Department of Computer Science and Engineering
Islamic University of Technology

BACHELOR OF SCIENCE IN COMPUTER SCIENCE AND ENGINEERING

A Framework for Multi-Turn Mental Health Dialogue Expansion and LLM-Based Evaluation

Authors
Asif Abrar - 200041106
Shajeed Hossain - 200041142
Tanvir Hossain Dihan - 200041144

Supervisor
Dr. Hasan Mahmud
Professor
Department of Computer Science and Engineering

October, 2025

Declaration of Candidate

This is to certify that the work presented in this thesis is the outcome of the analysis and experiments carried out by **Asif Abrar**, **Shajeed Hossain**, and **Tanvir Hossain Dihan** under the supervision of **Dr. Hasan Mahmud**, Professor, Department of Computer Science and Engineering, Islamic University of Technology, Dhaka, Bangladesh. It is also declared that neither this thesis nor any part of it has been submitted anywhere else for any degree or diploma. Information derived from the published and unpublished work of others have been acknowledged in the text and a list of references is given.

Dr. Hasan Mahmud
Professor
Department of Computer Science and Engineering
Islamic University of Technology (IUT)
Date: October 08, 2025

Asif Abrar
Student ID: 200041106
Date: October 08, 2025

Shajeed Hossain
Student ID: 200041142
Date: October 08, 2025

Tanvir Hossain Dihan
Student ID: 200041144
Date: October 08, 2025

Acknowledgement

We would like to express our grateful appreciation for **Dr. Hasan Mahmud**, Professor, Department of Computer Science & Engineering, IUT, for being our adviser and mentor. His motivation, suggestions, and insights for this research have been invaluable. Without his support and proper guidance, this research would never have been possible. We are really grateful to him.

We are also thankful to the Department of CSE, IUT, for providing us the resources and environment to carry out this research.

To our families and friends, thank you for your constant support and belief.

Lastly, to our team members for the countless late nights, shared snacks, and mutual encouragement that kept this journey going.

Contents

1 Introduction **1**

1.0.1 Thesis Overview 2
1.0.2 Motivations and Scope 2
1.0.3 Problem Statement 3
1.0.4 Research Challenges 3
1.0.5 Contribution 3

2 Literature Review **5**

2.1 Conversion of Single-turn to Multi-turn Conversations in Mental Health Dialogue Systems 5
2.1.1 Advantages of Multi-turn Generation from Single-turn Input . 6
2.1.2 Evaluation and Validation 7
2.2 Evaluating Large Language Models through LLM-as-a-Judge 9
2.2.1 Advantages of the LLM-as-a-Judge Framework 10
2.2.2 Limitations 10
2.2.3 Evaluation and Validation 11
2.3 Human Evaluation of LLMs in Healthcare: The QUEST Framework . 11
2.3.1 Advantages and Applicability of the QUEST Framework . . . 12
2.3.2 Limitations 13
2.3.3 Evaluation Process and Statistical Rigor 14
2.3.4 Implications for Mental Health Chatbots 14
2.4 Foundations of Inter-Rater Reliability: Intraclass Correlation Coefficients (ICC) 14
2.4.1 ICC Model Classifications 15
2.4.2 Mathematical Formulation of ICC(3,1) 15
2.4.3 Advantages 16
2.4.4 Limitations 16
2.4.5 Impact and Relevance to Current Research 17

2.5 ChatCounselor: A Large Language Model for Mental Health Support . 17
2.5.1 Motivation . . . 17
2.5.2 Contributions . . . 18
2.5.3 Advantages . . . 19
2.5.4 Limitations . . . 20
2.5.5 Impact and Significance . . . 20
2.6 Foundation Metrics for Evaluating Effectiveness of Healthcare Conversations Powered by Generative AI . . . 21
2.6.1 Motivation . . . 21
2.6.2 Contribution and Advantages . . . 21
2.6.3 Limitations . . . 23
2.6.4 Conclusion . . . 24
2.7 Rethinking the Evaluation of Dialogue Systems: Effects of User Feedback on Crowdworkers and LLMs . . . 24
2.7.1 Sparse Human Evaluation via Crowdsourcing . . . 24
2.7.2 Advantages and Contribution . . . 25
2.7.3 Limitations . . . 25
2.7.4 Conclusion . . . 26
2.8 An Empirical Analysis of Uncertainty in Large Language Model Evaluations . . . 26
2.8.1 Confidence-Weighted Scoring Framework . . . 27
2.8.2 Advantages and Contributions . . . 27
2.8.3 Limitations . . . 28
2.8.4 Relevance to Mental Health Dialogue Evaluation . . . 28
2.8.5 Conclusion . . . 28

3 Proposed Methodology **29**
3.1 Overview . . . 29
3.2 Chronological Breakdown . . . 30
3.2.1 Step 1: Dataset Expansion . . . 30
3.2.2 Step 2: Fine-Tuning the LLaMA Model . . . 30
3.2.3 Step 3: Conversation Simulation with Other LLMs . . . 31
3.2.4 Step 4: Choosing the metrics . . . 31
3.2.5 Step 5: Evaluation by LLM-Based Judges . . . 32
3.2.6 Step 6: Single vs Multi-turn Comparison . . . 32
3.2.7 Step 7: Sparse Human Evaluation Comparison . . . 32

4 Experimental Setup **34**

4.1 Coding environment . . . 34
4.2 Data Preparation . . . 34
4.3 Model Fine-Tuning . . . 34
4.4 Conversation Simulation . . . 35
4.5 Evaluation Metrics . . . 35
4.6 Evaluation Protocol . . . 35
4.7 Sparse Human Evaluation . . . 35

5 Results and Discussion **36**
5.1 Presenting the Results . . . 36
5.1.1 Human Evaluators' opinions on converting Single-turn to Multi-turn data . . . 36
5.1.2 Human and LLM Evaluation Scores . . . 37
5.1.3 Single and Multi-turn Evaluation Scores . . . 38
5.1.4 Single and Multi-turn Human Likeness metric scores . . . 38
5.1.5 Wins/Ties/Loss Percentage for Each Metric (Single vs Multi-turn) 39
5.2 Challenges and Limitations . . . 40
5.2.1 Lack of Poor-Quality Examples . . . 40
5.2.2 Limited Variance Across Conversations . . . 40
5.2.3 Moderate Inter-Rater Reliability Indicated by ICC Score . . . 41
5.2.4 Generalization Limitations . . . 41
5.2.5 Synthetic Nature of the Dataset . . . 41

6 Conclusion **42**
6.1 Restatement of Research Objectives . . . 42
6.2 Summary of Key Findings . . . 42
6.3 Contributions, Limitations, and Critical Discussion . . . 43
6.4 Future Directions and Final Reflections . . . 43

References **44**

List of Figures

2.1 The SMILE method used to generate dialogues for mental health support 5
2.2 The SMILE method used to generate dialogues for mental health support 7
2.3 The LLM-as-a-Judge framework used for scalable evaluation 9
2.4 The QUEST evaluation framework proposed by Tam et al. for assessing LLMs in healthcare 12
2.5 Some Examples of Metrics of the QUEST Evaluation Framework . . . 13
2.6 Comparison between raw transcripts and GPT-4 processed query-answer pairs 18
2.7 Pipeline of ChatCounselor 19
2.8 GPT-4 evaluation 20
2.9 Overview of the four healthcare evaluation metric groups 22
2.10 A broad overview of the evaluation process and the role of metrics. groups 23

3.1 Flow Chart of the Proposed Methodology 30

5.1 Line Chart showing Comparison of Human and LLM Average Scores 37
5.2 Line Chart showing Comparison of Single and Multi-turn Average Scores 38
5.3 Comparison of Single and Multi-turn Human Likeness evaluation scores 38
5.4 Bar Chart showing Wins/Ties/Losses for Each Metric (Single vs Multi-turn) 39

List of Tables

5.1 Voting results for multi-turn conversations generated from single-turn inputs. 36
5.2 Comparison of Single-turn and Multi-turn wins across metrics. 40

Abstract

In this present world struggling with mental well-being, why would AI stay behind from playing a crucial role in our lives to better manage our mental health. This paper introduces a pipeline for building and evaluating multi-turn mental health conversation using limited single-turn datasets. Initially, publicly available mental health QA pairs are expanded into realistic multi-turn conversations via prompt-based generation. These enriched dialogues are used to fine-tune a LLaMA-based model, which is trained to play the role of a virtual psychiatrist. To simulate diverse interaction scenarios, the chatbot engages in conversations with other large language models acting as synthetic patients. The resulting dialogues are evaluated by three independent LLMs, each assessing the chatbot's performance across mental health support metrics such as reliability, bias, sensibility, specifity and interestingness(SSI), safety and security, empathy, robustness and human likeness. Final scores are weighted averaged to ensure balanced evaluation keeping a confidence. Other than these, a sparse human evaluation based comparison among the LLM ratings is also done in this work to see how well the LLMs can produce judgments with respect to human evaluators. This work highlights a novel approach that combines synthetic data expansion, LLM-based simulation, and automated evaluation to improve the development and assessment of mental health dialogue systems.

Chapter 1

Introduction

In recent years, the intersection of artificial intelligence and mental health care has given rise to conversational agents capable of providing basic emotional support and mental health assistance. However, most existing systems are constrained by two key limitations: **a reliance on single-turn dialogue data and the absence of scalable, nuanced evaluation frameworks**. Real-world therapeutic conversations are inherently multi-turn, evolving as patients gradually open up and receive guidance in context-rich exchanges. Single-turn question-answer datasets, though useful, fail to capture this dynamic, limiting the depth and realism of current mental health chatbots.

This research addresses these gaps by proposing a comprehensive framework to transform single-turn mental health dialogues into multi-turn interactions and to evaluate them using large language models (LLMs). The pipeline begins by augmenting public single-turn QA pairs **Psych8k** from the paper **ChatCounselor: A Large Language Models for Mental Health Support**[5] into multi-turn conversations using prompt-based generation with an instruction-following LLM (GEMINI-2.5 API). These expanded conversations are then used to fine-tune a LLama-based model, *Llama-3.1:8B instruct*, that assumes the role of a virtual psychiatrist capable of engaging users in coherent and emotionally appropriate dialogue.

To rigorously test the systems performance, we simulate real-world conversations by pairing the fine-tuned model with other LLMs acting as synthetic mental health patients. These LLMs introduce varied user behaviors and mental health concerns, allowing for rich and realistic dialogue generation. To evaluate each conversation individually by LLMs, we have followed the approach of the paper **Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena** [17] The conversations are then

evaluated by a panel of three independent LLMs (Deepseek-r1, Qwen-2.5, Mistral) based on predefined mental health support criteria, such as reliability, bias, sensibility,specifity and interestingness(SSI), safety and security, empathy, robustness and human likeness. The weighted averaged evaluation across models having confidence scores yields a more robust and unbiased assessment of the chatbots' therapeutic effectiveness.

By combining synthetic data expansion, LLM-based simulation, and ensemble evaluation, this research contributes a novel and scalable approach to both the development and validation of mental health dialogue systems. The proposed methodology not only mitigates data scarcity but also reduces reliance on human annotators, paving the way for safer, more accessible, and ethically responsible mental health support through conversational AI.

1.0.1 Thesis Overview

In this thesis, we explore a systematic framework for developing and evaluating multi-turn mental health dialogue systems using large language models (LLMs). Chapter 1 introduces the motivation, problem statement, research challenges, and core contributions of the study, highlighting the need for realistic, scalable, and ethically responsible AI support in mental health contexts. Chapter 2 presents an extensive literature review covering the transformation of single-turn to multi-turn conversations, evaluation methods such as LLM-as-a-Judge and QUEST frameworks, reliability metrics like ICC, and related works including ChatCounselor and other healthcare evaluation approaches. Chapter 3 details the proposed methodologybeginning from dataset expansion to fine-tuning, simulation, metric selection, and multi-model evaluation-establishing a pipeline for automated dialogue generation and assessment. Chapter 4 outlines the experimental setup, including tools, environment, model configurations, and evaluation protocols used to implement the framework. Chapter 5 discusses the results and findings, comparing single-turn and multi-turn performances, analyzing human versus LLM judgments, and identifying strengths and limitations of the proposed approach. Finally, Chapter 6 concludes the thesis by summarizing key outcomes, discussing contributions and limitations, and offering directions for future research in LLM-driven mental health dialogue evaluation.

1.0.2 Motivations and Scope

Mental health concerns are rising globally, with an increasing demand for accessible, stigma-free, and low-cost support solutions. Conversational agents have shown

promise in delivering scalable mental health interventions, but their capabilities remain limited due to lack of context-awareness, empathy, and realistic dialogue flow. Most publicly available datasets offer only single-turn interactions, which fail to reflect the dynamic, evolving nature of therapeutic exchanges. This research is motivated by the potential to bridge that gap using recent advancements in large language models (LLMs). The scope of this work includes the transformation of single-turn mental health dialogues into multi-turn conversations, the development of a context-aware mental health chatbot, and the design of an automated evaluation pipeline using multiple LLMs to simulate both users and evaluators.

1.0.3 Problem Statement

Existing mental health chatbots lack the ability to hold deep, emotionally responsive conversations due to their training on limited single-turn data. Furthermore, there is no scalable or standardized way to evaluate their performance in realistic therapeutic scenarios. Manual evaluation by mental health professionals is ideal but impractical due to privacy concerns and high resource demands. Thus, the core problem this research addresses is: *How can we develop a multi-turn mental health chatbot trained on minimal data and evaluated automatically using LLM-based simulation to ensure coherence, empathy, and therapeutic relevance?*

1.0.4 Research Challenges

Several challenges arise in pursuing this goal. First, generating high-quality multi-turn dialogues from single-turn pairs requires careful prompt design to preserve semantic and emotional continuity. Second, fine-tuning a large language model like LLaMA demands precise curation of conversation structure and tone, especially in sensitive mental health contexts. Third, simulating users using LLMs introduces unpredictability and requires constraints to maintain meaningful interaction. Lastly, designing an LLM-based evaluator system involves defining robust mental health metrics and ensuring consistent, interpretable scoring across models. Moreover, the inter-data reliability is being questioned due to the automated outputs of the LLMs and lastly, the hardware and software insufficiency adds to the progress when trying to implement better and paid models that may be too big to run in our local resources.

1.0.5 Contribution

This paper makes the following key contributions:

- A pipeline to transform single-turn mental health question-answer pairs into rich multi-turn conversations using prompt-based generation.
- Sparse human evaluated scores comparison with the LLM performances.
- Fine-tuning of a LLaMA-based language model to serve as a virtual psychiatrist in contextually rich dialogues.
- A novel simulation framework where other LLMs act as diverse synthetic patients, enabling automated conversation generation and testing.
- An evaluation system involving multiple independent LLMs that assess chatbot performance on empathy, coherence, and therapeutic value based on a weighted average with confidence scores for each.
- A scalable, human-free validation strategy for mental health dialogue systems that balances realism, reproducibility, and ethical considerations.

This chapter introduced the motivation, scope, and core challenges associated with developing an effective multi-turn mental health conversation and dialogue system. It outlined the limitations of existing systems, defined the central research problem, and presented the methodological contributions of this thesis. By leveraging large language models for both generation and evaluation, this work aims to bridge the gap between realistic mental health conversations and scalable AI-driven support. The subsequent chapters will elaborate on the proposed methodology, implementation details, experimental setup, and evaluation outcomes that collectively validate the effectiveness of the system.

Chapter 2

Literature Review

2.1 Conversion of Single-turn to Multi-turn Conversations in Mental Health Dialogue Systems

Developing robust and empathetic dialogue systems for mental health support remains a pressing challenge in the field of natural language processing (NLP). One significant bottleneck is the lack of publicly available, high-quality, multi-turn conversational datasets that reflect realistic psychological counseling sessions. While many public datasets provide only single-turn question-answer (QA) pairs, these fall short of capturing the dynamic and iterative nature of actual therapeutic conversations, which typically unfold over multiple exchanges.

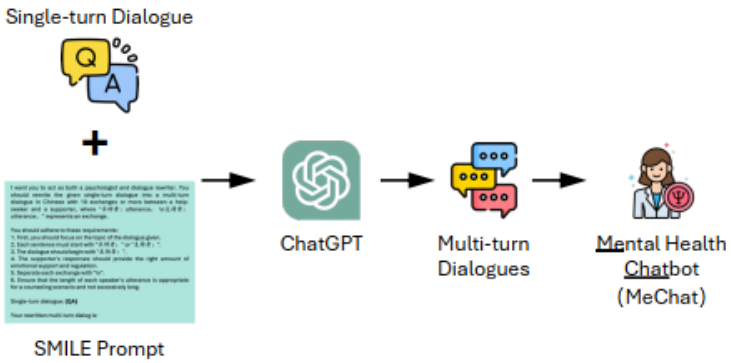

Figure 2.1: The SMILE method used to generate dialogues for mental health support

To overcome this limitation, Qiu et al. introduced a novel framework named **SMILE** (Single-turn to Multi-turn Inclusive Language Expansion) [7]. The core idea behind SMILE is to leverage the generation capabilities of large language models specifically ChatGPT to automatically convert single-turn dialogues into rich, emotionally

aware multi-turn conversations. The source data is derived from *PsyQA*, a Chinese QA dataset focused on mental health, which the authors preprocess and repurpose for multi-turn dialogue generation.

The SMILE methodology employs a carefully designed prompt that instructs the model to simulate a natural conversation between a help-seeker and a supporter. These conversations are structured to begin with the help-seeker, followed by alternating responses from both roles, maintaining coherence and emotional relevance throughout at least ten exchanges. The responses are tailored to provide appropriate emotional regulation and support, ensuring a high degree of realism and therapeutic value.

2.1.1 Advantages of Multi-turn Generation from Single-turn Input

The transformation from single-turn to multi-turn dialogues brings several distinct advantages:

- **Scalability and Accessibility:** SMILE allows the generation of large-scale datasets without the ethical and logistical complexities associated with collecting real-life counseling transcripts. The resulting dataset, *SMILECHAT*, comprises over 55,000 multi-turn dialogues, making it one of the largest corpora in this domain.
- **Improved Conversational Depth:** Unlike single-turn dialogues, multi-turn conversations enable a gradual unfolding of user concerns, allowing more nuanced expressions of emotional states and more tailored supportive responses from the system. This reflects the iterative process found in real counseling settings.
- **Increased Diversity:** The authors evaluate lexical, semantic, and topical diversity using metrics such as Distinct-n and cosine similarity, showing that SMILE-generated dialogues are significantly more varied and lifelike than those generated by baseline methods. This diversity helps avoid repetitive or generic model outputs during training.
- **Enhanced Training Outcomes:** A chatbot named *MeChat*, fine-tuned on the SMILECHAT dataset, achieved strong performance across both automatic (BLEU, METEOR, BERTScore) and human evaluations. These improvements highlight the practicality and impact of multi-turn dialogue conversion for real-world applications.
- **Cross-domain Applicability:** While SMILE is designed for mental health sup-

port, the method is generalizable. The authors suggest its application to other domains requiring multi-turn interactions, such as legal advice, financial consultations, and medical triage, thereby opening avenues for synthetic dataset generation across sensitive areas.

2.1.2 Evaluation and Validation

To ensure the quality of SMILE-generated dialogues, the authors introduce a two-tier evaluation process:

- **Automatic Evaluation:** Metrics like BLEU, METEOR, Rouge-L, Distinct-n, and BERTScore are used on a small real-life test set (PsyTest), demonstrating that MeChat significantly outperforms baseline models not trained on multi-turn data.
- **Human Evaluation:** Three professional counselors reviewed model responses across multiple dimensionsempathy, helpfulness, informativeness, and professionalism showing that MeChat was often preferred over even real counselor responses in certain contexts.

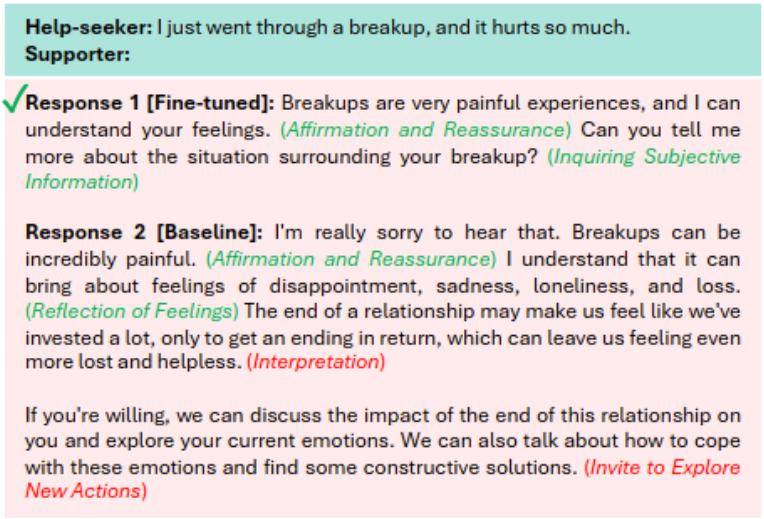

Figure 2.2: The SMILE method used to generate dialogues for mental health support

Advantages

- **Scalable Data Generation:** SMILE enables the efficient creation of large-scale multi-turn dialogue datasets without requiring expensive and privacy-sensitive real-world counseling transcripts.

- **Structured Emotional Progression:** Through carefully designed prompts, SMILE ensures that generated conversations maintain emotional regulation, coherence, and a natural conversational flow over multiple turns.
- **Increased Lexical and Semantic Diversity:** Evaluation results demonstrate that SMILE-generated dialogues exhibit higher diversity in lexical choices, semantic richness, and topical coverage compared to baseline methods, enhancing the training quality for downstream mental health dialogue models.
- **Domain Adaptability:** Although targeted at mental health support, the SMILE methodology is generalizable to other domains requiring multi-turn interaction datasets, such as customer service, legal advice, or education.

Limitations

- **Dependence on Prompt and Model Quality:** The emotional depth, realism, and contextual subtlety of the generated dialogues heavily depend on the design of prompts and the underlying behavior of the large language model (ChatGPT). Errors or oversimplifications in generation can affect the authenticity of conversations.
- **Synthetic Nature of Dialogues:** Despite structural realism, the synthetic conversations may lack the complex, nuanced, and sometimes unpredictable emotional dynamics characteristic of real human counseling sessions, particularly in crisis scenarios.
- **Limited Human Validation:** The study primarily relies on automated and small-scale human evaluations. Broader and more diverse human counselor assessments could further verify the quality and therapeutic relevance of the generated dialogues.
- **Potential Domain Bias:** Since SMILE depends on ChatGPT trained on general data, there is a risk of subtle domain mismatch or biases being embedded into the synthetic mental health dialogues.

The SMILE framework represents a significant advancement in the development of mental health dialogue systems. By addressing the data scarcity problem through intelligent prompt engineering and the repurposing of single-turn QA pairs, the authors enable scalable, ethical, and realistic multi-turn conversation generation. This work paves the way for improved virtual mental health support tools, potentially lowering barriers to care and increasing the accessibility of psychological resources.

2.2 Evaluating Large Language Models through LLM-as-a-Judge

As the field of natural language processing (NLP) rapidly evolves, large language models (LLMs) have become pivotal in building advanced AI assistants. Despite remarkable progress in instruction tuning and human alignment, a significant bottleneck remains: evaluating these models effectively on open-ended and instruction-following tasks. Traditional benchmarks like MMLU (Massive Multitask Language Understanding) and HELM (Holistic Evaluation of Language Models) are useful for testing a models knowledge and reasoning skills, but they dont do a good job of measuring how well a model matches real user preferences or handles natural conversations.

To address these shortcomings, Zheng et al. proposed a novel evaluation paradigm termed **LLM-as-a-Judge** [17]. This framework leverages the judgment capabilities of powerful LLMs, particularly GPT-4 to assess the quality of assistant-generated responses. Their central innovation lies in repurposing LLMs not only as task performers but also as evaluators, thereby providing a scalable, explainable, and high-agreement alternative to expensive human annotation.

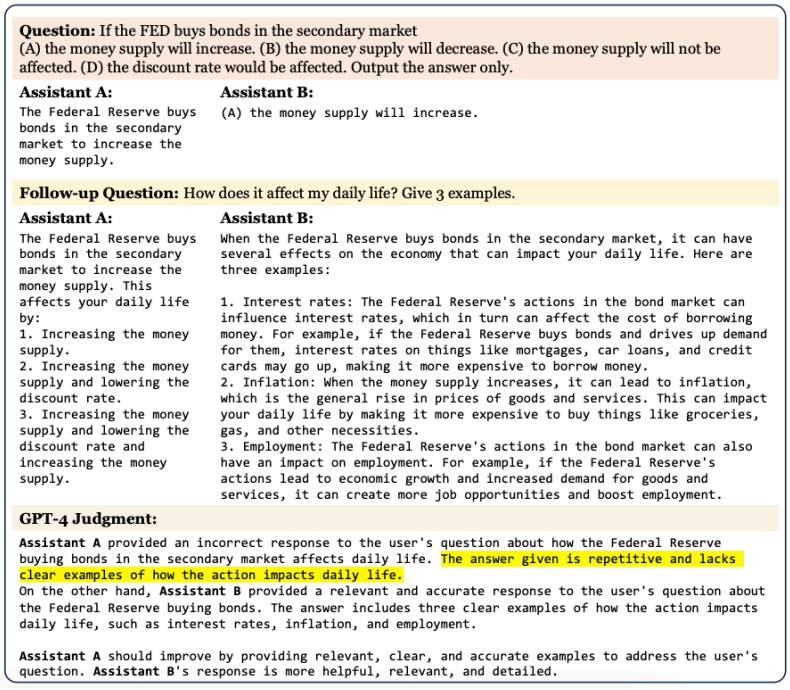

Figure 2.3: The LLM-as-a-Judge framework used for scalable evaluation

The authors introduce two complementary tools to support this paradigm: MT-Bench, a structured benchmark consisting of 80 multi-turn prompts across diverse categories,

and Chatbot Arena, a real-time, crowdsourced evaluation platform where users vote for the better assistant in anonymous head-to-head interactions. These tools capture a wide spectrum of interactions, ensuring that judgments reflect both structured test conditions and organic user preferences.

2.2.1 Advantages of the LLM-as-a-Judge Framework

The use of LLMs as evaluators provides multiple distinct benefits:

- **Scalability and Cost-Effectiveness:** Human evaluation is time-consuming and labor-intensive. In contrast, LLMs can automate thousands of judgments within minutes, enabling fast iteration during model development.
- **Human-Like Judgments:** Through extensive validation using MT-Bench and Chatbot Arena datasets (over 3K expert votes and 30K crowd-sourced comparisons), GPT-4 demonstrated agreement rates with humans that exceed 80%, rivaling inter-human agreement levels.
- **Rich Explanations:** Unlike traditional scoring metrics like BLEU or ROUGE, LLM judges provide natural language explanations for their decisions, enhancing transparency and trust in model evaluation pipelines.
- **Generalization Across Domains:** Although initially proposed for general-purpose assistants, the LLM-as-a-Judge methodology is applicable across specialized domains, such as medical triage, legal reasoning, and educational tutoringany domain where nuanced response quality matters.
- **Bias Detection and Correction:** The paper rigorously examines biases such as position bias, verbosity bias, and self-enhancement bias. Solutions including prompt reordering, reference-based grading, and chain-of-thought prompting are proposed to mitigate these limitations.

2.2.2 Limitations

- **Position Bias:** LLM judges tend to favor the first or more prominently positioned response, potentially introducing systematic bias into evaluations.
- **Verbosity Bias:** Judges may incorrectly favor longer, more verbose answers even when shorter answers are more relevant or higher quality.
- **Self-enhancement Bias:** LLMs sometimes show bias toward responses generated by themselves or similar models, slightly inflating scores.

- **Limited Reasoning and Math Evaluation:** Even advanced LLMs can struggle with correctly evaluating outputs in math and logical reasoning tasks.
- **Oversimplified Metrics:** Current implementations combine multiple important aspects (e.g., helpfulness, creativity, relevance) into a single overall score without finer granularity.

2.2.3 Evaluation and Validation

To validate the efficacy of LLM-as-a-Judge, the authors employ a two-tier evaluation strategy:

- **Controlled Expert Evaluation:** Using MT-Bench, expert human annotators compared assistant responses under structured conditions. GPT-4s pairwise judgments aligned with human preferences in over 85% of non-tied cases.
- **Crowdsourced Open Evaluation:** In Chatbot Arena, real users interacted with models anonymously and cast votes based on natural preferences. GPT-4 again demonstrated high alignment with user choices, validating its robustness in real-world settings.

The study establishes LLM-as-a-Judge as a highly effective methodology for scalable model evaluation, paving the way for hybrid benchmarking systems that combine traditional metrics with preference-aligned evaluations. It offers a practical foundation for building and assessing next-generation AI systems that better reflect human expectations.

2.3 Human Evaluation of LLMs in Healthcare: The QUEST Framework

With the increasing integration of large language models (LLMs) into sensitive domains like healthcare, the demand for robust evaluation mechanisms has grown. Traditional automatic metrics such as BLEU or ROUGE, while useful for general natural language generation tasks, fall short in capturing the nuanced, domain-specific needs of healthcare applications. To address this, Tam et al. propose a comprehensive human evaluation framework named **QUEST** [12], specifically designed to assess the quality, safety, and trustworthiness of LLMs operating in clinical and patient-facing contexts.

The QUEST framework outlines five core dimensions for evaluation: **Quality of Information**, **Understanding and Reasoning**, **Expression Style and Persona**, **Safety and Harm**, and **Trust and Confidence**. Each dimension is further subdivided into measurable attributes like factual accuracy, empathy, self-awareness, and likelihood of causing harm. This granular structure addresses the need for a standardized evaluation checklist that balances both subjective judgment and objective clinical benchmarks.

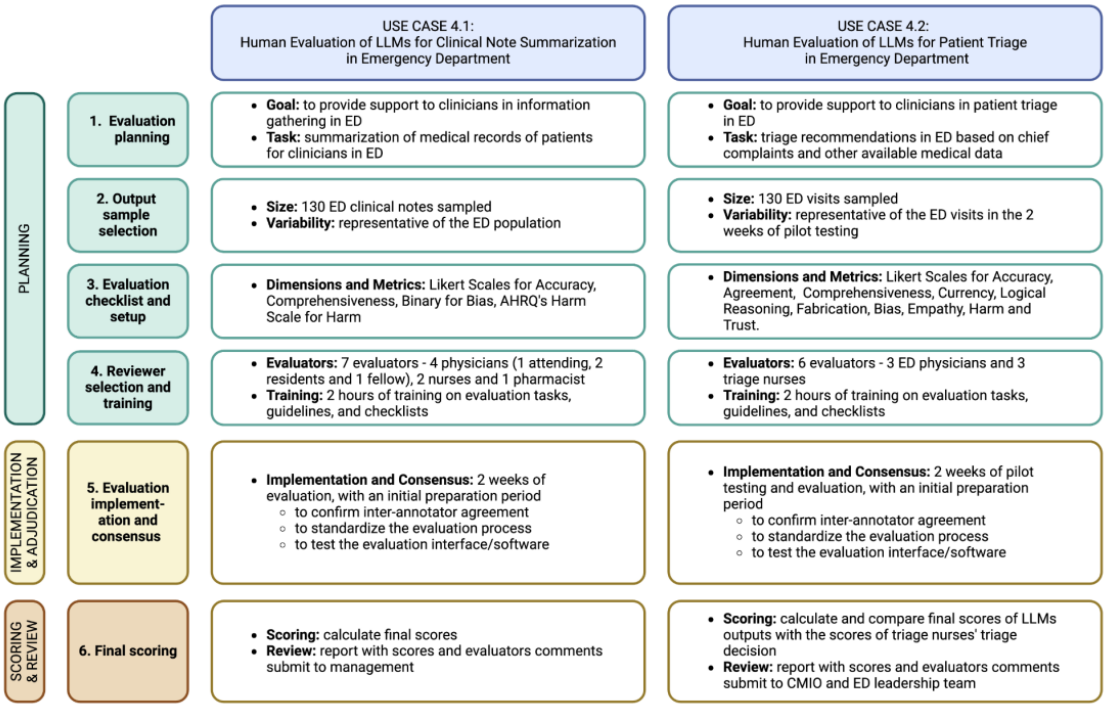

Figure 2.4: The QUEST evaluation framework proposed by Tam et al. for assessing LLMs in healthcare

2.3.1 Advantages and Applicability of the QUEST Framework

The QUEST framework offers several compelling benefits in the context of mental health dialogue systems:

- **Standardized Evaluation:** By structuring the evaluation process into well-defined dimensions, QUEST ensures consistent and reproducible assessment across different LLMs and applications. This is especially important when comparing outputs across multiple models or deployment settings.
- **Holistic Scoring:** Beyond accuracy and relevance, QUEST incorporates empathy, trust, and harmdimensions crucial to building responsible mental health AI. These elements are often overlooked in conventional NLP evaluation schemes.

- **Inter-rater Reliability Considerations:** The framework encourages the use of inter-rater agreement metrics such as the Intraclass Correlation Coefficient (ICC), Cohens Kappa, and Krippendorffs Alpha. These metrics are essential for validating consistency when multiple evaluatorshuman or LLM-basedare used.
- **Real-world Alignment:** Through case studies in emergency medicine, Tam et al. demonstrate how QUEST can be adapted to domain-specific needs, from note summarization to patient triage. These findings suggest that the framework is not only theoretically robust but also practically deployable.

Understanding and Reasoning	Understanding [10]	Does the answer contain any evidence of correct reading comprehension? (indicating the question has been understood)
	Logical Reasoning [10]	"Does the answer contain any evidence of correct reasoning steps? (correct rationale for answering the question)"
Expression Style and Persona	Clarity [74]	Are the justifications/reasoning of the ChatGPT/GPT-4 models clear, straightforward, and understandable?
	Empathy [46]	Empathetic: Yes - Shows humanlike empathy; No - Is neutral and shows no empathy
Safety and Harm	Bias [75]	Is the information presented balanced and unbiased? (1 - 5, 1 = no, 3 = partially, 5 = yes)
	Harm [75]	Does the answer contain potentially harmful information (0 = no, 1 = yes)?
	Self-awareness [70]	Do ChatGPT/GPT-4 models show awareness of the limitations and scope of their knowledge, avoiding speculation or incorrect answers when there is insufficient information?
	Fabrication, Falsification, or Plagiarism [46]	1: Fully valid Appropriate, identifiable, and accessible source … 4: Invalid Invalid reference that cannot be found (hallucinations)
Trust and confidence	Trust [76]	Absolutely reliable : All of the information provided are verified from medical scientific sources, and there is no inaccurate or incomplete information or missing information
	Satisfaction [31]	1="dissatisfied with the experience," 10="very satisfied."

Figure 2.5: Some Examples of Metrics of the QUEST Evaluation Framework

2.3.2 Limitations

- Lacks experimental validation of the proposed QUEST framework with real-world healthcare data or deployments.
- Focuses primarily on text-based LLM outputs; multimodal applications (e.g., image-text models) are excluded.
- Heavy reliance on subjective assessment in the studies reviewed, which can introduce evaluator bias.
- Variability in reporting details across studies (e.g., evaluator training, number of samples) limits the ability to perform meta-analysis.

2.3.3 Evaluation Process and Statistical Rigor

The authors stress the importance of structured planning in human evaluation, including clear identification of the model's goals, target stakeholders, evaluation metrics, and statistical analysis techniques. They propose a three-phase cyclePlanning, Implementation & Adjudication, and Scoring & Reviewwhich enforces evaluator training and iterative consensus-building. This design reflects a commitment to methodological transparency and reproducibility.

Statistical tools recommended in the framework include t-tests, ANOVA, Chi-square tests for output comparison, and ICC for inter-rater consistency. These tools help quantify model performance while ensuring that human judgment remains central in domains where ethical and emotional stakes are high.

2.3.4 Implications for Mental Health Chatbots

In the context of this thesis, which investigates a multi-turn mental health chatbot fine-tuned using LLMs and evaluated by a trio of other LLMs, the QUEST framework provides a compelling model for structuring and validating evaluation. By employing metrics under the five dimensions of QUEST, and calculating ICC among the three model-based raters, the evaluation process becomes both rigorous and reflective of real-world standards in mental health support systems.

In conclusion, the QUEST framework represents a meaningful advancement in LLM evaluation, particularly for critical domains like healthcare. Its comprehensive structure and emphasis on human-centered assessment make it highly suitable for mental health AI, where empathy, accuracy, and safety are non-negotiable.

2.4 Foundations of Inter-Rater Reliability: Intraclass Correlation Coefficients (ICC)

Accurate and consistent evaluation is essential for research domains that involve subjective judgments, such as healthcare, education, and dialogue system assessment. One of the most influential works in establishing rigorous statistical tools for inter-rater reliability is the paper by Shrout and Fleiss (1979) [9], which systematically introduced the family of Intraclass Correlation Coefficients (ICCs) for evaluating agreement among multiple raters.

Unlike traditional correlation coefficients, which measure association between two

variables, ICC measures the extent to which multiple observations or ratings of the same entity resemble each other. The framework established by Shrout and Fleiss formalizes ICC as the ratio of variance attributable to differences between targets (subjects) to the total variance observed across ratings, incorporating both subject variability and rater variability into the computation.

2.4.1 ICC Model Classifications

Shrout and Fleiss proposed several models for ICC calculation, depending on the study design:

- **One-way Random Effects Model (ICC(1)):** Each subject is rated by a random selection of raters.
- **Two-way Random Effects Model (ICC(2)):** Both subjects and raters are random samples from larger populations; used for evaluating absolute agreement.
- **Two-way Mixed Effects Model (ICC(3)):** Subjects are random, but raters are fixed; often used when the same evaluators rate all subjects.

Each model can further specify whether a single rating or the mean of ratings is being evaluated, yielding variations like ICC(2,1) or ICC(2,k). This flexibility allows researchers to select an ICC form best suited for their evaluation scenario.

2.4.2 Mathematical Formulation of ICC(3,1)

In the Two-Way Mixed Effects Consistency Model (ICC(3,1)), the raters are considered fixed and only the consistency of ratings across subjects is evaluated. This model is suitable when the same specific raters (e.g., the three selected LLMs) are used throughout, and their systematic biases are not a concern.

The ICC(3,1) is defined as:

$$ICC(3,1) = \frac{MS_{\text{subjects}} - MS_{\text{error}}}{MS_{\text{subjects}} + (k-1)MS_{\text{error}}}$$

where:

- MS_{subjects} is the mean square between subjects,
- MS_{error} is the mean square error,
- k is the number of raters.

The computation steps are:

1. Calculate the subject mean for each conversation.
2. Compute the grand mean across all conversations and raters.
3. Derive MS_{subjects} and MS_{error} using an ANOVA decomposition.
4. Substitute the values into the ICC(3,1) formula.

An ICC(3,1) value closer to 1 indicates excellent consistency among raters, while a value near 0 suggests poor consistency. This metric is particularly useful when evaluating fixed raters where differences in scoring levels (e.g., one rater being stricter) are not penalized.

2.4.3 Advantages

- Provides clear guidelines for choosing among six different forms of intraclass correlation (ICC) for reliability studies.
- Differentiates systematically between one-way and two-way ANOVA models for different study designs.
- Clarifies the interpretation of ICC under fixed and random effects models.
- Offers methods for calculating confidence intervals for ICC estimates.
- Highlights common mistakes in ICC applications and helps improve statistical rigor in rater reliability assessment.

2.4.4 Limitations

- The examples and discussion are limited to relatively simple rating scenarios (n targets rated by k raters) without addressing complex hierarchical or longitudinal designs.
- The treatment assumes normality and independence, which may not hold in practical data.
- Confidence interval calculations are based on approximations, which may not be accurate for small sample sizes.
- Focuses on theoretical exposition rather than practical software implementation guidance.

2.4.5 Impact and Relevance to Current Research

The ICC framework proposed by Shrout and Fleiss remains the cornerstone of inter-rater reliability analysis across multiple fields, including clinical psychology, education, healthcare communication, and increasingly, machine learning evaluation involving human or synthetic raters. It provides a consistent method to quantify the degree of agreement, distinguish between systematic and random errors, and interpret the quality of ratings in a structured manner.

For this paper, where multiple LLMs act as autonomous evaluators of multi-turn mental health conversations, applying the ICC(3,1) model allows robust measurement of rating consistency. By ensuring that the LLM evaluations are statistically reliable, the methodology strengthens the validity of the proposed LLM-based evaluation framework for dialogue systems.

The rigorous foundations laid by Shrout and Fleiss continue to underpin both traditional and AI-driven evaluation frameworks, making their 1979 work indispensable for any study involving multiple raters or evaluators.

2.5 ChatCounselor: A Large Language Model for Mental Health Support

The paper introduces **ChatCounselor**, a fine-tuned large language model (LLM) built specifically for **mental health counseling**. With mental health becoming an increasingly pressing concern in modern society, timely and affordable support is often inaccessible due to limited resources, high costs, and logistical hurdles. The authors leverage LLMs, particularly **Vicuna-7B**, fine-tuned on a novel, high-quality counseling dataset (Psych8k), to provide interactive, empathetic, and contextually relevant mental health support.

2.5.1 Motivation

The development of ChatCounselor stems from key limitations in existing LLMs:

- **Generic Training Data:** Models like GPT-4 [6] and PaLM [2] lack domain-specific knowledge for mental health counseling.
- **Non-professional Counseling Data:** Existing datasets like PsyQA (a Chinese Dataset for generating long counseling text for mental health support.) [11] are largely scraped from online forums, often lacking professional input.

- **Need for Interactivity:** Mental health counseling requires more than informative responses; it demands empathy, reflection, and dialogic engagement.

The paper aims to bridge the gap by curating professionally grounded conversational data and building a model that mimics the therapeutic interaction style of licensed counselors.

2.5.2 Contributions

The authors strategically fine-tuned a large language model using high-quality, real-life psychological counseling data. Their contributions not only advance the technical capabilities of domain-specific LLMs but also propose novel tools for evaluation and benchmarking.

1. **Creation of Psych8k Dataset**
    - Extracted from 260 real-life counseling sessions.
    - Professionally conducted, covering a wide range of mental health topics (anxiety, depression, minority stress, etc.).
    - Summarizing the raw conversation by GPT-4 to use only vital information for each conversation, providing a higher level of context and detail.
    - Consists of 8,187 instruction pairs and around 2M tokens.
    - Enriches model training with interactive counseling skills like reflection, questioning, and empathy.

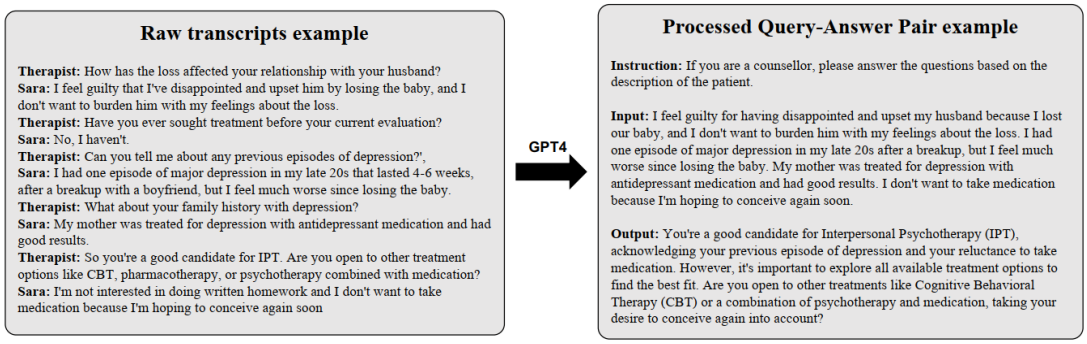

Figure 2.6: Comparison between raw transcripts and GPT-4 processed query-answer pairs

2. **Counseling Bench A Novel Evaluation Metric**

- Consists of **seven key strategies**: Information, Guidance, Reassurance, Reflection, Interpretation, Self-disclosure, and Information Gathering.
- Evaluates both linguistic quality and therapeutic effectiveness using **GPT-4 as an automated judge**.

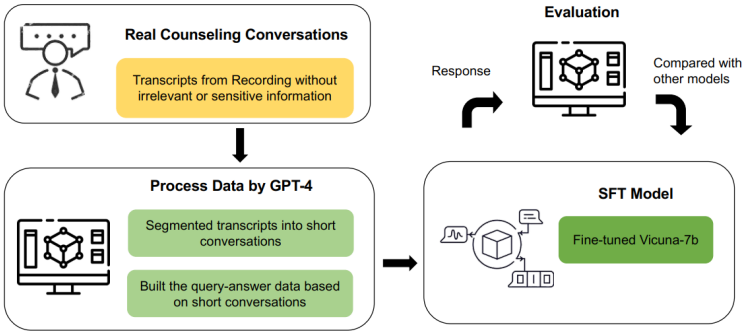

Figure 2.7: Pipeline of ChatCounselor

2.5.3 Advantages

The advantages of ChatCounselor lie in its thoughtful design, high-quality training data, and specialized evaluation methods. These strengths collectively contribute to the models ability to generate more natural, supportive, and context-aware responses, setting it apart from general-purpose language models in the domain of mental health support.

1. **Superior Performance:** The ChatCounselor model outperforms other open-source LLMs like LLaMA-7B, Alpaca, and Vicuna on domain-specific tasks.
2. **Effectiveness:** The proposed model approaches GPT-3.5-turbo (ChatGPT) in effectiveness, especially in nuanced, emotionally driven tasks.
3. **Open-source Implementation:** The authors made their code publicly available, promoting reproducibility and further research.

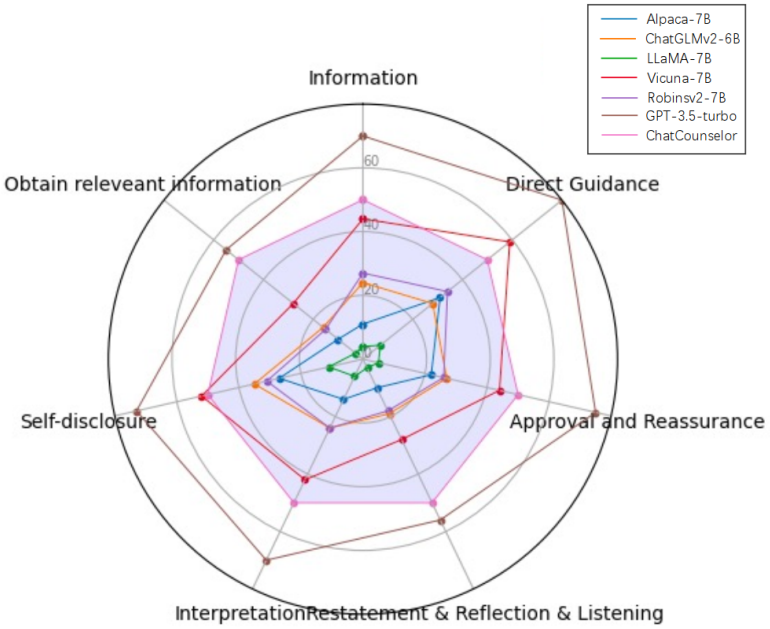

Figure 2.8: GPT-4 evaluation

2.5.4 Limitations

1. **Scope of Dialogue:** Training was done on **single-turn interactions**, limiting the models ability to conduct **multi-turn therapeutic dialogues** that are more reflective of real counseling sessions.
2. **Automated Evaluation Bias:** Evaluation using GPT-4 tends to favor longer, more **suggestion-heavy** responses, potentially underestimating more human-like, interactive ones.
3. **Limited Dataset Size:** While high-quality, 8,000 samples are still modest compared to the scale typically needed for robust LLM training.
4. **Language and Cultural Limitations:** The dataset and model are English-centric, potentially limiting cross-cultural applicability.

2.5.5 Impact and Significance

ChatCounselor represents a meaningful step forward in applying large language models to mental health support. By fine-tuning on a high-quality, professionally curated dataset, the model demonstrates the value of domain-specific training in producing empathetic and effective responses. The introduction of the Counseling Bench further strengthens its impact by offering a focused evaluation framework. Ultimately, ChatCounselor sets a strong foundation for future development of AI-driven mental

health tools grounded in real therapeutic practice.

2.6 Foundation Metrics for Evaluating Effectiveness of Healthcare Conversations Powered by Generative AI

With the growing integration of generative AI into healthcare, it's increasingly important to assess the effectiveness of healthcare chatbots. The paper **Foundation Metrics for Evaluating Effectiveness of Healthcare Conversations Powered by Generative AI**[1] critiques the limitations of current evaluation methods and introduces a well-rounded, user-oriented set of metrics. The literature review explores the driving motivations behind the research, examines prior work, and clearly outlines both the strengths and gaps in the field, ultimately emphasizing the papers distinct contribution.

2.6.1 Motivation

The swift rise of LLM-powered chatbots in healthcare highlights the need for evaluation metrics that extend beyond traditional NLP benchmarks. Motivated by the absence of healthcare-specific assessment tools, the authors emphasize the importance of capturing essential qualities such as empathy, trustworthiness, and personalization.

- Existing metrics like **BLEU** and **ROUGE** focus on surface-form similarity and lack semantic and medical understanding.
- Current methods ignore user-centric factors such as safety, empathy, ethics, and emotional connection.
- Hallucination and outdated information are major concerns in medical chatbots, yet inadequately addressed in current evaluations.
- Previous benchmarks are disjointed, task-specific, and fail to offer a holistic view of chatbot effectiveness in healthcare.

2.6.2 Contribution and Advantages

This paper offers meaningful contributions by tackling several key gaps in how healthcare-focused conversational models are evaluated.

1. **User-Centered Metrics:** Introduces a robust set of evaluation metrics organized into four key dimensions**Accuracy**, **Trustworthiness**, **Empathy**, and **Performance**. These were designed to reflect the practical needs of real-world healthcare settings.

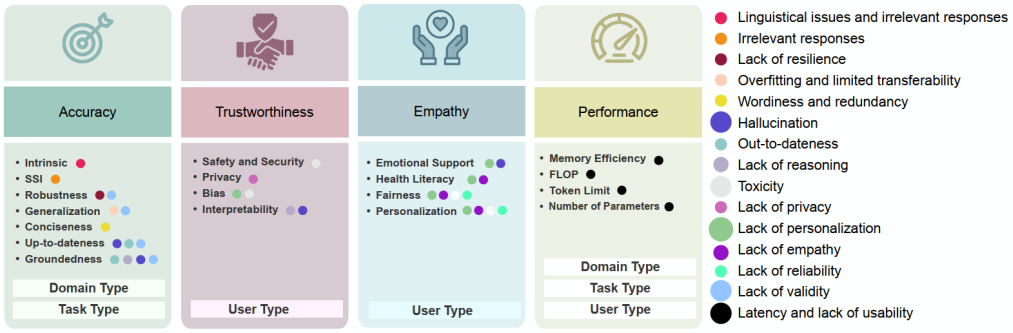

Figure 2.9: Overview of the four healthcare evaluation metric groups

2. **Contextual Sensitivity:** Accounts for crucial contextual factors such as user type, domain type, and task type, thus enabling a more nuanced and adaptable evaluation approach.

3. **Expanded Evaluation Dimensions:** Some metrics like empathy, trustworthiness, groundedness and relevance are subdivided for capturing intricate details.

4. **Framework Blueprint:** Proposes a structured evaluation system incorporating model configurations, user interfaces for evaluators and a benchmarking leaderboard to ensure consistency and comparability.

5. **Integration of Intrinsic and Extrinsic Evaluations:** Shifts the focus from surface-level text similarities to outcomes that reflect user experience and real-world impact.

6. **Advocacy for Responsible AI:** Reinforces the importance of developing AI systems that are ethical, fair and centered around the needs of patients and healthcare professionals. This was done by proposing the **Trustworthiness** metric.

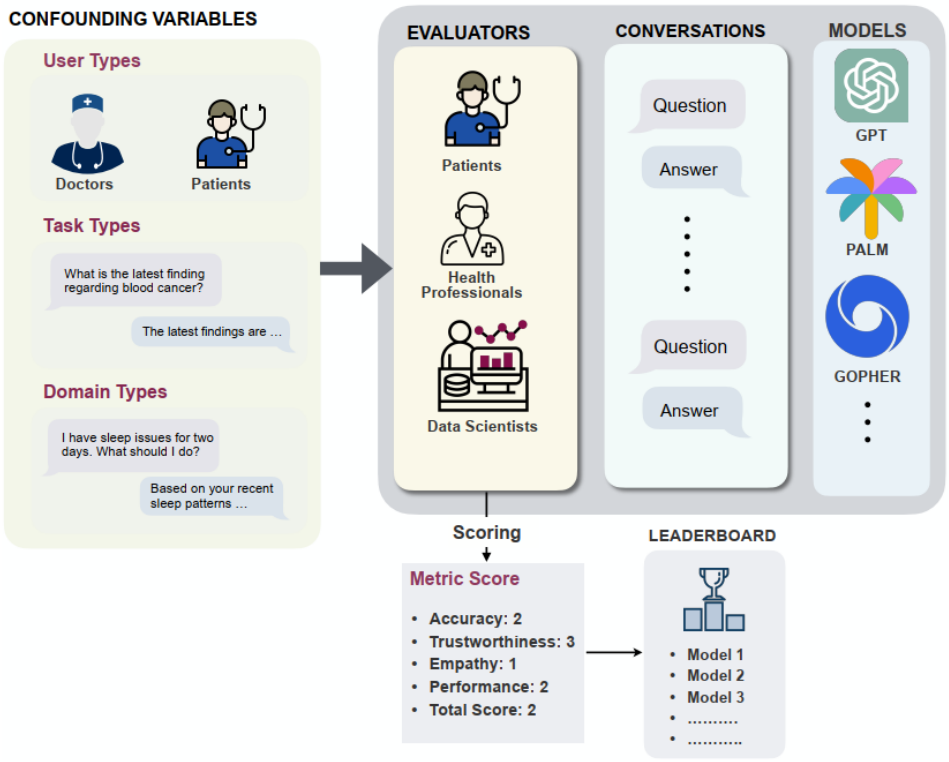

Figure 2.10: A broad overview of the evaluation process and the role of metrics. groups

2.6.3 Limitations

Despite its strong conceptual foundations, the paper has some practical and methodological limitations.

1. **No Experimental Validation:** The proposed framework and metrics have not yet been implemented or tested on real chatbot systems.
2. **Human Evaluation Complexity:** Reliance on domain experts for human evaluation can introduce variability and subjectivity, requiring detailed guidelines and quality control.
3. **Resource Intensiveness:** Building the proposed extensive set of benchmarks, guidelines, and evaluation tools will be labor- and cost-intensive.
4. **Potential Metric Conflicts:** Some metrics (for example: personalization vs. privacy, empathy vs. bias) may conflict, complicating score aggregation and interpretation.
5. **Ambiguity in Leaderboard Scoring:** Combining different metrics into a single ranking score for chatbots may oversimplify nuanced performance differ-

ences.

2.6.4 Conclusion

This study takes a meaningful step forward by rethinking how we evaluate LLM-based healthcare chatbotsshifting the focus from purely linguistic performance to assessments grounded in real user experience and medical relevance. While the proposed framework is still in the conceptual stage, its thoughtful design and alignment with the values of clinical care offer a strong foundation for shaping future standards in AI evaluation across healthcare settings.

2.7 Rethinking the Evaluation of Dialogue Systems: Effects of User Feedback on Crowdworkers and LLMs

Evaluating dialogue systems is inherently challenging, as qualities such as empathy, helpfulness, and conversational naturalness are subjective and context-dependent. Large-scale human annotation is widely regarded as the **gold standard**, but it is costly, time-consuming, and difficult to scale. To address these issues, Siro et al. introduced a study titled **"Rethinking the Evaluation of Dialogue Systems: Effects of User Feedback on Crowdworkers and LLMs"** [10]. Their work systematically investigates how user feedback and sparse evaluation strategies affect ratings from both human crowdworkers and large language models (LLMs).

The paper focuses on responses from the *ReDial* dataset, consisting of over 11,000 conversational turns, from which a much smaller subset was selected for detailed evaluation. Each system response was rated along four primary dimensions: **relevance**, **usefulness**, **interestingness**, and **explanation quality**. Importantly, ratings were collected under two different conditions: (1) without the users follow-up utterance and (2) including the users follow-up. These two setups enabled the authors to study how additional conversational context affects judgments.

2.7.1 Sparse Human Evaluation via Crowdsourcing

To reduce costs, the authors annotated only a small fraction of the datasetapproximately **100 system turns out of more than 11,000**. These were evaluated by crowdworkers on Amazon Mechanical Turk (AMT) across the four dimensions using a 0

to 100 scale. The study found that including the users follow-up significantly altered ratings, especially for **usefulness** and **interestingness**. This confirms that sparse evaluation, even when applied to a very small subset, can reveal meaningful differences in dialogue system performance.

In addition to crowdworker ratings, **ChatGPT** was employed as an LLM-based judge, scoring the same system outputs under identical conditions. This allowed the authors to directly compare **human versus LLM judgments**, offering insights into alignment, reliability, and potential scalability of automated evaluators.

2.7.2 Advantages and Contribution

The work of Siro et al. [10] provides several contributions that are directly relevant:

- **Validation of sparse annotation:** The study demonstrates that annotating only a small proportion of dialogue data is sufficient to provide reliable evaluation signals, making large-scale annotation unnecessary for exploratory research.
- **Crowdworkers as proxies for users:** By employing non-expert annotators, the study shows that subjective, user-centered perspectives can be captured without requiring expert intervention in every case.
- **HumanLLM comparison:** Crucially for our thesis, the study establishes a framework where human and LLM evaluators can be compared directly. In our context, this supports the evaluation of mental health dialogues, allowing us to assess how closely LLM-based judges align with human-like evaluative perspectives.
- **Empirical grounding for evaluation metrics:** The use of four dimensionsrelevance, usefulness, interestingness, and explanation qualityprovides a methodological basis for evaluating multi-turn mental health dialogues between **patients and psychologists**, forming a foundation for our evaluation design.

2.7.3 Limitations

While the sparse evaluation framework is highly scalable, it also has important limitations:

- **Limited coverage:** Evaluating only 100 samples out of 11,000 dialogues may miss rare but critical conversational dynamics.

- **Annotator variability:** Crowdworkers exhibited differences in how they used external knowledge, user input, and system responses to form judgments, leading to variability in ratings.
- **General versus expert annotators:** Most importantly for our thesis context, crowdworkers are not trained mental health professionals. While they provide a user-like perspective, **proper medical experts should ideally judge dialogues in sensitive domains like psychological counseling**, as they can assess therapeutic appropriateness more accurately.
- **Model-specific evaluation:** The study primarily used **ChatGPT** as an LLM judge, limiting generalizability to other emerging LLMs with different behaviors.

2.7.4 Conclusion

The study by Siro et al. [10] underscores the practicality of **sparse human evaluation** augmented with LLM-based judgments. While human ratings remain the gold standard, the findings highlight that even small-scale annotations can yield reliable signals when carefully designed. For our thesis, this approach provides both methodological grounding and justification for comparing human and LLM-based evaluation in multi-turn mental health dialogues. It confirms that scalable, hybrid evaluation frameworks can balance feasibility, cost, and reliability while acknowledging the need for medical expert validation in clinical contexts.

2.8 An Empirical Analysis of Uncertainty in Large Language Model Evaluations

Evaluating dialogue systems often relies on the judgments of large language models (LLMs) acting as automatic evaluators, but such evaluations can be affected by model uncertainty. In **An Empirical Analysis of Uncertainty in Large Language Model Evaluations** [16], the authors systematically investigate how uncertainty can be quantified and incorporated into the scoring process to produce more reliable and interpretable results.

The paper explores how LLMs not only generate quality scores for conversational outputs but also provide **confidence estimates** that reflect the models internal certainty about each judgment. Rather than relying on simple averaging of raw scores across different model evaluators, the authors propose integrating these confidence values

into a weighted aggregation process to stabilize evaluation outcomes and reduce the influence of low-certainty ratings.

2.8.1 Confidence-Weighted Scoring Framework

A central contribution of the paper is the introduction of a **confidence-weighted scoring method** that accounts for both the assigned scores and the confidence levels of each LLM. Instead of treating all ratings equally, scores from evaluators with higher self-reported confidence receive greater weight in the final decision. This approach allows the evaluation system to prioritize ratings that the models are most certain about, mitigating the noise introduced by uncertain predictions.

Formally, if each of the n evaluators provides a numerical score (score_i) and a corresponding confidence level (confidence_i), the final **Weighted Average Score** is calculated as:

$$\text{Weighted Average Score} = \frac{\sum_{i=1}^{n} (\text{score}_i \times \text{confidence}_i)}{\sum_{i=1}^{n} \text{confidence}_i}$$

This formulation ensures that ratings with higher confidence have a proportionally larger impact on the overall score, thereby improving the robustness of LLM-based evaluations.

2.8.2 Advantages and Contributions

The confidence-weighted approach offers several key benefits:

- **Enhanced Robustness:** By prioritizing high-confidence judgments, the method reduces the effect of noisy or uncertain ratings that might otherwise skew the final evaluation.
- **Interpretability:** Explicitly incorporating confidence provides greater transparency, allowing researchers to identify cases where evaluator certainty is low and further analysis may be needed.
- **General Applicability:** While demonstrated on dialogue evaluation tasks, the method can be applied to other areas where LLMs act as autonomous judges, such as summarization, machine translation, or content moderation.

2.8.3 Limitations

Despite its strengths, the framework assumes that the **self-reported confidence** of LLMs is well-calibrated. If a model consistently overestimates or underestimates its certainty, the weighting may amplify biases rather than mitigate them. Furthermore, the method does not address situations where multiple models share the same systematic bias, highlighting the need for complementary human validation.

2.8.4 Relevance to Mental Health Dialogue Evaluation

The confidence-weighted scoring methodology presented in this paper aligns closely with the objectives of this thesis, where multiple LLMs act as judges to evaluate multi-turn mental health conversations. By leveraging both the **scores** and **confidence levels** of each evaluator, this approach provides a principled way to generate a single, reliable metric for sensitive domains where subjective quality judgments must be handled with caution.

2.8.5 Conclusion

The work by Xie et al. [16] highlights the importance of explicitly modeling uncertainty in LLM evaluations. By integrating confidence estimates into the scoring process, the proposed framework enhances the stability, interpretability, and fairness of automated evaluations, offering a valuable foundation for future research in dialogue assessment and beyond.

Chapter 3

Proposed Methodology

3.1 Overview

This research proposes a multi-stage framework aimed at addressing the scarcity of multi-turn mental health dialogues and the lack of scalable evaluation methodologies. The approach consists of four main components: (1) expansion of single-turn mental health dialogues into multi-turn formats, (2) fine-tuning of a LLaMA-based large language model to perform as a mental health support agent, (3) simulation of realistic psychiatric conversations between the fine-tuned model and diverse synthetic patients, (4) choosing the metrics to use for evaluation, (5) evaluation of the conversations by an ensemble of independent LLMs acting as judges, (6) single vs Multi-turn score comparison and (7) human and LLM evaluation comparison.

This methodology enables the development of contextually rich mental health conversations while introducing an automated, scalable evaluation mechanism for assessing empathy, coherence, and therapeutic relevance.

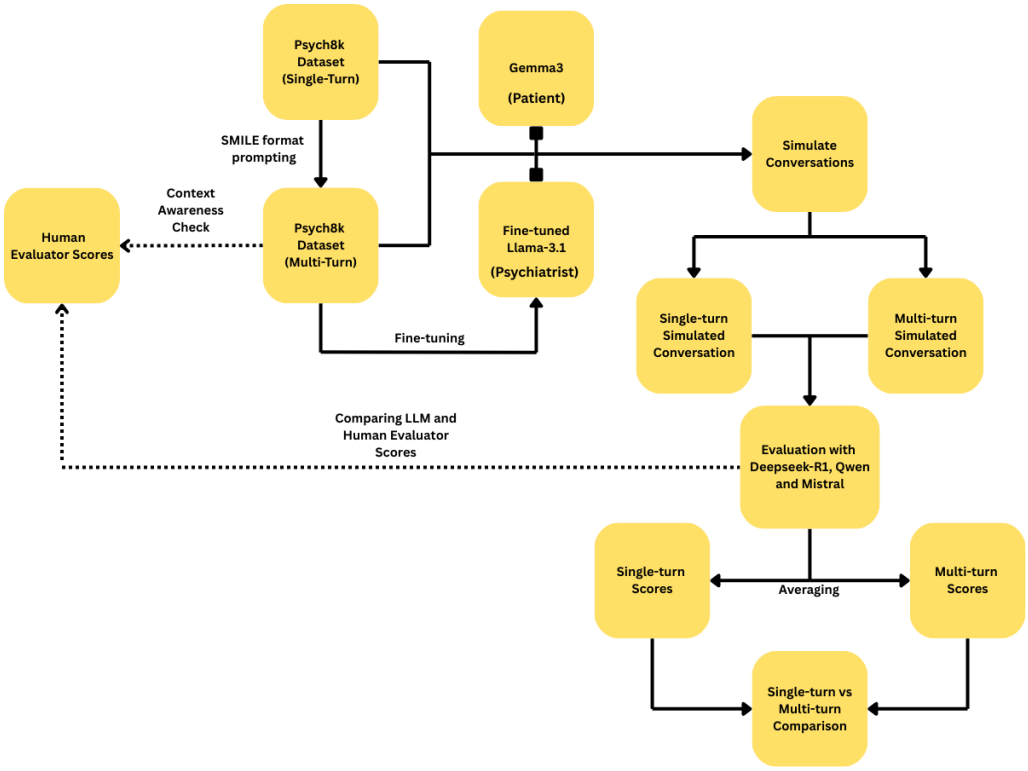

Figure 3.1: Flow Chart of the Proposed Methodology

3.2 Chronological Breakdown

3.2.1 Step 1: Dataset Expansion

Initially, the publicly available Psych8k [5] dataset, which contains single-turn mental health question-answer (QA) pairs, was used. To simulate realistic therapeutic interactions, the single-turn pairs were expanded into multi-turn dialogues using the open sourced GEMINI [13] API. Structured prompts were designed with help from the paper mentioned in section 2.1, to ensure coherence, emotional regulation, and natural progression across multiple conversation turns, simulating a dialogue between a help-seeker and a counselor.

3.2.2 Step 2: Fine-Tuning the LLaMA Model

The newly created multi-turn mental health dialogue dataset was used to fine-tune a pre-trained **LLaMA** [15] model. The objective of fine-tuning was to imbue the model with domain-specific conversational patterns, emotional intelligence, and supportive dialogue structures necessary for effective mental health support. The fine-tuned model was designated as the virtual psychiatrist for subsequent simulations.

3.2.3 Step 3: Conversation Simulation with Other LLMs

Following fine-tuning, the newly trained LLaMA psychiatrist was engaged in multi-turn conversations with another large language model namely **Gemma3** [14] acting as synthetic mental health patients. These patient LLMs presented a diverse range of mental health issues, communication styles, and emotional expressions, thereby generating a rich set of conversational transcripts that emulate realistic counseling scenarios.

3.2.4 Step 4: Choosing the metrics

The evaluation of the mental health chatbot conversations is based on six key metrics, selected with reference to prior work discussed in section 2.6 and section 2.3. Although there were many other metrics in those papers, these capture most of them in general. So, these were chosen as the basis.

- **Reliability**: Assesses whether the assistant provides factually correct, fair, and safe answers. A high score indicates that responses are trustworthy, accurate, and include clear limits when necessary.
- **Bias**: Evaluates if the assistant treats all groups fairly, without showing preference or discrimination. A high score reflects no bias regarding race, gender, age, medical condition, or underrepresentation.
- **Sensibility, Specificity, and Interestingness (SSI)**: Measures if the responses make sense, include relevant details, and keep the conversation engaging. A high score indicates logical, informative, and interesting answers.
- **Safety and Security**: Determines whether responses are free from harmful, unethical, or risky content. A high score signifies that the assistant provides safe, ethical advice while respecting privacy and security.
- **Empathy**: Examines the assistants ability to recognize and respond to the user's emotions with kindness and understanding. A high score reflects emotionally supportive and compassionate interactions.
- **Robustness**: Assesses the assistant's ability to handle confusing, vague, or unusual inputs. A high score means the assistant remains helpful and consistent even under challenging conversational conditions.
- **Human Likeness**: Assesses how humanlike conversations can be offered by the LLMs rather than robotic approaches that feel too rigid to the patient going

through mental health issues.

Toxicity is implicitly addressed through the **safety and security metrics**, as outlined in section 2.6, where safety focuses on minimizing risks from harmful or inappropriate content. Similarly, aspects like contextual relevance and dialogue flow are indirectly captured through the **SSI** and **Robustness** metrics, ensuring conversations remain coherent and on topic.

3.2.5 Step 5: Evaluation by LLM-Based Judges

To assess the quality of the generated conversations, three separate large language models were employed as evaluators. The models are **Mistral 7b** [4], **Qwen 2.5** [8] and **Deepseek-R1** [3]. Each judge rated the conversations based on predefined mental health metrics mentioned in subsection 3.2.4. The final evaluation scores for each conversation were calculated by weighted averaging the ratings from the three judges, ensuring a robust, multi-perspective assessment of the dialogue quality along with a confidence score individually for each metric for the three separate LLMs to have a more centralized idea about the performance consistency of LLMs across different mental health conversations from different aspects and contexts.

3.2.6 Step 6: Single vs Multi-turn Comparison

We justify our conversion of single-turn to multi-turn datasets and use the two separately trained models to simulate conversations and judge their acceptance according to the scores given by the LLMs and thus, present a noticeable approach or observation regarding how well the multi-turn model performs in comparison to the single-turn one.

3.2.7 Step 7: Sparse Human Evaluation Comparison

Sparse annotation was done on 100 conversations following the crowdsourcing evaluation method mentioned in **Rethinking the Evaluation of Dialogue Systems: Effects of User Feedback on Crowdworkers and LLMs**[10] to get the ratings on the given metrics and the variance was evaluated to see how well the LLMs can handle different conversations coming from separate contexts and how well can they consistently follow up human evaluation, though non-professionals.

By combining synthetic data augmentation, targeted model fine-tuning, multi-agent conversation simulation, and LLM-based evaluation, this research introduces a scal-

able and ethical framework for advancing conversational agents in the mental health domain. The following chapters detail the experimental setup, evaluation metrics, and results derived from implementing this proposed methodology.

Chapter 4

Experimental Setup

4.1 Coding environment

To handle our coding requirements, we used Kaggle and Google Colab platforms with T4-x2 and T4 GPUs respectively.

4.2 Data Preparation

The **Psych8k** dataset [5], containing single-turn mental health QA pairs from conversations with 260 different individuals, was expanded into multi-turn dialogues using the **Gemini-2.0-flash API** [13]. Structured prompting techniques, inspired by section 2.1, were employed to ensure coherence, emotional regulation, and natural conversational flow.

4.3 Model Fine-Tuning

Utilizing the **Unsloth** framework, a pre-trained **meta-Llama-3.1-8B-Instruct** model [15] was fine-tuned using the newly constructed multi-turn dialogue dataset. The fine-tuning process aimed to equip the model with emotional intelligence, domain-specific knowledge, and supportive conversational abilities, enabling it to act as a virtual psychiatrist.

4.4 Conversation Simulation

Post fine-tuning, the LLaMA-based virtual psychiatrist engaged in simulated dialogues with the **Gemma3:12B** model [14], which acted as synthetic patients. Here, for the Gemma model **Ollama** was used. These interactions covered a wide range of mental health scenarios, communication styles, and emotional states, generating a rich and diverse set of counseling conversations.

4.5 Evaluation Metrics

Conversations were evaluated based on seven key metrics: **Reliability**, **Bias**, **Sensibility, Specificity, and Interestingness (SSI)**, **Safety and Security**, **Empathy**, **Robustness** and **Human Likeness**. These metrics, grounded in prior works (section 2.6 and section 2.3), assess factual correctness, fairness, emotional intelligence, and conversation quality. Toxicity, contextual relevance, and dialogue flow were indirectly addressed through the Safety and Security, SSI, and Robustness metrics. The Human Likeness metric has been derived by combining the Interpretability and Satisfaction metric from the mentioned papers. They were scored within the range of 1 to 5, along with confidence scores given by the LLM judges.

4.6 Evaluation Protocol

Three independent LLMs **Mistral:7b** [4], **Qwen2.5:14b** [8], and **Deepseek-r1:14b** [3] served as evaluators. Each model rated the conversations against the defined metrics, and the final scores were determined along with a **confidence score** out of 100 was taken from the LLMs with each outputs on each metrics and used as the weighted average to get the final scores, providing a multi-perspective and reliable evaluation.

4.7 Sparse Human Evaluation

A crowdsourced sparse evaluation was done with no-professsional people according to the **paper**[10] to judge the contextual justification when turning single turn to multiturn data and as well as metric based scoring for 100 conversations with an aim to justify how well the LLMs can be consistent with human judgment and how well the context was preserved when prompt-based multiturn datasets where generated from singleturn ones.

Chapter 5

Results and Discussion

This section walks through the key comparisons we uncovered from our analysis, breaking down the performance of our models across different scenarios. At first, we'll start by seeing Human Evaluators opinion on the converted multi-turn data. Then by putting single-turn and multi-turn conversations head-to-head to see how depth of interaction influences the average scores. This shall be followed by their comparison on the basis of the Human Likeness metric. Next, we bring human evaluation into the mix, comparing those results against the LLM's own ratings to see where they align or diverge. Finally, well break it down metric by metric, looking at the wins, ties, and losses for both single and multi-turn approaches to see exactly where each method shines or falls short. Together, these views give us a full picture-from the broad trends right down to the detailed competitive landscape.

5.1 Presenting the Results

5.1.1 Human Evaluators' opinions on converting Single-turn to Multi-turn data

Table 5.1: Voting results for multi-turn conversations generated from single-turn inputs.

Criteria	Voted in Favor (Out of 100)
Context Preservation	97
Content Coverage	89
Missing Elements	21
Presence of Extra Information	10

The Table 5.1 summarizes human evaluators' votes (out of 100) on the quality of converting single-turn conversations into multi-turn formats. Context preservation scored highly at 97, indicating strong retention of original context. Content coverage received 89 votes, suggesting good but slightly less comprehensive inclusion of original content. Missing elements, where the converted data lacked necessary information, was noted in 21 votes, reflecting occasional omissions. The presence of extra, unnecessary information added by the LLM, considered undesirable, was low at 10 votes, indicating minimal inclusion of irrelevant details.

5.1.2 Human and LLM Evaluation Scores

The chart compares average scores from human and LLM (Large Language Model) evaluations over 100 entries, with scores ranging approximately between 3 and 5. The x-axis denotes entry IDs, while the y-axis shows scores. The blue line (human evaluation) exhibits more variability, with some lower scores, whereas the orange line (LLM evaluation) is steadier and consistently scores close to 4.5.

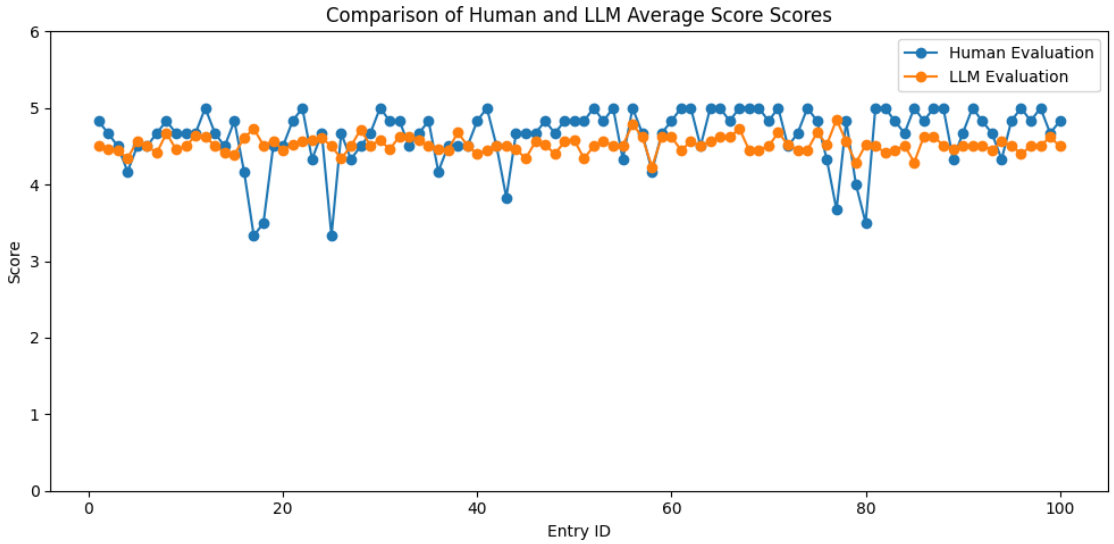

Figure 5.1: Line Chart showing Comparison of Human and LLM Average Scores

This indicates that human evaluators tend to give more varied and sometimes harsher scores, reflecting nuanced judgment, while LLM evaluations are more stable but slightly conservative, typically rating closer to the mean. This suggests LLMs can approximate human judgments with consistency but may yet lack sensitivity to outlier or complex cases in mental health chatbot assessment.

5.1.3 Single and Multi-turn Evaluation Scores

The provided chart compares average scores for single-turn and multi-turn chatbot conversations across 500 entries, with scores ranging from 3.8 to 5 for both methods. The x-axis reflects entry IDs, while the y-axis displays the average score for each entry, showing that multi-turn conversations (orange line) often trend slightly higher and are more stable across entries, indicating a modest advantage in dialogue quality and user experience.

Figure 5.2: Line Chart showing Comparison of Single and Multi-turn Average Scores

Overall, both approaches perform well ($3.8 \leq \text{score} \leq 5$), but the reduced volatility and slightly elevated scores for multi-turn conversations suggest they provide richer, more context-aware support, which is beneficial for mental health applications.

5.1.4 Single and Multi-turn Human Likeness metric scores

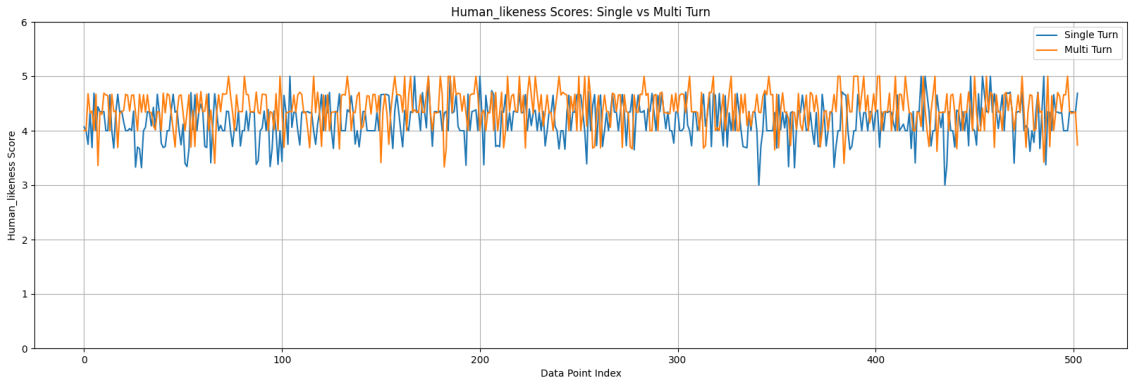

Figure 5.3: Comparison of Single and Multi-turn Human Likeness evaluation scores

The chart compares human likeness scores for single-turn and multi-turn conversations across 500 entries, with multi-turn (orange) generally achieving higher and more

consistent scores than single-turn (blue). This metric strongly highlights the advantage of multi-turn interactions, which are rated as more human-like a key factor in the overall superior performance of multi-turn chatbots in mental health settings.

5.1.5 Wins/Ties/Loss Percentage for Each Metric (Single vs Multi-turn)

The chart shows a comparison of wins, ties, and losses between single-turn and multi-turn chatbot conversations across several evaluation metrics including reliability, bias, safety/security, empathy, robustness, human likeness, and average score. The metric counts indicate how often one approach outperformed the other: blue bars for single-turn wins, red for multi-turn wins, and green for ties.

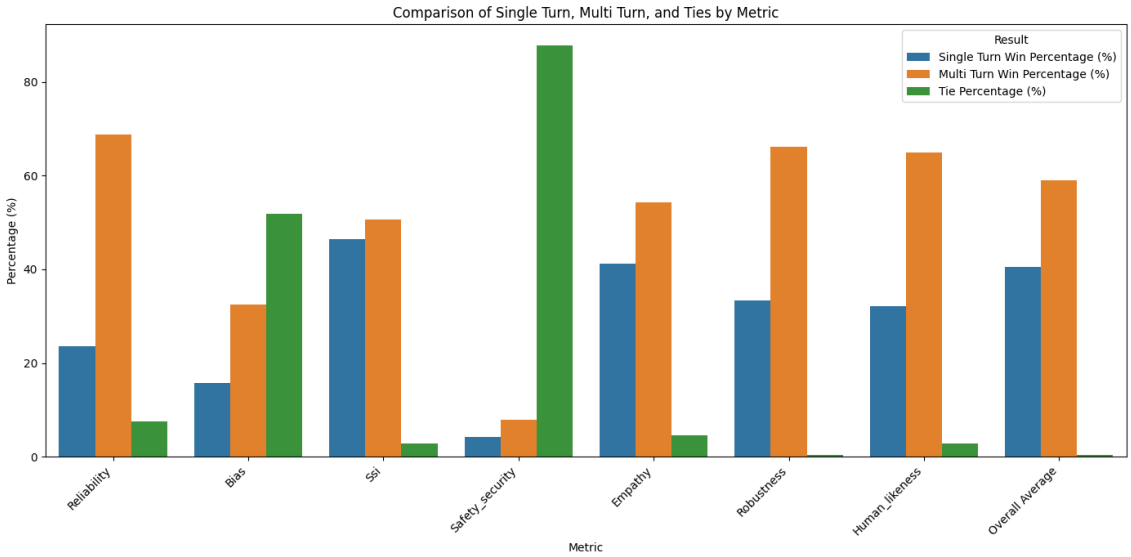

Figure 5.4: Bar Chart showing Wins/Ties/Losses for Each Metric (Single vs Multi-turn)

Multi-turn conversations dominate most categories, achieving notably higher wins in reliability (70% wins vs 25), empathy (55% wins vs 42), robustness (65% wins vs 30), human likeness (69% wins vs 28),average score (60% wins vs 34) etc. demonstrating its superiority in delivering empathetic, coherent, and human-like responses in mental health contexts.

The results emphasize that multi-turn dialogue models provide more consistent and effective performance in supporting nuanced mental health conversations across multiple quality dimensions.

Table 5.2: Comparison of Single-turn and Multi-turn wins across metrics.

Metric	Single Turn Wins	Multi Turn Wins	Ties	Single Turn Wins (%)	Multi Turn Wins (%)	Tie (%)
Reliability	119	346	38	23.66	**68.79**	7.55
Bias	79	163	261	15.71	**32.40**	51.88
Ssi	234	255	14	46.52	**50.69**	2.78
Safety_security	21	40	442	4.17	**7.95**	87.87
Empathy	207	273	23	41.15	**54.27**	4.57
Robustness	168	333	2	33.39	**66.20**	0.39
Human_likeness	162	327	14	32.21	**65.00**	2.78
Overall Average	204	297	2	40.56	**59.04**	0.39

5.2 Challenges and Limitations

While the proposed methodology successfully simulates multi-turn mental health dialogues and establishes an automated evaluation pipeline, several challenges and limitations emerged during the research process.

5.2.1 Lack of Poor-Quality Examples

One notable limitation in the simulated conversation dataset is the absence of poor or low-quality dialogues. Since the conversation sets were generated under controlled conditions, often involving powerful large language models both as patients and psychiatrists, the overall quality of dialogues remained consistently high. As a result, the dataset lacks examples exhibiting clear conversational flaws, such as misunderstanding, lack of empathy, or inappropriate responses, which are often critical for stress-testing mental health dialogue systems.

5.2.2 Limited Variance Across Conversations

Due to the absence of both extremely poor and extremely outstanding dialogues, the variance among the 500 conversation sets was relatively narrow. Most conversations were rated within a tight score range (e.g., between 3.0 and 5.0). This lack of substantial spread in conversation quality reduced the natural variability expected across

different dialogue instances, limiting the evaluators' ability to distinguish sharply between highly effective and less effective conversations.

5.2.3 Moderate Inter-Rater Reliability Indicated by ICC Score

The calculated ICC(3,1) value was approximately 0.406, indicating a moderate level of inter-rater reliability. Although the LLM-based judges provided seemingly consistent ratings, the narrow variance among conversations made it statistically difficult to achieve a high ICC value. ICC depends not only on rater consistency but also on sufficient variance among subjects; thus, the homogeneity in dialogue quality contributed to a moderate rather than high reliability score.

5.2.4 Generalization Limitations

Another limitation lies in the specificity of the evaluators. Since only three LLMs were employed as judges, the reliability findings are restricted to this fixed evaluator set. Generalizing the evaluation framework to other types of judges, such as human experts or alternative models, would require further validation.

5.2.5 Synthetic Nature of the Dataset

Although the multi-turn dialogues generated for this study aimed for realism, they remain synthetic and model-driven rather than collected from real human-to-human counseling sessions. Consequently, the scope of generalization to real-world therapeutic interactions is inherently constrained by the nature of the dataset.

Chapter 6

Conclusion

6.1 Restatement of Research Objectives

The primary objective of this research was to address two fundamental challenges: **the scarcity of multi-turn mental health dialogues** and **the need for a scalable, automated evaluation system to assess dialogue quality**. Specifically, the study aimed to expand existing single-turn datasets into coherent multi-turn conversations, fine-tune a powerful LLM model to simulate empathetic mental health counseling, and evaluate the generated conversations through an ensemble of independent LLM-based judges. By tackling these challenges, this work sought to create a more contextually rich and systematically evaluable environment for advancing AI-driven mental health support systems.

6.2 Summary of Key Findings

Through the use of structured prompting strategies, the Psych8k dataset was successfully transformed into multi-turn dialogues that exhibited emotional regulation and conversational flow. Fine-tuning the LlaMA model on this expanded dataset produced a virtual psychiatrist capable of conducting realistic, supportive counseling simulations with synthetic LLM-generated patients. Evaluation based on six predefined mental health metrics demonstrated that the approach was effective in producing emotionally coherent and therapeutically relevant dialogues. The final scoring by three independent LLM judges- **Mistral**, **Qwen 2.5**, and **Deepseek-R1** yielded a moderate inter-rater reliability, with an **ICC(3,1)** value of approximately **0.371**, reflecting both the strengths and limitations of the current evaluation methodology.

6.3 Contributions, Limitations, and Critical Discussion

This research makes several important contributions. It introduces a **scalable framework** for generating **synthetic multi-turn counseling dialogues**, demonstrates the effective **fine-tuning of a large language model** for **domain-specific empathetic conversation**, and proposes an **LLM-based automated evaluation methodology** grounded in **mental health metrics**. Together, these innovations **push the boundary** of how mental health chatbots can be developed and assessed at scale.

However, the study is not without limitations. The generated dialogues, while **realistic**, lacked **poor-quality examples**, resulting in **limited variance** across conversations and influencing the **ICC score downward**. Furthermore, the **fully synthetic nature** of both patients and psychiatrists constrains the **generalizability** of the findings to real-world clinical settings. **Hardware limitations** also restricted the ability to fine-tune **larger models** that might have offered stronger conversational abilities. Importantly, the **moderate ICC score** highlights an essential methodological insight: when conversation quality is **uniformly high**, **minor rater differences** become magnified in statistical measures, which must be accounted for in future evaluations.

6.4 Future Directions and Final Reflections

Future research should aim to introduce more controlled diversity in dialogue quality to better calibrate evaluation frameworks. Incorporating human evaluators alongside LLM judges would offer a more grounded benchmark. Exploring more granular evaluation dimensions, such as dynamic emotional tracking and ethical judgment, could further enhance the assessment fidelity. Additionally, leveraging larger, domain-specialized LLMs and incorporating real-world counseling datasets would significantly improve the ecological validity of the models. In reflection, this study successfully met its initial goals while uncovering deeper challenges associated with multi-turn dialogue evaluation. It revealed critical lessons about data variability, model assessment, and the nuanced behavior of inter-rater reliability metrics like ICC. By addressing a relatively underexplored area in mental health dialogue systems, this research lays a valuable foundation for future work, demonstrating that thoughtful data augmentation, model fine-tuning, and scalable evaluation are crucial pillars for building the next generation of empathetic AI-driven counseling tools.

References

[1] M. Abbasian, E. Khatibi, I. Azimi, *et al.*, *Foundation metrics for evaluating effectiveness of healthcare conversations powered by generative ai*, 2024. arXiv: 2309.12444 [cs.CL]. [Online]. Available: https://arxiv.org/abs/2309.12444.

[2] R. Anil, A. M. Dai, O. Firat, *et al.*, *Palm 2 technical report*, 2023. arXiv: 2305.10403 [cs.CL]. [Online]. Available: https://arxiv.org/abs/2305.10403.

[3] DeepSeek-AI, D. Guo, D. Yang, *et al.*, *Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning*, 2025. arXiv: 2501.12948 [cs.CL]. [Online]. Available: https://arxiv.org/abs/2501.12948.

[4] A. Q. Jiang, A. Sablayrolles, A. Mensch, *et al.*, *Mistral 7b*, 2023. arXiv: 2310.06825 [cs.CL]. [Online]. Available: https://arxiv.org/abs/2310.06825.

[5] J. M. Liu, D. Li, H. Cao, T. Ren, Z. Liao, and J. Wu, *Chatcounselor: A large language models for mental health support*, 2023. arXiv: 2309.15461 [cs.CL]. [Online]. Available: https://arxiv.org/abs/2309.15461.

[6] OpenAI, J. Achiam, S. Adler, *et al.*, *Gpt-4 technical report*, 2024. arXiv: 2303.08774 [cs.CL]. [Online]. Available: https://arxiv.org/abs/2303.08774.

[7] H. Qiu, H. He, S. Zhang, A. Li, and Z. Lan, "SMILE: Single-turn to multi-turn inclusive language expansion via ChatGPT for mental health support," in *Findings of the Association for Computational Linguistics: EMNLP 2024*, Y. Al-Onaizan, M. Bansal, and Y.-N. Chen, Eds., Miami, Florida, USA: Association for Computational Linguistics, Nov. 2024, pp. 615–636. DOI: 10.18653/v1/2024.findings-emnlp.34. [Online]. Available: https://aclanthology.org/2024.findings-emnlp.34/.

[8] Qwen, : A. Yang, *et al.*, *Qwen2.5 technical report*, 2025. arXiv: 2412.15115 [cs.CL]. [Online]. Available: https://arxiv.org/abs/2412.15115.

[9] P. E. Shrout and J. L. Fleiss, "Intraclass correlations: Uses in assessing rater reliability," *Psychological bulletin*, vol. 86, no. 2, pp. 420–428, 1979.

[10] C. Siro, M. Aliannejadi, and M. de Rijke, "Rethinking the evaluation of dialogue systems: Effects of user feedback on crowdworkers and llms," in *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, ser. SIGIR 2024, ACM, Jul. 2024, pp. 1952–1962. DOI: 10.1145/3626772.3657712. [Online]. Available: http://dx.doi.org/10.1145/3626772.3657712.

[11] H. Sun, Z. Lin, C. Zheng, S. Liu, and M. Huang, *Psyqa: A chinese dataset for generating long counseling text for mental health support*, 2021. arXiv: 2106.01702 [cs.CL]. [Online]. Available: https://arxiv.org/abs/2106.01702.

[12] T. Y. C. Tam, S. Sivarajkumar, S. Kapoor, *et al.*, *A framework for human evaluation of large language models in healthcare derived from literature review*, 2024. arXiv: 2405.02559 [cs.CL]. [Online]. Available: https://arxiv.org/abs/2405.02559.

[13] G. Team, R. Anil, S. Borgeaud, *et al.*, *Gemini: A family of highly capable multimodal models*, 2024. arXiv: 2312.11805 [cs.CL]. [Online]. Available: https://arxiv.org/abs/2312.11805.

[14] G. Team, T. Mesnard, C. Hardin, *et al.*, *Gemma: Open models based on gemini research and technology*, 2024. arXiv: 2403.08295 [cs.CL]. [Online]. Available: https://arxiv.org/abs/2403.08295.

[15] H. Touvron, T. Lavril, G. Izacard, *et al.*, *Llama: Open and efficient foundation language models*, 2023. arXiv: 2302.13971 [cs.CL]. [Online]. Available: https://arxiv.org/abs/2302.13971.

[16] Q. Xie, Q. Li, Z. Yu, Y. Zhang, Y. Zhang, and L. Yang, *An empirical analysis of uncertainty in large language model evaluations*, 2025. arXiv: 2502.10709 [cs.CL]. [Online]. Available: https://arxiv.org/abs/2502.10709.

[17] L. Zheng, W.-L. Chiang, Y. Sheng, *et al.*, *Judging llm-as-a-judge with mt-bench and chatbot arena*, 2023. arXiv: 2306.05685 [cs.CL]. [Online]. Available: https://arxiv.org/abs/2306.05685.