarce data and bias challenges in mental health and therapeutic counselling. In *Proceedings of the First International Conference on Natural Language Processing and Artificial Intelligence for Cyber Security*, pages 238–251, Lancaster, UK. International Conference on Natural Language Processing and Artificial Intelligence for Cyber Security.

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.

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.

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.

Emre Sezgin. 2023. Artificial intelligence in healthcare: Complementing, not replacing, doctors and healthcare providers. *Digit. Health*, 9:20552076231186520.

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

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.

Xin Sun, Jiahuan Pei, Jan de Wit, Mohammad Aliannejadi, Emiel Krahmer, Jos T.P. Dobber, and Jos A. Bosch. 2024. Eliciting motivational interviewing skill codes in psychotherapy with LLMs: A bilingual dataset and analytical study. In *Proceedings of* the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 5609–5621, Torino, Italia. ELRA and ICCL.

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

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.

Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, Nathan Cooper, Griffin Adams, Jeremy Howard, and Iacopo Poli. 2024. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference.

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.

Yuqi Wu, Kaining Mao, Yanbo Zhang, and Jie Chen. 2024. Callm: Enhancing clinical interview analysis through data augmentation with large language models. *IEEE Journal of Biomedical and Health Informatics*, 28(12):7531–7542.

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.

Xuhai Xu, Bingsheng Yao, Yuanzhe Dong, Saadia Gabriel, Hong Yu, James Hendler, Marzyeh Ghassemi, Anind K. Dey, and Dakuo Wang. 2024. Mental-llm: Leveraging large language models for mental health prediction via online text data. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 8(1).

Kailai Yang, Tianlin Zhang, Ziyan Kuang, Qianqian Xie, Jimin Huang, and Sophia Ananiadou. 2024. Mentallama: Interpretable mental health analysis on social media with large language models. In *Proceedings of the ACM Web Conference 2024*, WWW '24, page 4489–4500, New York, NY, USA. Association for Computing Machinery.

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.

Manzil Zaheer, Guru Guruganesh, Kumar Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, and Amr Ahmed. 2020. Big bird: Transformers for longer sequences. In *Advances in Neural Information Processing Systems*, volume 33, pages 17283–17297. Curran Associates, Inc.

Ethical Considerations

This work focuses on methodological evaluation and benchmarking rather than deployment in real-world clinical settings. All MI data used in this study are de-identified and handled in accordance with applicable data protection guidelines. While our use of large language models enables scalable data augmentation, such models may generate hallucinated or misleading content, particularly in sensitive mental health contexts. Accordingly, the generated data and trained models should not be used for diagnostic or therapeutic purposes without rigorous human oversight. Any future application of these methods in practice would require careful validation, domain expert involvement, and ethical review to ensure patient safety and responsible use.

A Related Work

The research field of integrating LLMs into therapeutic settings, particularly MI, is still evolving, and there is not much work done to answer specific research questions (Wu et al., 2024; Yuan et al., 2023; Li et al., 2023b; Tang et al., 2023), such as data augmentation at the MI session level, using LLMs for mass annotation tasks, etc. In fact, since 2016, only a few works (Pérez-Rosas et al., 2016; Welivita and Pu, 2022; Kumar et al., 2023; Wu et al., 2023) have contributed to the development of public datasets, as mentioned in (Sun et al., 2024). (Balloccu et al., 2024a) can be considered a holistic work to generate the HAI-Coaching dataset by collecting approximately 2,400 real-world dietary struggles and around 97,000 supportive responses using LLMs, evaluated by nutrition experts for safety and quality. Among a few available works, (Basar et al., 2025) evaluates the ability of LLMs to generate reflective listening responses in MI dialogues. The study assesses whether LLM-generated responses align with human therapists' reflections and effectiveness in therapeutic conversations. Early credits can be given to (Brown et al., 2023) to invoke MI-based conversation to help smokers move toward quitting smoking. In the subsequent work, (Brown et al., 2024) investigates how LLMs can be fine-tuned to generate MI-style reflections for use in therapeutic chatbots. It evaluates the effectiveness of AI-generated reflections in maintaining engagement and supporting behavior change. The work in (Basar et al., 2024) focuses on the quality, coherence, and therapeutic value of AI-generated reflections compared to human-generated ones. It provides insights into how LLMs can generate meaningful reflections in counseling. As evident from the above works, there are several gaps in the research of LLM integration in effective and reliable healthcare services. Therefore, with this work, we aim to provide a clear path to integrate LLMs in the therapeutic scenario for benchmarking, data augmentation, automatic evaluation of MI sessions, using SOTA, and further improvement.

B High- and Low-quality MI Sessions

C Risk Minimization Problem

D Defining An Augmenter for TTs

E Prompt-based Generation Example

F Distribution of Talk-turns in LLM-LLM-augmented datasets

G Classification Metrics For Different Models and LLM-generated Augmented Data.

H Evaluation Scheme MI Dialogues Generated by LLMs

I Fine-grained Evaluation of LLMs Performance

MI Dialogue (High-quality)

Therapist	What else has brought you here?
Client	They found a little bag of weed.
Therapist	Mm-hmm.
Client	It's not like I was gonna sell it or anything. It's like
Therapist	Just enough for your own use?
Client	Yeah, whatever, you know. It's California, no one cares.
Therapist	Yeah. Yeah. Yeah.
Client	How long is this gonna take?
... ...	... ...
Therapist	Okay, mm-hmm. Well, I'm wondering then if it's okay with you and-and I have a form that we can use if you'd like, um, just to put out some different ideas on an agenda that we could share as we're working together so that I-I can support you to be successful in getting this behind you. Would-would this feel like a natural time to-to move towards figuring out where we might go together?

MI Dialogue (Low-quality)

Therapist	Hi. You are?
Client	Yeah.
Therapist	Okay. Um, what brings you in?
Client	Well, I'm supposed to come to learn about my-- how to manage my diabetes. Just got diagnosed.
Therapist	Mm-hmm. I see- I see that your A1C or your blood glucose is very high.
... ...	... ...
Client	That's, uh, that's a huge amount of stuff to do and, I'm not- I'm not certain I can do all that, and on top of it, it's-it's-- if I don't get it all done, all this stuff happens to me?
Therapist	So, it sure can.
Client	[sighs] So, uh, you know, we-we need to start looking at doing all those different things to get your diabetes under control 'cause your-your
Therapist	number is very high.
Client	Well, I-I

Figure 1: A sample of a). high-quality and b) low-quality MI sessions.

$$
\begin{aligned}
\theta^*(\phi) &= \arg\min_{\theta} \sum_{(S,q)\in\mathcal{D}_{train}\cup\phi(\mathcal{D}_{train})} \mathcal{L}\big(f_\theta(\Gamma(S)), q\big) \\
\phi^* &= \arg\min_{\phi\in\Phi} \sum_{(S,q)\in\mathcal{D}_{test}} |f_{\theta^*(\phi)}(\Gamma(S)) - q|
\end{aligned}
\tag{3}
$$

Where $\mathcal{L}(\cdot,\cdot)$ is a training loss function for classification tasks, e.g., binary cross-entropy. In other words, our objective is to identify the augmenter that can raise the bar in MI quality classification by minimizing the misclassification error on a fixed testing set.

$$
\begin{aligned}
\phi^S_{TT}(S_i) = \phi^S_{TT}(t_{i1}, t_{i2}, \ldots, t_{iM_i}) &= \left(\phi^t_{TT}(t_{i1}), \phi^t_{TT}(t_{i2}), \ldots, \phi^t_{TT}(t_{iM_i})\right), \\
\phi_{TT}(\mathcal{D}_{train}) = \left\{(\phi^S_{TT}(S_i), q_i)\right\}_{i=1}^{L} &= \left\{\left(\phi^t_{TT}(t_{i1}), \ldots, \phi^t_{TT}(t_{iM_i})\right), q_i\right\}_{i=1}^{L}
\end{aligned}
\tag{4}
$$

Where $\phi^t_{TT} : \mathbb{R}^e \to \mathbb{R}^e$ denotes a TT sub-function that operates at the talk level. Note that the TTs adopted at the utterance level imply $|\mathcal{D}_{Aug}| = |\mathcal{D}_{train}|$, unless ϕ^t_{TT} is repeatedly applied to the same talk. In other words, we constrained TTs augmenters to generate the same amount of utterances included in the original training set, maintaining also the same order.

System Prompt

You are an expert psychologist specializing in motivational interviewing with exceptional rephrasing abilities and a deep understanding of therapeutic techniques.
Your task is to rephrase each line of a therapy session to enhance clarity, empathy, and professional tone without altering the core meaning or intent of the conversation.

Task Instructions:
- Review the context of each interaction between the client and therapist.
- Identify implicit emotional or cognitive themes in the client's responses.
- Rephrase each line to improve tone, empathy, or clarity, reflecting motivational interviewing principles, including open-ended questions, reflective listening, and affirmation.
- Do not create any new content or expand on any ideas.

Chain of Thought Process:
- Step 1: Carefully review each client and therapist utterance.
- Step 2: Interpret the client's underlying emotions or cognitive themes.
- Step 3: Rephrase each response thoughtfully, ensuring it aligns with the original meaning while subtly improving the dialogue.
- Step 4: Preserve the original conversational flow and structure without adding new lines or content.

Output Format:
- Respond only with rephrased dialogue lines in the format:
Therapist: [rephrased therapist response]
Client: [rephrased client response]
- Avoid any narrative descriptions, explanations, or additional text outside of the direct dialogue lines.

User Prompt

Please rephrase the following therapy session, focusing on improving clarity, empathy, and tone while preserving the original meaning and flow.

Here is the session to rephrase:

Client: I've been feeling like my wife has been on my case a lot lately about improving my health habits, but to be honest, I don't really see what the issue is.
Therapist: Just to make sure I understand, you're feeling a bit puzzled by your wife's concerns about your health habits?
Client: Yeah, I feel fine. I mean, my job is physically demanding, so I'm always on the go and sweating out anything unhealthy I might eat.
Therapist: That makes sense. It sounds like you're relying on your physical activity to balance out your diet.

Sample MI Session

Provide only rephrased dialogue in the format:
Therapist: [rephrased therapist response]
Client: [rephrased client response]

Do not add new dialogue or text outside the provided lines.

LLM Output

Therapist: Just to clarify, it sounds like you're feeling uncertain about why your wife is so concerned about your health habits
Client: Yeah, I feel fine. My job keeps me physically active, so I figure I'm burning off anything unhealthy I eat.
Therapist: I see. It sounds like you rely on your level of physical activity to counterbalance your diet.

Figure 2: A structured chat interface showcasing a system prompt, user prompt, and LLM output in a MI rephrasing task. The system prompt instructs the model to enhance clarity, empathy, and professional tone while preserving meaning. The user provides a sample therapy session, and the LLM generates a refined version, demonstrating its ability to improve conversational quality while maintaining fidelity to the original dialogue.

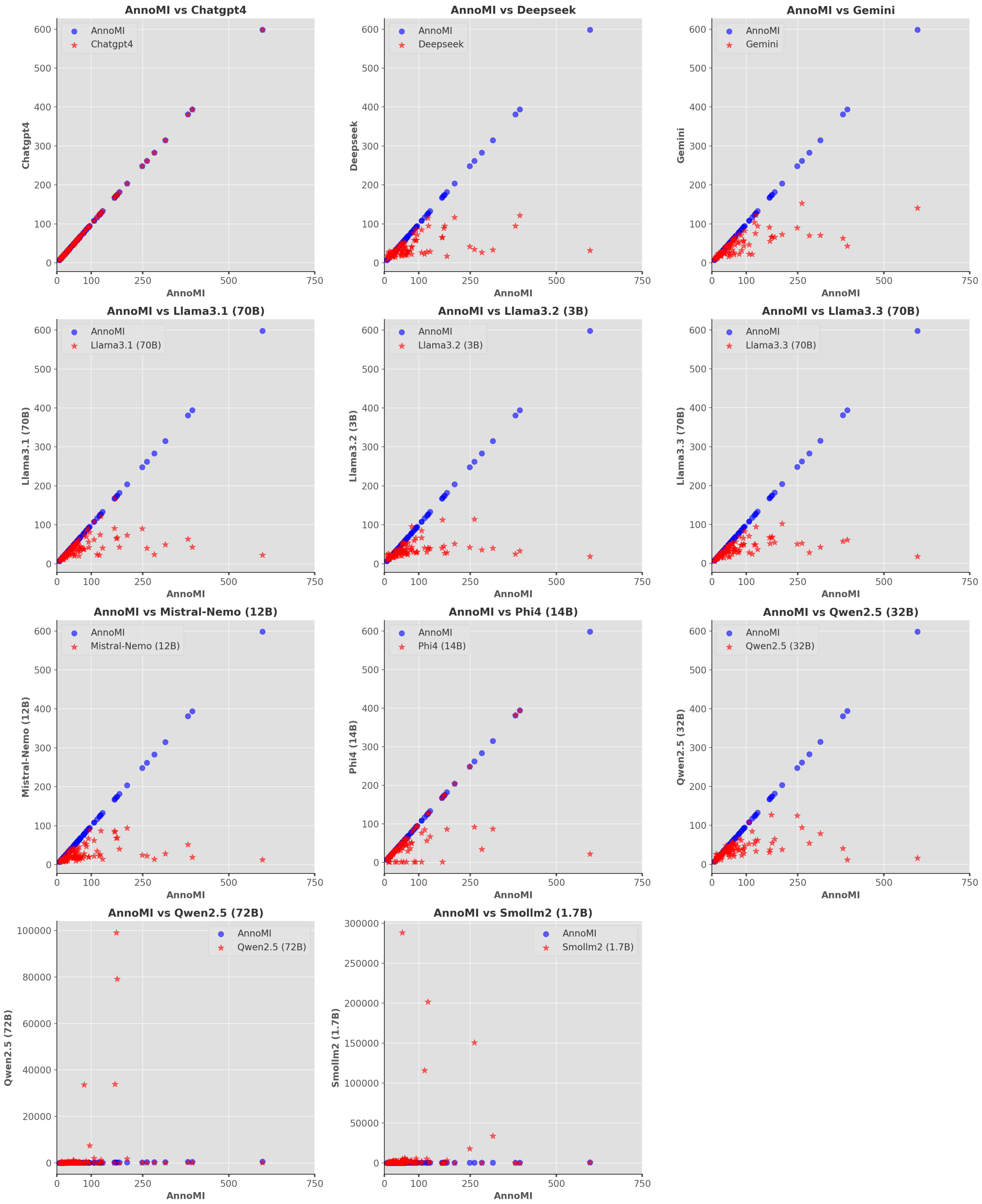

Figure 3: Distribution of the number of turn talks for each $MI_{orig.}$ vs $MI_{syn.}$ for all LLM-augmented datasets.

Model	Augmented Data	Acc	Bal. Acc.	F1	Prec	Rec
BigBird	AnnoMI (Base Data)	0.8824	0.9231	0.8892	0.9216	0.8824
	Backtranslation	**0.9706**	**0.9808**	**0.9712**	**0.9739**	**0.9706**
	Basic style	0.8529	0.7308	0.8400	0.8475	0.8529
	ProtAugment	0.9412	0.8750	0.9381	0.9454	0.9412
	Synonym substitution	0.8235	0.6250	0.7797	0.8566	0.8235
	Tweet style	0.9118	0.8558	0.9097	0.9097	0.9118
Longformer	AnnoMI (Base Data)	0.7647	0.5000	0.6627	0.5848	0.7647
	Backtranslation	0.8824	0.8798	0.8865	0.8975	0.8824
	Basic style	0.7941	0.5625	0.7263	0.8378	0.7941
	ProtAugment	**0.9412**	**0.9183**	**0.9412**	**0.9412**	**0.9412**
	Synonym substitution	0.8235	0.7548	0.8235	0.8235	0.8235
	Tweet style	**0.9412**	**0.9183**	**0.9412**	**0.9412**	**0.9412**
ModernBERT	AnnoMI (Base Data)	0.9118	0.8125	0.9040	0.9209	0.9118
	Backtranslation	**0.9412**	**0.9183**	**0.9412**	**0.9412**	**0.9412**
	Basic style	0.7941	0.6490	0.7760	0.7740	0.7941
	ProtAugment	0.8235	0.6683	0.8004	0.8137	0.8235
	Synonym substitution	0.8529	0.7308	0.8400	0.8475	0.8529
	Tweet style	0.9118	0.8558	0.9097	0.9097	0.9118

Table 3: Classification Metrics for different models and NLP-augmented data

Model	Semantic Drift			Hallucination			Session Length			Prompt Compliance			Fluency & Emotion		
	High	Avg	Low	High	Avg	Low	Longer	Equal	Shorter	High	Avg	Low	High	Avg	Low
ChatGPT			✓			✓		✓		✓			✓		
DeepSeek			✓			✓			✓	✓				✓	
Gemini		✓			✓				✓		✓		✓		
Llama		✓			✓				✓		✓			✓	
Mistral Nemo		✓			✓				✓		✓			✓	
Nemotron Mini			✓			✓	✓			✓				✓	
Phi-4		✓			✓				✓		✓		✓		
Qwen		✓				✓			✓		✓			✓	
SmolLM2		✓			✓		✓				✓			✓	

Table 4: Evaluation scheme for analyzing the quality of synthetic MI dialogues generated for each LLM.

Model	Semantic Drift	Hallucination	MI Session Length	MI Adherence	Fluency & Emotion
ChatGPT	Very low drift, preserves original intent	Minimal hallucination, rarely adds content	Matches original length closely	Excellent compliance with MI structure	Highly fluent with strong emotional understanding
DeepSeek	Moderate drift, restructures reflections	Rare hallucination but simplifies responses	Slightly condensed compared to original	Adheres to MI but compresses content	Fluent but lacks natural expressiveness
Gemini Flash	Minimal drift, sometimes reinterprets tone	Rare, but adds motivational cues	Slightly longer due to clarifications	Strong adherence but slightly directive	Highly fluent with emotional depth
Llama	Moderate drift, explains rather than reflects	Occasionally adds interpretations	Inconsistent, some responses very brief	Follows MI structure but shifts therapist tone	Fluent but robotic in phrasing
Mistral Nemo	Moderate drift, rewords client intent	Occasionally adds emotional details	Inconsistent, some responses long	Adheres to MI but can be directive	Fluent but lacks emotional nuance
Nemotron Mini	Low-Moderate drift, moderate rewording	Rare hallucination but adds details	Mostly longer than original	Good adherence but over explanation	Fluent but lacks emotional context
Phi4	Low-moderate drift, reframes statements	Adds clarifications, making responses longer	Slightly longer than original	Strong MI adherence but overuses affirmations	Well-structured with moderate emotional depth
Qwen2.5	Low drift, over-summarizes at times	Avoids hallucination but loses nuance	Shorter, condenses responses	Good adherence but loses open-ended depth	Fluent but less expressive
Qwen2.5	Minimal drift, slight rewording	Doesn't add but removes details	Generally shorter than original	Good compliance but oversimplifies	Fluent but lacks emotional depth
Smollm2	Moderate-High drift, restructures statements	Moderate hallucination, adds extra details	Mostly longer than original	Adheres to MI but expands content	Fluent but lacks emotional context

Table 5: An intuitive evaluation of LLMs based on the LLM-augmented data performance.