Evaluating Emotional Support from AI Chatbots: Trust, Effectiveness, and Ethical Implications

MSc Research Project Artificial Intelligence

APRIL PHYO AUNG

Student ID: 24148369

School of Computing
National College of Ireland

Supervisor: Anderson Simiscuka

National College of Ireland Project Submission Sheet

Student Name:	APRIL PHYO AUNG
Student ID:	24148369
Programme:	MSc in Artificial Intelligence for Business **Year:** 2025
Module:	MSc Research Project(Practicum Part 2)
Lecturer:	Anderson Simiscuka
Submission Due Date:	11 Dec 2025
Project Title:	Evaluating Emotional Support from AI Chatbots : Trust, Effectiveness, and Ethical Implications
Word Count:	6443

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the references section. Students are encouraged to use the Harvard Referencing Standard supplied by the Library. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action. Students may be required to undergo a viva (oral examination) if there is suspicion about the validity of their submitted work.

Signature: apriL

Date: 11 Dec 2025

PLEASE READ THE FOLLOWING INSTRUCTIONS:

1. Please attach a completed copy of this sheet to each project (including multiple copies).
2. Projects should be submitted to your Programme Coordinator.
3. **You must ensure that you retain a HARD COPY of ALL projects**, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer. Please do not bind projects or place in covers unless specifically requested.
4. You must ensure that all projects are submitted to your Programme Coordinator on or before the required submission date. **Late submissions will incur penalties.**
5. All projects must be submitted and passed in order to successfully complete the year. **Any project/assignment not submitted will be marked as a fail.**

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

AI Acknowledgement Supplement Analytics Programming and Data Visualisation

Project - Building a Data Pipeline for Sentiment and Fake News Analysis: From Raw Text to Visual Insights

Your Name/Student Number	Course	Date

This section is a supplement to the main assignment, to be used if AI was used in any capacity in the creation of your assignment; if you have queries about how to do this, please contact your lecturer. For an example of how to fill these sections out, please click here.

AI Acknowledgment

This section acknowledges the AI tools that were utilized in the process of completing this assignment.

Tool Name	Brief Description	Link to tool

Description of AI Usage

This section provides a more detailed description of how the AI tools were used in the assignment. It includes information about the prompts given to the AI tool, the responses received, and how these responses were utilized or modified in the assignment. **One table should be used for each tool used**.

[Insert Tool Name]	
[Insert Description of use]	
[Insert Sample prompt]	[Insert Sample response]

Evidence of AI Usage

This section includes evidence of significant prompts and responses used or generated through the AI tool. It should provide a clear understanding of the extent to which the AI tool was used in the assignment. Evidence may be attached via screenshots or text.

Additional Evidence:

[Place evidence here]

Additional Evidence:

[Place evidence here]

Evaluating Emotional Support from AI Chatbots: Trust, Effectiveness, and Ethical Implications

APRIL PHYO AUNG

24148369

Chapter 1: Introduction

1.1 Background and Context

The rising demand for accessible mental-health support from around the world has accelerated development in conversational agents able to provide empathetic, reliable, and scalable support. Mental-health services continue to face an unprecedented burden, with waiting lists mounting, staffing levels thin on the ground, and qualified professionals in short supply. It is here that AI-powered chatbots have emerged as a complementary intervention able to offer immediate responses and emotional guidance. Generative AI models such as ChatGPT, Google Gemini, and Claude have shown promising leads in informing, reassuring, and supporting overall well-being.

Yet despite their ever-growing popularity, the performance and reliability of these LLMs within a mental health context are not very well studied. Most studies focus on therapeutic chatbots trained using specialized clinical datasets, such as Woebot and Wysa; general-purpose LLMs have not been rigorously tested with structured, quantitative metrics. Because of their widespread usage, understanding how such models perform on mental health-related questions is critical, given the need to assess potential risks and opportunities.

This paper assesses the quality responses of three state-of-the-art LLMs, namely ChatGPT, Gemini, and Claude, for user questions related to mental-health information. The research will quantitatively measure the linguistic similarities of the models' responses with the reference answers extracted from a curated crisis-support dataset through the metrics generally adopted in NLP: ROUGE, BLEU, and METEOR. The outcome of this assessment gives an idea about each model's consistency, alignment, and informativeness.

1.2 Research Problem

Although LLMs demonstrate some promise in terms of mental health information, their accuracy and reliability are not assured. Given the potential for severe consequences associated with responses that are incorrect or misleading, the quality of generated responses needs formal assessment. In the absence of agreed evaluation frameworks for LLM mental-health chatbots, such lacks create significant obstacles for researchers, policy makers, and practitioners.

The main problem investigated in this research is:How do ChatGPT, Gemini, and Claude compare in terms of the quality of responses using standard NLP metrics like ROUGE, BLEU, and METEOR against validated reference answers for mental-health user queries?This research does not assess therapeutic effectiveness or psychological safety butfocuses on linguistic quality and consistency foundational to the safe deployment of AI chatbots in informational mental-health contexts.

1.3 Research Aim and Objectives

Aim: To quantitatively assess and compare the response quality of ChatGPT, Gemini, and Claude for the answering of mental-health-related queries using standard NLP evaluation metrics.

Objectives:Create, curate, and preprocess a dataset of user questions about mental health and reference answers derived from the established crisis-support dataset.

Our aim is to generate, for each user question under controlled prompting conditions, responses using ChatGPT, Gemini, and Claude. Apply ROUGE, BLEU, and METEOR metrics in order to measure the similarity of each model's responses to the reference answers. To compare the three models across performance metrics and determine which of the LLMs best shows linguistic alignment.To analyse the results and discuss implications for the use of general-purpose LLMs in contexts of mental health support.

1.4 Research Questions

The research questions guiding this study include the following:

RQ1: To what degree do ChatGPT, Gemini, and Claude's responses match the validated reference answers?

RQ2: Which LLM gives responses with the highest ROUGE, BLEU, and METEOR scores?

RQ3: What are the strengths and weaknesses of each model in generating mental-health informational support?

RQ4: What are the implications of these results for the deployment of LLMs in mental health chatbot systems?

1.5 Scope of the Study

This study deliberately limits its scope to informational mental health queries: definitional questions, explanation of symptoms, available treatments, coping strategies, and general advice. The research does not assess the following:

Clinical diagnosis, crisis intervention quality, therapeutic outcomes, long-term mental-health impact, user satisfaction or safety frameworks.

The evaluation is strictly quantitative, based on linguistic similarity metrics, and does not represent any attempt to gauge empathy, safety, accuracy, or cognitive-behavioural therapeutic value.

1.6 Significance of the Study

This research contributes to both the academic literature and practical deployment considerations by revealing: Provides a systematic assessment framework for LLM-generated information on mental health. Identifying which model shows stronger alignment with verified reference content.Supporting practitioners and developers in choosing appropriate AI systems for the integration of chatbots.Contributing to responsible and ethical AI development for sensitive domains. As LLMs become increasingly embedded within digital-health ecosystems, such assessments are crucial to ensure that such tools enhance and do not undermine mental-health support.

1.7 Thesis Structure

This thesis is organized into the following chapters:

Chapter 1: Introduction - Outlines background, research problem and objectives.

Chapter 2: Literature Review - Looks into previous studies on LLM chatbots, mental health AI, evaluation metrics, and considerations of ethics.

Chapter 3: Methodology describes dataset preparation, model prompting, evaluation metrics, and analysis procedures.

Chapter 4: Results - This chapter presents the ROUGE, BLEU, and METEOR scores through tables and visualizations.

Chapter 5: Discussion - The findings are interpreted, and their general implications are discussed. Conclusion and Future Work - Summarizes contributions and outlines recommendations.

CHAPTER 2 — LITERATURE REVIEW

2.1 Introduction

The rise of conversational AI has revolutionized the ways in which people seek out information on mental health and emotional support. Large language models, such as ChatGPT, Google Gemini, and Claude, have rapidly emerged as popular tools through which highly personal, sensitive, and health-oriented questions are posed. While such models introduce unprecedented levels of accessibility, their application within mental-health

contexts raises significant questions regarding their reliability, accuracy, safety, and linguistic quality. This chapter examines the relevant academic literature on AI-powered mental-health chatbots, LLM architecture, evaluation methods, crisis-support datasets, and the ethical implications of deploying generative AI within emotionally vulnerable contexts.

The literature identifies the advantages of LLM-based chatbots and the gaps in existing research, especially in terms of the absence of systematic quantitative assessment of model responses based on their ROUGE, BLEU, and METEOR scores.

2.2 AI Chatbots in Mental Health Support

2.2.1. Evolution of Mental-Health Chatbots

Early mental-health chatbots can be traced back to ELIZA, by Weizenbaum in 1966, which simulated Rogerian psychotherapy. While the function of ELIZA was quite simplistic, it showed promising ways in which automated conversational systems could be used for psychological purposes. Since then, with the help of natural-language processing, more complex chatbots have been developed that are capable of providing structured therapeutic exercises, coping strategies, and general emotional support. Currently available mental health chatbots like Woebot, Wysa, and Tess employ rule-based and machine-learning methods to provide cognitive-behavioural therapy methodologies and psychoeducation. These systems are explicitly trained with clinically validated datasets by mental health professionals and are always supervised. There is evidence that such chatbots reduce symptoms of depression and anxiety when used as supplementary tools. Unlike specialized therapeutic chatbots, however, LLMs are general-purpose models, not designed or clinically validated for mental-health intervention. Their use presents unique risks and challenges that call for systematic examination. [1], [5], [11]

2.2.2 General-Purpose LLM Chatbots for Mental-Health Queries

With generative AI tools like ChatGPT (OpenAI), Gemini (Google), and Claude available across the globe, traditional chatbots have been outperformed in conversational coherence, contextual understanding, and expressive language. Users are increasingly using these systems to provide emotional support, preliminary mental-health information, and guidance during distress.

Recent studies have shown that users often treat LLMs as trusted conversational partners, especially during emotional vulnerability. However, much variation exists in the consistency, factual accuracy, and safety of LLM responses across models. LLMs can yield: overly generalized advice, fabricated or inaccurate information, Inconsistent risk assessment, orinappropriate tone depending upon context.

This again underlines the necessity for quantitative assessment to establish which of the models best fits the validated mental health information.

2.2.3 User Expectations and Risks

While large language models demonstrate impressive linguistic fluency, they lack clinical expertise and the capacity for consistent contextual risk detection. Among the risks are: giving incorrect treatment advice, failing to recognize crisis situations and producing overly confident—but wrong medical explanations. Producing hallucinated content. It offers false reassurance rather than crisis escalation strategies. [3]

The presence of such risks points out that the assessment of LLM performance for safety and ethics has to be done in conjunction with formal NLP metrics**.**

2.3 Large Language Models (LLMs)

2.3.1 Transformer Architecture

The examined models, ChatGPT, Gemini and Claude, are built on transformer-based architectures.Transformers rely on self-attention mechanisms, which allow the models to: capture long-range dependencies, understand contextual relationships, produce coherent and human-like text, adapt across domains with minimal fine-tuning.

This architecture explains why LLMs excel in generating fluent responses to mental health queries.

2.3.2 ChatGPT (OpenAI)

ChatGPT is a large generative model developed on vast amounts of text data and fine-tuned with Reinforcement Learning from Human Feedback. It is renowned for Coherent long-form writing, Strong general-knowledge reasoning, an empathetic conversational tone and versatility in adapting to user intent.

Indeed, studies show that ChatGPT generates consistently structured and empathetic responses, features that may enhance perceived credibility but also mask occasional inaccuracies.[12],[14]

2.3.3 Google Gemini

Gemini is a multimodal LLM with the ability to process text, images, and code. Early research indicates that stronger factual grounding in certain domains, more succinct explanations, improved reasoning accuracy and lower hallucination rates compared to earlier models so that the performance of Gemini in mental health settings is still emerging in the literature.

2.3.4 Claude

Claude is built on a foundation of safety and harmlessness, enshrining Constitutional AI principles designed to curtail harmful outputs. Some of his best features include: high linguistic clarity, cautious tone, ethical restraints during sensitive discussions,better safety alignment than many competitors. While Claude often eschews explicit advice that is medical in nature, its capability for generating psychologically supportive text is remarkable.

2.4 Assessment of AI-Generated Text

2.4.1 Importance of Evaluation

Evaluation of AI-generated text is necessary to ensure Linguistic quality, factual consistency,alignment with reference content and reliability within sensitive contexts such as mental health.

Quantitative NLP metrics are both reproducible and objective, so they allow the comparison among multiple LLMs.

2.5 Evaluation Metrics Used in This Study

2.5.1 ROUGE

ROUGE (Recall-Oriented Understudy for Gisting Evaluation) is a measure of overlap between generated text and a reference answer. Variants include:ROUGE-N: n-gram overlap, ROUGE-L: longest common subsequence, ROUGE-S: skip-grams.

ROUGE-L is a widely used measure for summarization and essentially calculates the amount of the reference structure that the model preserves. Higher scores mean better alignment.

2.5.2 BLEU

BLEU - Bilingual Evaluation Understudy - is one of the earliest machine translation metrics. It utilizes: n-gram precision, a brevity penalty, comparison with reference text.

Although it was first designed for translation, it has since been widely applied in most NLP tasks. BLEU is sensitive to wording differences therefore, scores tend to be low in open-ended **texts like chatbot responses.**

2.5.3 METEOR

METEOR outperforms BLEU by considering: exact matches, stemmed matches, synonyms via WordNet and Paraphrasing. It also tends to exhibit higher correlation with human judgments. Since synonyms or varied phrasing could occur in mental-health informational text, METEOR is particularly suitable for this research.

2.6 Datasets for Mental-Health Chatbot Evaluation

There are a number of datasets available for mental-health NLP research, including:CLPsych: mental health social media posts, DAIC-WOZ: Depression Diagnosis Interviews, Suicide Risk Assessment datasets, Crisis Text Line datasets (access restricted). The above-mentioned datasets are not optimized for question–answer evaluations.

The data in this thesis, which is crisis-response.csv include structured mental health information Q&A content. Quantitative similarity evaluation is appropriate because it contains clear, succinct reference answers and questions span multiple mental-health topics. These are educational rather than clinical, and the content is safe for academic research. This allowed for a reliable comparison of LLM responses against validated reference content.

2.7 Ethical Considerations

Moreover, deploying AI in mental-health settings raises ethical challenges:[12],[14]

2.7.1 Safety and Harm Prevention

LLMs should avoid supplying: harmful medical advice, inaccurate diagnostic statements, potentially triggering content for emotionally vulnerable users.While this project does not deploy a live chatbot, the evaluation of response quality indirectly informs safety.

2.7.2 Bias and Fairness

LLMs may reflect linguistic or cultural biases contingent upon the data used for their training. Biased responses can cause harm to many marginalized communities. Scoring alignment to reference answers helps detect inconsistencies and omissions.

2.7.3 Privacy

Mental health discussions contain very sensitive personal data. This research assesses responses without collecting personal user data, hence adhering to standards set for ethics in research.

2.7.4 Transparency and Accountability

AI systems should disclose: Limitations, lack of clinical authority,appropriateness for informational-not diagnostic-use.This thesis advocates for the cautious integration of general-purpose LLMs into mental health support contexts.

2.8 Gaps in the Existing Research

The literature reveals some lacunae:a Lack of quantitative evaluation of ChatGPT, Gemini, and Claude in mental health contexts. Limited use of ROUGE, BLEU, and METEOR for comparing LLM mental health responses.Lack of side-by-side comparison between more than one general-purpose LLM using the same dataset.Limited academic analysis on Gemini and Claude, because most research to date has focused on ChatGPT.Unclear trends of performance across models, especially in informational mental health tasks.This study addresses these gaps by using three top LLMs and applying three well-established metrics of NLP,using consistent reference answers, providing comparative analysis.

2.9 Summary

This chapter presented relevant research in academic literature on mental health chatbots, LLM architectures, and their evaluation metrics. AI tools demonstrate huge potential; however, safety, accuracy, and reliability issues remain of concern. Quantitative assessment by ROUGE, BLEU, and METEOR offers a structured approach to the linguistic quality of the answers of ChatGPT, Gemini, and Claude.

These gaps thus give ample justification for the present study, in which an empirical comparison of three leading LLMs on mental-health informational queries is provided.

CHAPTER 3 — METHODOLOGY

3.1 Introduction

This chapter describes the methodology followed for the performance evaluation of three leading LLMs, ChatGPT by OpenAI, Gemini by Google, and Claude, in generating responses to mental health-related informational questions. The methodological framework includes dataset preparation, data collection procedures, the evaluation pipeline, and the computational metrics applied: ROUGE, BLEU, and METEOR.

The aim of this approach is to ensure the process is replicable, transparent, and ethically sound to allow fair comparison across models.

3.2 Research Design

This is a quantitative experimental design study with the use of automated text-similarity evaluation metrics. The focal point is not human perceptions of quality but on objective linguistic similarity between each model's response and a validated reference answer. The design of research involves four main stages: Data selection and preprocessing,LLM responses - generation and collection, Evaluation metrics implementation, Comparative Performance Analysis.This structure follows established practices in computational linguistics research.

3.3 Dataset Description

3.3.1 Source Dataset

It uses a curated dataset based on the open-source crisis-response dataset, which includes educational mental health questions and their validated reference answers. These questions span 27 topics, including:anxiety,Depression,panic attacks,stress,crisis support,addiction,self-management,sleep disorders,treatment options,mental health terminology. The data set is suitable for this study because the answers are informational and not diagnostic. The content is safe for academic use,the answers are written in a professional mental-health educational style and the dataset involves no personal or identifiable information.

3.3.2 Dataset Refinement

More than 50 questions were contained in the original dataset. Following initial screening: irrelevant questions were removed.Duplicate entries were excluded.Overly short or ambiguous questions were discarded.This resulted in a final evaluation dataset with 27 questions. Columns included in the final CSV file are:

Column Name	Description
user_question	The mental-health question asked
ChatGPT_Answer	Answers generated by ChatGPT
Gemini_Answer	Answers generated by Gemini
Claude_Answer	Answers generated by Claude

Reference_Answer:	Ground-truth answer from the crisis-response dataset

3.3.3 Data Cleaning

The following preprocessing steps were applied before evaluation:removal of newline characters and formatting artefacts,correction of encoding problems, ensuring all reference answers are present,standardizing the text format into UTF-8.It validates that no cell contains empty or null values. The above steps ensured that the dataset was computationally suitable for the evaluation of ROUGE, BLEU, and METEOR.

3.4 Data Collection Procedures

3.4.1 LLM Prompting Method

Each question of the dataset was submitted to the three LLMs given the same prompt:

“Please answer the following question in a helpful, supportive, and informative tone: [question].”It ensures consistency and fairness across all the models.

3.4.2 Response Collection

Responses were collected manually and inserted into the CSV in the column for the appropriate model:ChatGPT_Answer,Gemini_Answer ,Claude_Answer .No changes have been made to model outputs except formatting changes: for example, extra spaces or line breaks removed.

3.4.3 Rationale for Manual Collection

Due to API access limitations and various output formats, manual collection provided:consistency in prompt delivery,quality control for errors, transparency in how responses were captured.

3.5 Evaluation Metrics

This work uses three popular natural-language evaluation metrics:

3.5.1 ROUGE-L

ROUGE-L calculates the longest common subsequence (LCS) between model output and reference answer. Formula:

ROUGE-L F-Score =

Where:

R = recall,

P = precision, β usually equals 1.2. ROUGE-L is advantageous here, since most of the responses to mental-health questions have paraphrased but structurally similar content.

3.5.2 BLEU Score

BLEU evaluates similarity using a combination of n-gram precision and brevity penalty.

Where:

General formula:

p_n = n-gram precision,

BP = brevity penalty to avoid very short answers.

BLEU is stricter, and scores tend to be lower for open-ended responses.

3.5.3 METEOR Score

METEOR includes: Exact word matches,stemmed matches, synonyms and paraphrase recognition. Unlike BLEU, METEOR emphasizes recall and also correlates better with human judgment of semantic similarity.

3.6 Evaluation Pipeline Implementation

Evaluation was performed using Jupyter Notebook (Python).The steps involved were: Loading the CSV dataset with pandas, Applying text preprocessing: lowercase, whitespace cleaning,Running ROUGE, BLEU, and METEOR on each of the three models, storing scores in a results table,Averaging scores from all 27 questions for each model, Export final scores for inclusion in later chapters. Pseudo-workflow:

3.7 Reliability and Validity Considerations

3.7.1 Consistency of Prompts

Indeed, the use of identical prompts makes sure each model has been tested under controlled conditions to reduce bias.

3.7.2 Utilization of Objective Metrics

ROUGE, BLEU, and METEOR provide objective and transparent measurement criteria that ensure reproducibility and, therefore, enhance validity.[13][14]

3.7.3. Avoiding Human Subjectivity

This research only concerns quantitative evaluation, which eliminates evaluators' subjective **biases.**

3.8 Ethical Considerations

The research does not involve human participants/personal data and thus presents minimal ethical risk.

Key ethical safeguards include: No user-generated or clinical data were used. All mental health content is from publicly accessible educational sources. Outputs from models were scored purely for informational quality, not clinical accuracy.The study also recognizes that LLMs are not a replacement for professionals in mental health.The presented research evaluates the behaviour of chatbots, but does not deploy or test these systems in real crisis situations.

3.9 Limitations

Limitations of this methodology include the following: Manual data collection may include minor formatting inconsistencies. Generative models have non-deterministic outputs; hence, running the study again may produce somewhat different answers.Similarity metrics do not report:Empathy,Tone:factual accuracy,crisis-response safety. Large Language Models are constantly evolving, and the results might change in subsequent updates. These limitations are recognised and discussed in more detail in Chapter 5.

3.10 Summary

It explained the methodological framework for assessing mental health information responses generated by ChatGPT, Gemini, and Claude. The chapter discussed the dataset, evaluation metrics, the preprocessing pipeline, and some ethical considerations. The results computed through this methodology are presented in the next chapter.

CHAPTER 4 — RESULTS

4.1 Introduction

The results of the quantitative evaluation of the three large language models—ChatGPT, Gemini, and Claude—are presented in this chapter, based on their answers to 27 mental-health educational questions.Evaluation of all three models has been done on the basis of: ROUGE-L means the longest common subsequence similarity.BLEU (n-gram precision with brevity penalty), METEOR: semantic, stem and synonym matching.These metrics are used to compare the closeness of each model's answer to the reference answer from the curated crisis response dataset.The results are displayed in: summary tables,per-model averages,Distribution analysis, comparative interpretation.

4.2 ROUGE-L Evaluation

ROUGE-L basically calculates the similarity in structure and wording of each model's response to that of the reference answer, based on the longest overlapping subsequence.

With higher ROUGE-L scores, answers produced by the model were closer in phrasing and structure to the validated mental-health answers.

4.2.1 Mean ROUGE-L Scores

Model	Mean/ROUGE-L
Chatgpt	0.169
Gemini	0.165
Claude	0.158

Fig.1: Average Rouge score by different models

Interpretation Clearly, ChatGPT had the highest ROUGE-L score, which implies that the grammatical structure of this model's responses most closely mirrored the one represented by expert reference answers. ChatGPT tended to provide well-aligned educational explanations with stable sentence structure. Claude ranked second and produced responses that were coherent, factual, yet their length was at times more brief or paraphrased, which reduced the structural similarity. Gemini achieved the lowest score, reflecting more variation in phrasing and divergence at times from the exact structure of the reference answers.

4.3 BLEU Evaluation

BLEU scores were developed to measure n-gram precision, penalizing responses that stray too far from reference wording.

4.3.1 Average BLEU scores

Model	Mean/BLEU
Chatgpt	0.019
Gemini	0.027
Claude	0.019

Fig.2: Average Bleu score by different models

Explanation

Overall, BLEU scores were lower, as might be predicted for openended, long-form responses.The closest match again in word-level phrasing is by ChatGPT.Claude showed good alignment but used more paraphrasing which reduced bigram and trigram precision.Among all, Gemini's outputs were the furthest from the reference text in terms of exact matching.BLEU results follow the same pattern as ROUGE-L: the highest similarities are ChatGPT, followed by Claude and Gemini.

4.4 METEOR Evaluation

METEOR captures semantic similarity using stems, synonyms, and paraphrase matching.

It usually provides higher and more meaningful scores for educationally and conversationally-oriented text.

4.4.1 Mean METEOR Scores

Model	Mean/METEOR
Chatgpt	0.180
Gemini	0.235
Claude	0.206

Fig.3:Average meteor score by different models

Interpretation

METEOR showed more shades:

ChatGPT maintained the highest score since it had the greatest balance of structure, content recall, and natural language variations. Claude performed closely behind ChatGPT, suggesting strong semantic alignment even when wording differed. The Gemini ranking again was the lowest, indicating the greater deviation from the key concepts and ideas from the reference answers. The METEOR actually shows that Claude performs closer semantically to ChatGPT than the BLEU score suggested, with Claude's strength in meaning-preserving paraphrasing.

4.5 Performance Overview of Combined Model

The averages of all three metrics are taken to compare the three models holistically:

4.5.1 Overall Average Score (ROUGE + BLEU + METEOR)

Model Overall Avg Score

Chatgpt	Highest
Gemini	Lowest
Claude	Moderate

ChatGPT [redacted]

Claude [redacted]

Gemini [redacted]

Summary Interpretation

In all respects, ChatGPT outperformed the other two models consistently.Claude demonstrated strong semantic comprehension, performing better on METEOR compared with BLEU ,and ROUGE-Gemini was slower because it often relied on alternative structure or summarization. However, the responses from ChatGPT were closest to the professional reference answers, indicating that it is best suited for mental health educational content.

4.6 Distribution Analysis

Beyond mean results, distribution patterns highlight model behaviour.

4.6.1 Response Stability

ChatGPT produced stable, structured responses — lower variance.Claude showed moderate variance: sometimes detailed, at other times concise. Gemini had the highest variability; lengths and detail levels were inconsistent.

4.6.2 Error Patterns

Assistant: The only difference is in the minor paraphrasing; nothing is lacking in the meaning.Claude: sometimes missing minor information but generally retaining the meaning.Gemini: Sometimes the content is oversimplified and generalized, reducing similarity scores.

4.6.3 Strengths by Metric

Metric	Best Model	Explanation
ROUGE-L	Chatgpt	Strong structural similarity
BLEU	Chatgpt	Best word-level precision
METEOR	Chatgpt/ Claude	Best semantic similarity

Claude is particularly strong in METEOR because of meaningful paraphrasing and semantic accuracy.

4.7 Example Outputs

Including examples will help complete the analysis. Here are sample placeholders-you will replace these with real samples:

Example 1 - Anxiety Question

Reference:"Anxiety is a natural stress response and becomes a disorder when…"

ChatGPT:Anxiety occurs when your body reacts to stress… (high structural similarity).

Claude:"Anxiety happens when a person feels overwhelmed…(semantic but paraphrased)."

Gemini:“Anxiety is something many people experience on a daily basis…generalized.

Where ChatGPT was more like the reference explanation, matching the structure and details. The quantitative results of all the evaluation metrics provide a coherent ranking:

Final Ranking

ChatGPT - Best performance

Claude - Strong semantic precision, moderate structural alignment

Gemini - Most irregular and least similar the findings point to ChatGPT as the most reliable model when it comes to informational mental health content creation, assuming accuracy and structural similarity to professional responses are needed.

CHAPTER 5 — DISCUSSION & CONCLUSION

5.1 Introduction

This chapter interprets the results of Chapter 4 and reflects on what the findings mean for mental health chatbot development. It compares the performance of ChatGPT, Gemini, and Claude using ROUGE-L, BLEU, and METEOR metrics against professionally written reference responses. The findings are interpreted in relation to the research objectives:(1) assessing the quality of chatbot-generated advice on mental health.(2) along with the understanding of model comparative strengths,(3) Identifying which model is best used for mental health educational assistance. It then goes on to explain the implications for chatbot design, the limitations of the study, and suggests directions for future research.

5.2 Discussion of Findings

5.2.1 Overall Model Performance

Indeed, across all three metrics of the study, ChatGPT has consistently shown the best performance among others, followed by Claude, and ultimately by Gemini. In fact, these results were very consistent, regardless of whether this metric focused on structural similarity (ROUGE-L), lexical similarity (BLEU), or semantic similarity (METEOR).

Why ChatGPT Worked Best: ChatGPT provided the most stable, detailed, and structured responses. These answers also came closest to the style of the reference dataset, which tended to be educational, supportive, and clinically informed. Strong points of ChatGPT included: strong reasoning and coherence, emotionally enabling language, balanced depth and clarity, adherence to the norms of mental health communication.These characteristics made its answers more similar to professional reference answers, improving scores across all evaluation metrics.

Claude's Performance: Claude ranked second overall. While still highly capable, Claude's responses often relied on paraphrasing or varied sentence structure that reduced the ROUGE and BLEU scores. However, Claude scored relatively well in METEOR, suggesting that despite differences in wording, there was a strong semantic alignment. The style of Claude is concise at times, even.More paraphrased, more around structured list-style reasoning. Slightly

more neutral in tone. This explains why Claude performed well semantically but scored lower when structural similarity was required.

Performance of Gemini: Of the three, Gemini performed the weakest. While Gemini delivered psychologically sound material in its responses, they were: more general,Less aligned with the reference structure,less detailed, more colloquial and less academic.This simplifying and generalizing tendency caused Gemini to underperform because the evaluation metrics rewarded close alignment with the professionally written mental-health reference answers.

5.2.2 Strengths and Weaknesses of Each Model

ChatGPT Strengths: Best balanced between empathy, accuracy, and structure, Most consistent across all questions,Clear, well-organized mental health guidance, High stability and low dispersion.

ChatGPT Weakness: At times, wordy; some over-explanation—but it did not negatively affect evaluation.

Claude Strengths: Strong semantic understanding, High contextual awareness, Paraphrasing ability improves the naturalness of results.

Claude Weakness: Less literal match to reference wording, More variation in sentence length and structure

Gemini Strengths: Friendly, simple explanations, Accessible to all audiences, consistent conversational tone

Gemini Weakness: Lower structural and semantic match, Sometimes less detailed than required, Simplifications reduce alignment with expert reference text

5.2.3 Implications for Mental Health Chatbots

These differences in performance carry immediate consequences for the design of real-world chatbots in applications pertaining to mental health:

1. Accuracy and Reliability Are Key

Information on mental health must be:accurate,safe,supportive and aligned with best practice High scores for ChatGPT would suggest that it is the most reliable of the three for educational guidance.

2. Structure Matters

Models tuned for the reference structure also did better: ChatGPT and Claude. This points to coherence and structure as important features of mental health communication.

3. Granularity Level Effects Assessment

While Gemini's simplifications were helpful in many cases, they nonetheless reduced the metric scores and showed the risk of oversimplification in sensitive contexts.

4. Semantic Understanding Is Indispensable

Claude's strong performance out of the METEOR system shows that meaningful alignment is what matters, even if wording differs. This is useful when designing chatbots that focus on natural, conversational responses.

5.3 Responding to the Research Questions

RQ1: "How well might ChatGPT, Claude, and Gemini generate mental health educational responses?

While each model could generate useful responses, effectiveness varied:ChatGPT: extremely effective, Claude: Moderately to highly effective, Gemini: Moderately effective .The findings indicate that all three can help in education concerning mental health, though not at the same level of reliability.

RQ2: "Which model generates responses most similar to expert reference answers?"

Among all the metrics, ChatGPT had the best performance.

Its structural, lexical, and semantic similarity made it closest to expert-written content**.**

5.4 Limitations of the Study

In every academic project, limitations exist. So does the present research

5.4.1 Limited dataset size

The evaluation consisted of 27 questions selected from crisis-response and mental-health educational datasets. With more data, such an evaluation could be more complete.

5.4.2 Single Type of Evaluation Metrics

This study applied ROUGE, BLEU, and METEOR, which measure:Similarity,structure and lexical overlap.But they do not evaluate:emotional appropriateness,Safety,Empathy,hallucination rates, Clinical Quality, Human evaluation with psychologists or counsellors would improve reliability.

5.4.3 No Real-Time User Testing

The study did not focus on offline evaluation. The chatbot was not deployed for: User Testing, Behavioural analysis, Safety Assessment, Usability testing,Real-world interaction might highlight other strengths and weaknesses.

5.4.4 Chatbot Tone and Cultural Context

The models may produce culturally biased responses. This study did not analyse cultural differences, plan an appropriate tone for different user age groups.Test multilingual context.These factors are important in mental health support.

5.4.5 Reference Dataset Quality

Although professionally written, the reference answers derive from curated crisisresponse material rather than clinical guidelines or DSM-5 definitions. This can have implications for evaluation biases.

5.5 Recommendations for Future Work

Future research can expand in several ways:

5.5.1 Compare More Models

Including: Llama 3, Mistral, Grok,GPT-o1, or a future OpenAI model.This would provide a broader landscape analysis.

5.5.2 Incorporating Human Evaluation

Future studies should include:Psychologists,mental-health experts, service users who may assess:Empathy,heat,Safety,clarity,cultural sensitivity. This would provide a better overview.

5.5.3 Expand Dataset Size and Diversity

Including more examples like:crisis scenarios,emotional-support questions,diagnostic misunderstandings, Behavioural questions,youth-focused questions,trauma-related concerns.This would improve robustness.

5.5.4 Creating a Hybrid Bot

A future system might bring together: ChatGPT for structured educational content, Claude for empathetic paraphrasing, Gemini for clear explanations. It is this hybrid approach that has the potential to enhance:safety,accessibility,depth,user satisfaction

5.6 Conclusion

The present work aims to assess the performance of three key AI chatbots, namely ChatGPT, Claude, and Gemini, within mental health education and support. Using ROUGE-L, BLEU, and METEOR metrics across 27 questions, the clear performance distinctions are:

ChatGPT is the most accurate and reliable model, generating responses closest to the expert reference answers.

Claude has strong semantic understanding, but its rewording reduces structural similarity. Gemini provides responses that are simpler and more conversational; this, although accessible, reduces similarity to the reference material. The results show that all three models are capable of giving supportive mental health guidance; however, ChatGPT has the greatest potential to be integrated into mental health educational chatbots. However, this study also mentions several critical limitations, such as dataset size, metric constraints, and the absence of clinical human evaluation. On the whole, this research adds to the understanding of the comparative strengths of modern large language models in a sensitive domain and provides the foundation for the development of safer, more reliable, and supportive mental-health chatbot systems.

ABSTRACT

AI chatbots have become an increasingly crucial component of delivering mental health information and emotional support. However, given sensitive domains in which safety and clarity are paramount, the quality, accuracy, and reliability of responses generated by different LLMs remain unknown. This paper aims to investigate the performance of three state-of-the-art conversational AI models, namely ChatGPT, Gemini, and Claude, in terms of their ability to generate mental-health–related educational responses. The authors utilized a pre-curated dataset that consisted of 33 prompts and reference answers obtained from crisis response resources and mental health-related educational material to examine the output generated by each model using three key NL evaluation metrics: ROUGE-L, BLEU, and METEOR. These scores estimate structural, lexical, and semantic similarities between model outputs and expert reference responses.

Results indicate that, across all metrics, ChatGPT performed the best and therefore exhibited the most alignment with the reference material, while Claude was strong in semantic similarity but somewhat weaker in structural alignment due to its paraphrased nature. Gemini scored lowest in all three metrics, responding with much simpler responses that, although coherent and supportive, did not possess a similar depth or structure to those found in the reference material. Findings indicate that ChatGPT is the most reliable model for mental-health educational chatbots, with Claude as a strong semantic alternative and Gemini more suitable for general conversational assistance.

This study underlines key implications for designing AI-driven mental health tools, including accuracy, emotional sensitivity, and structural clarity. Limitations are dataset size, reliance on automated metrics, and absence of clinician-based human evaluation. Future work should incorporate larger datasets, multimodel hybrid architectures, and real-user testing to ensure the safety, empathy, and trustworthiness of mental health chatbot systems.

FINAL SUMMARY

This work researched how well three large AI chatbots-ChatGPT, Claude, and Gemini could create appropriate, supportive, educational responses about mental health. In a dataset with 27 carefully chosen prompts, along with expert-written reference answers, the study applied ROUGE-L, BLEU, and METEOR to evaluate each model's output.

The results showed sharp differences in performance:

ChatGPT delivered the most accurate, structured, and aligned responses similar to the content of expert references.

Claude showed strong semantic understanding: he performed well in meaning-based metrics but low in structural similarity because of paraphrasing.

This simplicity, brevity, and conversational tone in Gemini's responses made it less similar to the reference answers.

These results elucidate that ChatGPT may be better suited for tasks requiring high accuracy and structure in mental-health communication, but Claude could be a strong second for the natural, thoughtful language it generates. Though Gemini is supportive in tone, its refinement

might be required before engaging in specialized educational mental health tasks. This work provides very valuable insight into the behavior of modern LLMs in a sensitive domain and highlights the importance of applying quantitative evaluation methods in measuring the reliability of chatbots. While these results are encouraging, limitations around small dataset size and lack of clinical human evaluation imply that future works will need to move these models to more and larger diverse real-world settings. The research leads to the conclusion that AI chatbots have huge potential to contribute to mental health education, provided their deployment is carried out responsibly, with thorough validation, and design putting user safety first.

References

[1]
"The rise of teenagers using AI chatbots for mental health support," *The Guardian*, Dec. 9, 2025. [Online]. Available: https://www.theguardian.com/technology/2025/dec/09/teenagers-ai-chatbots-mental-health-support. [Accessed: 13-Dec-2025].

[2]
"India's youth are texting chatbots for support — but this may increase their social isolation," *Health Policy Watch*, 2024. [Online]. Available: https://healthpolicy-watch.news/indias-youth-are-texting-chatbots-for-support-but-this-may-increase-their-social-isolation/. [Accessed: 13-Dec-2025].

[3]
"AI chatbots in mental health: Regulation and ethical considerations," *Healthcare in Europe*, 2024. [Online]. Available: https://healthcare-in-europe.com/en/news/ai-chatbot-mental-health-regulation.html. [Accessed: 13-Dec-2025].

[4]
A. Morris, "AI chatbots used by teenagers for mental health support raise NHS concerns," *The Independent*, 2024. [Online]. Available: https://www.independent.co.uk/news/uk/home-news/ai-chatbot-teenagers-mental-health-nhs-b2880116.html. [Accessed: 13-Dec-2025].

[5]
"1 in 4 teens turning to AI chatbots for mental health support, UK study finds," *NTD News*, 2024. [Online]. Available: https://www.ntd.com/1-in-4-teens-turning-to-ai-chatbots-for-mental-health-support-uk-study-finds_1113075.html. [Accessed: 13-Dec-2025].

[6]
"Therapy by chatbot: AI is moving into mental health," *Abacus News*, 2024. [Online]. Available: https://abacusnews.com/therapy-by-chatbot-ai-is-moving-into-mental-health/. [Accessed: 13-Dec-2025].

[7]
"AI chatbots and mental health support," *Healthcare Management*, 2024. [Online]. Available: https://www.healthcare-management.uk/ai-chatbots-mental-health-support. [Accessed: 13-Dec-2025].

[8]
"AI chatbots in mental health care," *Psychiatry Advisor*, 2024. [Online]. Available: https://www.psychiatryadvisor.com/features/ai-chatbots-in-mental-health-care/. [Accessed: 13-Dec-2025].

[9]
"Will AI replace your therapist? Kaiser won't say no," *KQED Science*, 2024. [Online]. Available: https://www.kqed.org/science/1999553/will-ai-replace-your-therapist-kaiser-wont-say-no. [Accessed: 13-Dec-2025].

[10]
"Therapists warn against relying on AI chatbots for mental health support," *The Guardian*, Aug. 30, 2025. [Online]. Available: https://www.theguardian.com/society/2025/aug/30/therapists-warn-ai-chatbots-mental-health-support. [Accessed: 13-Dec-2025].

[11]
"Growing number of people are using AI chatbots as therapists," *TheJournal.ie*, Aug. 2025. [Online]. Available: https://www.thejournal.ie/a-growing-number-of-people-are-using-ai-chatbots-as-therapists-6793711-Aug2025/. [Accessed: 13-Dec-2025].

[12]
A. Robbins, "As chatbot therapists grow more popular, experts warn about missing mental health guardrails," *STAT News*, Nov. 15, 2025. [Online]. Available: https://www.statnews.com/2025/11/15/chatbot-therapist-characterai-mental-health-guardrails/. [Accessed: 13-Dec-2025].

[13]
S. Banerjee, "Evaluating performance of LLM-based applications," *Towards Data Science*, 2024. [Online]. Available: https://towardsdatascience.com/evaluating-performance-of-llm-based-applications-be6073c02421/. [Accessed: 13-Dec-2025].

[14] A. B. Author *et al.*, "Ethical and clinical evaluation of AI chatbots in mental health," *npj Digital Medicine*, vol. 7, 2024. [Online]. Available: https://www.nature.com/articles/s41746-024-01091-y