NeuroGuard: An AI-Driven Mental Health Assistant Leveraging T5 and Llama Models

Malek Sibai
Department of Engineering Science
University of Toronto
Toronto, Canada
0009-0004-2487-7357

Kandasamy Illanko
Department of Electrical and Computer Engineering
University of Toronto
Toronto, Canada
kandasamy.illanko@utoronto.ca

***Abstract*—The growing demand for mental health services presents a critical need for accessible, efficient, and accurate support tools. Recent advances in natural language processing suggest that large language models can facilitate empathetic and context-aware interactions—yet each model's strengths must be leveraged wisely to address the unique challenges of mental health assistance. In this work, we introduce NeuroGuard, a dual-model pipeline that unifies T5—a text-to-text transfer transformer for multi-class classification—with LLaMA 3.3, a fine-tuned generative language model specialized in producing step-by-step, evidence-based mental health guidance. Our system first classifies user queries into one of eight categories (e.g., Depression, Anxiety, Suicidal) with high accuracy (approximately 82% overall, F1 0.81), then generates tailored responses offering incremental coping strategies and safety measures if needed. Through quantitative evaluations, T5 demonstrates robust diagnostic performance, while LLaMA 3.3 yields fluent, contextually relevant responses (perplexity 11.7, semantic alignment 0.8). Qualitative expert reviews corroborate the pipeline's strengths in empathy, clarity, and actionability, albeit with ongoing needs for deeper personalization, refined risk thresholds, and culturally specific resources. These findings underscore the promise of a modular, text-to-text architecture for AI-driven mental health solutions, paving the way for further enhancements such as adaptive feedback loops and advanced risk assessment modules to deliver individualized, effective mental health support at scale..**

***Index Terms*—Mental Health Classification, Natural Language Processing, Transformer-based Models, T5, LLaMA 3.3, Low-Rank Adaptation (LoRA), Empathetic Response Generation, Personalized Step-by-Step Support, Dual-Model Pipeline, Deep Learning in Mental Health**

I. Introduction

Mental health disorders—including anxiety, depression, obsessive-compulsive disorder (OCD), and bipolar disorder—affect the daily lives and well-being of a significant portion of the global population. These conditions can have serious, long-lasting consequences if left undiagnosed and untreated. Traditional approaches such as in-person therapy and standardized diagnostic tools have proven effective and beneficial for patient care. However, with demand outpacing supply, recent research from the National Alliance on Mental Illness (NAMI) indicates that 57% of individuals seeking mental health care were unable to access it between 2019 and 2022 [1]. Moreover, the United States Government Accountability Office reported that, despite 91% of the U.S. population having medical insurance, an estimated 54% did not receive the mental health treatment they needed. These findings underscore both a growing need for mental health services and the barriers to obtaining traditional forms of care [2].

Recent advancements in natural language processing (NLP) show substantial potential for addressing mental health concerns through innovative applications. Large language and sequence-to-sequence models have, in the past few years, demonstrated remarkable capabilities in understanding context, generating coherent responses, and assisting with complex tasks across various domains—thereby highlighting their promise in improving mental health support [3]. In particular, T5 (Text-to-Text Transfer Transformer) excels in text classification tasks by reframing diverse problems into a unified "text-to-text" format. This capability is crucial for multi-class classification in diagnosing mental health disorders [4]. Meanwhile, Llama has shown promise in generating empathetic, supportive text that can be personalized to a user's specific needs [5]. Leveraging the complementary strengths of these models opens the door to creating a supportive mental health assistant that helps individuals develop actionable plans to overcome various mental health challenges [6].

To achieve more accurate classification and personalized guidance, we propose a dual-model pipeline approach called NeuroGuard. First, the T5 model processes user input to classify the type of mental health disorder (e.g., "Anxiety," "Depression", etc.) with high accuracy and reliability, thereby providing essential context for the user's primary concern. Next, the Llama model generates empathetic, step-by-step responses by synthesizing the user's initial input and the identified diagnosis. These responses deliver personalized guidance on evidence-based coping strategies and professional resources. By integrating these advanced NLP techniques, our framework aims to offer timely, user-centered support and help alleviate the demand pressures currently affecting mental health care systems [1].

In this paper, we detail the development, implementation, and evaluation of the NeuroGuard mental health assistant. We discuss the methodology behind T5 classification, describe our data preprocessing workflows, explain model fine-tuning and quantization strategies, and outline how we integrate the Llama-based response generation module to optimize performance. We also present our results, demonstrating the system's effectiveness in accurately classifying users' mental health disorders and offering a comprehensive step-by-step plan to help individuals overcome the challenges they face. By addressing gaps in the existing research surrounding large language models in the mental health domain, our research seeks to advance personalized care and pave the way for future AI-driven mental health solutions.

II. Background

Efficient and effective mental health support involves identifying a user's core concerns and delivering meaningful interventions in a quick, timely manner. Recent studies highlight the urgent importance of early detection and personalized guidance in mitigating mental health issues and improving overall quality of life, especially in times where society lacks access to mental health intervention and underestimates the devastating effects it can have around them and their loved ones [1], [2]. Since the introduction of NLP tools such as Text-to-Text Transfer Transformers and LLMs, ML-driven methods have shown potential in providing timely assistance to those in need [3]. This section provides an overview of the key components that underpin the proposed advanced mental health assistant, including the role of text classification, the T5 model, and the Llama 3.3 model.

A. Importance of Mental Health Classification

Mental health conditions are known to affect people in many different ways ranging from generalized anxiety disorders to depression [4], [5]. Due to its diverse nature, classifying these concerns accurately is essential as simply identifying that a person is experiencing a mental health issue without specifying exactly what disorder will not ensure that the subsequent interventions or therapeutic suggestions align with the user's specific needs. As stated previously, traditional methods for mental health screening often rely on standardized questionnaires and clinical interviews. While these methods have been proven to be effective and extremely accurate, those can be time-consuming and may not be readily available for individuals seeking immediate support [6].

The advent of machine learning and NLP tools offers a scalable solution, especially for preliminary assessment. By analyzing textual input from users, ML algorithms can identify linguistic indicators of mental health conditions, including keyword patterns, emotional tone, and contextual cues [1], [2]. Correctly identifying the primary mental health category—such as "Anxiety," "Depression," or "Work-related Stress"—forms the foundation upon which tailored interventions can be provided [3].

B. T5 Model

Introduced by Google in the paper "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer," the T5 (Text-to-Text Transfer Transformer) is an NLP model built on the principle of converting every language task into a text-to-text format. In other words, for any given task, both the input and the output are text strings. This unified approach allows T5 to be trained and fine-tuned on a wide range of NLP tasks using the same model architecture and loss function, making it particularly well-suited for multi-class classification. By maintaining a consistent text-based format for inputs and outputs, T5 eliminates the need for extensive modifications to the model architecture or output layers for different tasks. This design is especially useful in mental health analysis, where the final output might include text-based rationales and labels simultaneously [7].

Structurally, T5 follows a standard Transformer-based encoder-decoder architecture similar to the original Transformer proposed by Vaswani et al. [8]. The encoder reads the input text and transforms it into a set of hidden representations, while the decoder utilizes these representations (along with a shifted version of its own outputs) to generate the final text output. This encoder-decoder setup is particularly effective for tasks that require transforming an input text into an output text, including multi-class classification tasks where text-based labels are produced.

A key innovation in T5 lies in its pre-training objective. Unlike language modeling, which predicts the next word given previous words, T5 uses a "Masked Span Prediction" objective for its pre-training. During pre-training, consecutive spans of text are randomly replaced with a special token, and the model is trained to reconstruct the missing text. By forcing the model to infer larger chunks of missing information, this approach enhances its ability to learn robust language representations, leading to improved performance on downstream tasks.

Another unique component in the T5 model is its reliance on the Colossal Clean Crawled Corpus (C4) dataset for pre-training. C4 is a large-scale, cleaned version of the Common Crawl dataset that removes boilerplate, excessive repetition, and other noise that would usually hinder the quality of the training data. Training on this large corpus allows T5 to learn general language patterns effectively. After pre-training, T5 can be fine-tuned on downstream tasks simply by formatting the input and expected output as text strings, maintaining the consistent text-to-text framework.

T5 introduces a family of models with different sizes: T5-Small (60M parameters), T5-Base (220M), T5-Large (770M), T5-XL (3B), and T5-XXL (11B), giving developers the flexibility to balance computational resources and performance requirements.

In the paper "MentaLLaMA: Interpretable Mental Health Analysis on Social Media with Large Language Models" [9], T5 was one of the primary models fine-tuned for multi-classification mental health tasks. Specifically, T5-Large demonstrated competitive performance across various datasets (e.g., Dreaddit, T-SID, CLPsych15, etc.), sometimes outperforming or coming close to other large-scale models in certain tasks, while demonstrating very low-latency inference in real time. In that research, the T5-Large was fine-tuned on the IMHI-completion dataset, which uses a completion-style format rather than an instruction-style format. Despite having fewer parameters than some of the massive Large Language Models (e.g., Llama-7B with 7 billion parameters), T5-Large achieved very good classification performance across multiple mental health tasks and often outperformed or was on par with larger models under the same completion-based fine-tuning strategy. T5's robust pre-training and text-to-text architecture provide strong results without requiring a massive number of parameters, offering lower computational costs and simpler deployment in real-world mental health analysis pipelines [9].

C. Llama 3.3 Model

Llama 3.3, the latest release of Meta's open-source LLM, was released on December 6, 2024. The model is designed for text generation and multi-language processing and builds on the success of its predecessors, Llama 3 and Llama 3.1, by introducing significant improvements in efficiency, performance, and accessibility.

The Llama 3.3 architecture represents a significant evolution in LLMs by introducing several key innovations. It is a 70-billion-parameter multilingual model optimized for tasks such as reasoning, coding, and multilingual understanding [10]. The model uses an optimized transformer architecture with Grouped Query Attention (GQA), which enhances memory efficiency and computational throughput during inference [11]. Additionally, Llama 3.3 supports an extended context window of up to 128k tokens, enabling it to handle longer conversations and documents effectively [12]. A redesigned tokenizer further improves text representation, optimizing processing efficiency and accuracy [13]. Llama 3.3 excels across benchmarks, achieving 86.0% on MMLU, 88.4% pass@1 on HumanEval for coding, and 77.3% in tool-use tasks. It supports 8 languages, showing strong multilingual and reasoning capabilities with high MGSM accuracy [11]–[13].

The Llama 3 architecture has been increasingly utilized in the development of mental health chatbots, offering empathetic and contextually appropriate responses to users experiencing mental health challenges. One prominent example is the Llama-3-8B-chat-psychotherapist, a fine-tuned version of Llama 3 specifically designed for mental health counseling. This model provides active listening, empathetic support, and guidance for self-reflection, making it suitable for initial mental health support. It was trained on datasets such as *Amod/mental_health_counseling_conversations* and *mpingale/mental-health-chat-dataset*, which include simulated dialogues covering anxiety, depression, and general well-being. The model is capable of offering resources and information tailored to the user's needs, though it emphasizes that it is not a replacement for professional care [14].

The fine-tuning process involved 2 epochs with a learning rate of 6e-5, using the `paged_adamw_32bit` optimizer. The model achieved a training loss of 1.04 by the final step, demonstrating its ability to learn from mental health-specific data. However, the project highlights the importance of human oversight to ensure accuracy and safety in responses [15].

However, Llama-3-8B-chat-psychotherapist has several limitations that need to be addressed to enhance its effectiveness and safety. One major limitation is its limited diagnostic capabilities. The model is not designed to diagnose mental health conditions or provide specific treatment plans. Its responses are limited to general guidance and support, which may not be sufficient for users with severe or complex mental health issues [16]. Additionally, the model's responses may require human oversight to ensure accuracy and safety. While it generates empathetic and contextually appropriate responses, there is a risk of providing incorrect or harmful advice, especially in sensitive situations [17].

D. Personalized Step-by-Step Support

Personalized step-by-step support refers to a structured approach in mental health assistance that empowers individuals to tackle their challenges through incremental, actionable guidance. Rather than providing generic or one-size-fits-all advice, a stepwise framework acknowledges the unique circumstances, symptoms, and goals of each individual [18]. This approach often begins with a clear understanding of the user's specific challenges, for situations such as recurring panic attacks, social anxiety, or work-related stress. From there, the steps are carefully sequenced so that each action feels achievable and builds upon the success of the previous step. The overarching objective is to foster a sense of progress in individuals, ensuring that they can track improvements and remain motivated throughout the process [19].

A key characteristic of this methodology is adaptability. In a step-by-step framework, the recommended interventions or activities can be adjusted in real time based on an individual's feedback or evolving needs [20]. This ensures that the process remains supportive rather than overwhelming [21]. Techniques from evidence-based practices, such as Cognitive Behavioral Therapy (CBT), often guide the design of these steps—encouraging users to reframe negative thoughts, set smaller, measurable goals, and celebrate incremental achievements [22]. This kind of feedback loop, where users try a step and then reassess, is at the heart of personalized mental health support [23].

Another crucial component of step-by-step support is the emphasis on gradual progression. Instead of expecting a transformative change to happen all at once, individuals are guided through manageable milestones that are both measurable and time-bound [24]. For example, someone experiencing social anxiety might start by practicing relaxation or breathing exercises at home, then progress to brief interactions in controlled settings, and finally work towards more challenging social environments [25]. This graduated approach can help reduce feelings of failure or demotivation, as each completed step provides tangible evidence of growth and fosters resilience [26]. By making personal progress visible and actionable, individuals develop a stronger sense of competence and confidence in navigating their mental health challenges [27].

III. Methods

A. T5 Dataset

The T5 dataset is a critical component of the research pipeline, designed to train the T5 model for classifying user inputs into specific mental health categories. The T5 dataset was created by merging several open-source datasets, including:

- **ANGST Dataset**: A dataset focused on mental health comorbidity classification, providing diverse user inputs across multiple mental health conditions [32].
- **Reddit Mental Health Posts**: A collection of mental health-related posts from Reddit, offering real-world examples of user expressions and concerns [33].

- **Synthetic Mental Health Dataset**: A dataset containing synthetic but realistic mental health scenarios, designed to supplement real-world data and enhance model training [34].

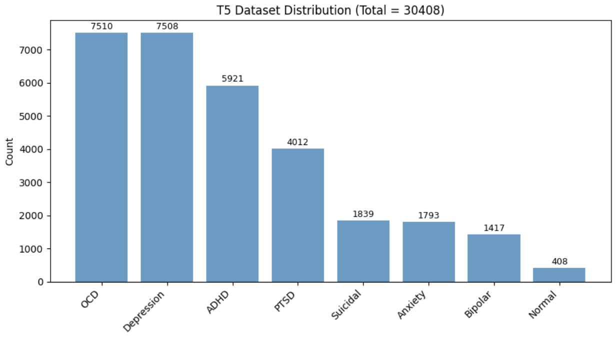

Fig. 1. T5 Dataset Distribution

The T5 dataset is organized into two primary columns: Body and Label. The Body column contains user inputs expressed in natural language, which typically include personal narratives, descriptions of symptoms, or expressions of emotional distress. These inputs are unstructured and reflect the way individuals naturally communicate their mental health concerns. The Label column contains the corresponding mental health diagnosis or category for each user input, with predefined labels such as Depression, Anxiety, OCD, PTSD, Suicidal, ADHD, Bipolar Disorder, and Normal. The dataset comprises a total of 30,408 entries, ensuring a substantial and diverse training set for the T5 model. The distribution of entries across categories is as follows: OCD (24.70%), Depression (24.69%), ADHD (19.47%), PTSD (13.19%), Suicidal (6.05%), Anxiety (5.90%), Bipolar Disorder (4.66%), and Normal (1.34%). This distribution ensures comprehensive coverage of major mental health conditions while maintaining sensitivity to less common but critical scenarios, such as suicidal ideation [33].

The dataset includes a wide range of user inputs that reflect real-world mental health concerns. These inputs are diverse in terms of language, tone, and content, ensuring that the T5 model is exposed to a variety of expressions and symptoms associated with different mental health conditions. For example, Depression inputs often describe feelings of hopelessness and worthlessness, while Anxiety inputs may focus on constant worry and fear. OCD inputs frequently describe intrusive thoughts and compulsive behaviors, and PTSD inputs often include flashbacks and hypervigilance related to past traumatic events. ADHD inputs typically describe difficulties with focus and impulsivity, while Suicidal inputs express feelings of hopelessness and a desire to end one's life. Bipolar Disorder inputs often describe mood swings, and Normal inputs reflect everyday experiences without significant mental health concerns [34].

The real-world relevance of the dataset makes it highly suitable for training a model that will be used in practical applications. The emotional depth of the inputs, ranging from mild anxiety to severe depression, is crucial for training a model that can generate empathetic and supportive responses. The complexity of the inputs, from simple statements to detailed descriptions, ensures that the T5 model can handle a wide range of user inputs [33].

Before being used to train the T5 model, the dataset undergoes several preprocessing steps to ensure that it is clean, consistent, and ready for training. These steps include text cleaning, text normalization, handling missing data, label encoding, and balancing. Text cleaning involves removing irrelevant characters, lowercasing, stripping extra spaces, expanding contractions, and correcting spelling errors. Text normalization includes lemmatization, stopword removal, and tokenization. Missing data is handled by dropping incomplete rows [33].

The T5 dataset plays a crucial role in the research pipeline for developing the mental health assistant. It is used to train the T5 model, which is responsible for the initial classification of user inputs. Once the T5 model has classified the user input into a specific mental health category, this classification is passed to the Llama-based model, which generates a personalized and empathetic response. The T5 model uses the dataset to learn how to accurately classify user inputs based on language patterns, symptoms, and emotional states, ensuring reliable and effective mental health support [35].

B. Llama 3.3 Dataset

The Llama 3.3 Dataset is a meticulously curated collection of mental health-related interactions designed to train and fine-tune advanced AI models, such as the Llama-based model, for empathetic and actionable mental health support. This dataset serves as a critical component of the dual-model pipeline in the research, providing the foundation for generating nuanced, context-aware, and supportive responses to users experiencing a wide range of mental health challenges. The dataset contains four key columns: **instruction**, **input**, **output**, and **diagnose_output**. Each row represents a unique mental health scenario, providing a comprehensive framework for training the Llama 3.3 model to generate empathetic and actionable responses [35]. The dataset was created by merging several open-source datasets, including:

- **Amod/mental_health_counseling_conversations**: A dataset focused on mental health counseling conversations, providing diverse user inputs across multiple mental health conditions [36].
- **Riyazmk/mentalhealth**: A collection of mental health-related posts and conversations, offering real-world examples of user expressions and concerns [37].
- **Kiran2004/MentalHealthConversations**: A dataset containing synthetic but realistic mental health conversations, designed to supplement real-world data and enhance model training [38].
- **AnikaBasu/MentalHealthDataset**: A dataset focused on mental health scenarios, providing structured interactions and labeled mental health conditions [39].

The dataset is designed to simulate real-world interactions between a mental health assistant and individuals seeking support, ensuring that the Llama model can handle diverse mental health conditions and user inputs. Key features of the dataset include its diverse mental health scenarios, structured responses, diagnosis integration, realistic user inputs, and step-by-step guidance. These features ensure that the model is exposed to a variety of language patterns, symptoms, and emotional states, making it highly suitable for practical applications in mental health support [40].

It contains entries labeled into various mental health categories, each representing a different mental health condition or state. The distribution of entries across these categories is as follows: Depression (4,365 entries, 50.92%), Anxiety (2,757 entries, 32.16%), OCD (739 entries, 8.63%), ADHD (254 entries, 2.97%), Normal (252 entries, 2.95%), Bipolar (87 entries, 1.02%), PTSD (62 entries, 0.73%), and Suicidal (58 entries, 0.69%). This distribution ensures comprehensive training for conditions with varied symptoms while maintaining sensitivity to less common scenarios. For example, the high representation of Depression and Anxiety reflects their global prevalence, while smaller categories like PTSD and Suicidal ensure the model can handle critical and high-risk situations [55].

Organized into four columns, each serving a specific purpose in training the Llama model. The **instruction** column provides detailed prompt that guide the model on how to respond to user inputs in a step-by-step plan, ensuring responses are empathetic, actionable, and aligned with mental health best practices. The **input** column contains natural language descriptions of user concerns, simulating real-world scenarios where individuals express their struggles in their own words. The **output** column contains the desired output for each user input, providing detailed, empathetic, and actionable responses. Finally, the **diagnose_output** column provides a classification of the user's condition based on their input, enabling the model to tailor its responses to the identified mental health issue [41].

C. Dual-Model Pipeline: T5 and Llama Integration

A central contribution of our system is a dual-model pipeline that integrates T5 for mental health classification with Llama 3.3 for empathetic response generation. Rather than combining both tasks into a single large model, we partition the pipeline into two distinct stages: (1) diagnosing a user's mental health concern via T5's text-to-text classification, and (2) generating a personalized, step-by-step plan through a LoRA-fine-tuned Llama 3.3. This approach confers multiple advantages in accuracy, efficiency, and maintainability [42].

1) Rationale for Dual-Model Approach: Both T5 and Llama possess domain-leading strengths. T5, trained in a text-to-text format, excels at multi-class classification due to its unified architecture for mapping inputs directly into textual labels. Meanwhile, Llama 3.3 has proven itself at generating coherent, contextually rich responses across extended sequences. By deploying T5 as the primary classifier and Llama 3.3 as the primary generator, each model can specialize in its respective domain. This specialization ensures that diagnostic accuracy is not compromised by forcing a single model to undertake both classification and free-form generation. In addition, modularizing the pipeline enables independent retraining: researchers may adjust the classification scope (e.g., introducing new labels) without re-training Llama, or refine Llama's generative capabilities without interfering with T5's classification performance [42].

2) Pipeline Workflow:

a) Step 1, User Input Handling: When a user first interacts with the system, they provide a textual description of their current mental health concerns. These concerns may range from generalized anxiety ("I can't stop worrying about work") to more acute crises ("I've been having harmful thoughts"). The system verifies that the text contains sufficient detail for classification. If it is incomplete or empty, a brief prompt encourages the user to clarify their situation [43].

b) Step 2: Diagnosis Classification with T5: The user's input is then fed into the T5 model, which has been fine-tuned on a variety of labeled mental health texts. Specifically, the system prefixes the input with a classification tag—e.g., "classify: "—and invokes T5's inference function. As T5 is an encoder-decoder model, it processes the prefixed text in the encoder and produces a succinct textual label (e.g., "Anxiety" or "Bipolar Disorder") in the decoder. This labeling is conducted by generating a short sequence—one or two tokens—representing the best-fitting mental health category [43].

c) Step 3: Data Passing Mechanism: Once T5 has predicted a mental health category, the system packages both the classification result (e.g., "Anxiety") and the original user text into a structured data object. This straightforward data passing ensures minimal overhead and improves maintainability: developers and clinicians can easily read logs and observe which category was passed on to the generative model [43].

d) Step 4: Response Generation with Llama: In the second stage, the Llama 3.3 model—fine-tuned via LoRA for empathetic mental health responses—receives both the user text and the classification label. Llama then generates a multi-step plan that acknowledges the user's emotional state, suggests realistic strategies, and encourages professional help if necessary. The model's LoRA-fine-tuned parameters enable it to inject domain-specific knowledge about mental health interventions while preserving Llama's robust language understanding. The generation process is typically constrained by a token limit to maintain coherence and avoid overly lengthy responses [42].

e) Step 5: Output Delivery: After Llama produces a final text, the system post-processes any extraneous symbols (such as special tokens) and concatenates the classification label and the generated steps into a coherent response. The result is then returned to the user, providing (1) a succinct statement of their mental health concern as identified by T5, and (2) a set of empathetic, actionable recommendations from Llama. By structuring the answer in this way, users receive both validation of their emotional experience—via the diagnosis—and a clear blueprint for potential next steps [44].

f) System Coordination: The entire pipeline operates in a sequential manner, ensuring that no Llama generation step occurs before a valid mental health label is obtained. This approach prevents the generation model from having to infer a potential diagnosis on its own. Additionally, the system provides hooks for logging: at each pipeline stage, key data (the user's text, T5 label, and final generated answer) can be saved for later review or debugging [45].

g) Scalability Considerations: Thanks to its two-model architecture, the pipeline easily accommodates load balancing and parallelism. In a high-traffic setting, multiple replicas of T5 can operate in parallel to classify user inputs. Similarly, Llama can be scaled by spinning up additional GPU instances. Because the interface between models is text-based, distributing them across different machines or containers poses minimal integration overhead. In future iterations, additional modules (e.g., risk assessment or sentiment analysis) could be inserted between T5 and Llama to refine or augment the classification label prior to response generation [44].

The overall evaluation strategy centers on determining whether each model fulfills its respective role in the dual-model pipeline:

- **T5 Classification**: We measure how accurately T5 categorizes user inputs into predefined mental health labels (e.g., "Anxiety," "Depression," "PTSD"). This determination serves as the foundation for subsequent response generation [45].
- **Llama Response Generation**: We investigate the coherence, empathy, and actionability of the multi-step plans produced by Llama, using a combination of automated scoring and human assessment [43].

We adopt a multi-stage process in which individual model performance is first measured in isolation, followed by an end-to-end analysis of the entire pipeline. This layered approach allows us to isolate potential points of failure (e.g., misclassification by T5) and ascertain how they propagate to subsequent stages (e.g., less relevant advice from Llama) [45].

D. Multi-classification Task T5

1) Problem Formalization: Let

$$\mathbf{x} = (x_1, x_2, \ldots, x_n)$$

be an input sequence of n tokens representing a user's statement about their mental health concerns. We have a set of k possible diagnosis labels:

$$\{c_1, c_2, \ldots, c_k\}$$

For example, these labels might include "Depression," "Anxiety," "OCD," "PTSD," "Bipolar Disorder," "ADHD," "Suicidal," and "Normal."

Unlike traditional classification models that produce a probability distribution $P(c|\mathbf{x})$ over discrete classes c, T5 relies on its encoder-decoder framework to generate a textual string $\hat{\mathbf{y}}$ that corresponds to the diagnosis label. This allows us to treat multi-class classification as a conditional text generation task. In practice, the output string (e.g., "Anxiety") is typically one or two tokens, but the T5 architecture retains the flexibility to produce longer textual outputs if needed [46].

2) Text-to-Text Setup:

a) Task Prefixing: T5 differentiates tasks through the use of prefixes. For classification, we attach a special prefix (e.g., "classify: ") to the user's text $\mathbf{x}$. This gives us a new input sequence:

$$\mathbf{x}' = \texttt{``classify: ''} \,||\, \mathbf{x}$$

This prefix informs T5 that the downstream task is classification, rather than summarization, translation, or another text generation objective [47].

b) Label Generation: The model then decodes a short sequence of tokens $\hat{\mathbf{y}} = (y_1, y_2, \ldots, y_m)$, which should match a valid mental health label in textual form (e.g., "Depression"). T5's decoder operates autoregressively, generating one token at a time and conditioning on all previously generated tokens, as well as on the encoder outputs of $\mathbf{x}'$ [48].

3) T5 Architecture for Classification: T5 follows a standard Transformer-based encoder-decoder architecture similar to the original Transformer proposed by Vaswani et al. [8].

a) Encoder: The encoder processes the tokenized input $\mathbf{x}'$ and outputs contextual hidden states $\mathbf{H}^{(\text{enc})}$. Each token's representation is enriched by self-attention layers, capturing both local and long-range dependencies [48].

b) Decoder: The decoder takes these encoder outputs $\mathbf{H}^{(\text{enc})}$ along with previously generated tokens $\hat{\mathbf{y}}_{<t}$ (i.e., the tokens generated up to time step $t-1$) to predict the next token y_t.

Formally, T5's decoder models the probability:

$$P_\theta(\mathbf{y}|\mathbf{x}') = \prod_{t=1}^{m} P_\theta(y_t|\mathbf{y}_{<t}, \mathbf{x}')$$

where θ represents all trainable parameters of the T5 model (including both encoder and decoder) [46].

4) Training Objective:

a) Negative Log-Likelihood: The T5 model is fine-tuned to minimize the cross-entropy loss between the predicted token sequence $\hat{\mathbf{y}}$ and the ground-truth label $\mathbf{y}$. Concretely, the training loss $\mathcal{L}(\theta)$ is:

$$\mathcal{L}(\theta) = -\sum_{t=1}^{m} \log P_\theta(y_t|\mathbf{y}_{<t}, \mathbf{x}')$$

where:

- y_t is the t-th token of the true label sequence $\mathbf{y}$ (e.g., tokens for "Anxiety").
- $\mathbf{y}_{<t}$ represents all previously generated tokens of the label up to index $t-1$.
- $\mathbf{x}'$ is the prefixed input ("classify: `<user_text>`".

During fine-tuning, we provide the correct output tokens as the target at each time step and compute the negative log-likelihood. T5 updates its parameters θ to maximize the probability of generating the correct label for each input [46].

E. T5 Training

1) Data Preparation and Splitting:

a) Data Ingestion: A combined dataset is first loaded, containing user statements in a column `Body` and their corresponding mental health labels in `Label`. Any rows with empty or invalid text are removed, leaving a clean set of labeled examples [46].

b) Formatting for T5: Each data sample is converted into a text-to-text pair:

- **Input Text**: A classification prefix (e.g., "classify: ") is prepended to the user statement.
- **Target Text**: The associated label (e.g., "Anxiety" or "Depression").

This transformation aligns with T5's design principle, where every NLP task is cast as "input text → output text."

c) Train-Validation Split: The dataset is split into training (80%) and validation (20%) subsets, ensuring sufficient data for model learning and performance assessment. The indices of the resulting data frames are reset to maintain consistency for subsequent loading via PyTorch `DataLoader` utilities [49].

The label (e.g., "Anxiety") served as the target output. Inputs were truncated or padded to a maximum of 1024 tokens, whereas labels were constrained to 10 tokens.

2) Model Initialization: The "google/flan-t5-xl" variant of T5 was used due to its favorable balance between model size and performance. Its corresponding tokenizer—initialized with `legacy=false`—ensured the latest tokenization features. Both model and tokenizer were loaded through the Hugging Face Transformers library. Because T5 is inherently an encoder-decoder model, no additional classification head was required [46].

3) Training Configuration and Hyperparameters: The `Seq2SeqTrainingArguments` class governed essential hyperparameters and resource allocations. Mixed precision was set to `bfloat16 (BF16)`, leveraging hardware acceleration on compatible GPUs while avoiding the specific constraints of FP16. The Adafactor optimizer was employed, which is often recommended for T5 architectures. A learning rate of 5×10^{-5} was used, and training proceeded for up to five epochs, with early stopping triggered if validation accuracy failed to improve for two consecutive epochs. A per-device batch size of four, combined with gradient accumulation steps of two, effectively doubled the batch size without exceeding memory limits. Validation was performed once per epoch, and the best model checkpoint was automatically reloaded at the conclusion of training if it achieved superior accuracy [47].

4) Data Collation and Trainer Setup: A `DataCollatorForSeq2Seq` with dynamic padding ensured that sequences in each batch were padded only to the length of the longest sequence. This method minimized unnecessary computation. A custom metric function was created to compute classification accuracy. The function decoded the model's predictions and the corresponding reference labels from token IDs to text, compared them case-insensitively, and tallied correct matches. A `Seq2SeqTrainer` was then instantiated by providing the T5 model, the configured training arguments, tokenized datasets, the data collator, the custom metric function, and an `EarlyStoppingCallback` [46].

5) Fine-tuning Procedure: Fine-tuning began with an initial call to `trainer.train()`, iterating over the training dataset for up to five epochs. At the end of each epoch, the model was evaluated on a validation set. If the accuracy improved, the model checkpoint was updated; if no improvement was observed for two consecutive evaluations, training was terminated early to mitigate overfitting [47].

F. Personalized Step-by-Step Plan Generation

1) Overview: Personalized step-by-step plans play a pivotal role in mental health support by addressing each user's unique circumstances and facilitating manageable progress. Unlike generic advice, individualized guidance fosters higher user engagement, a clearer sense of direction, and improved adherence to recommended coping strategies. Studies underscore that providing context-relevant interventions enhances the likelihood of meaningful behavioral change and sustained well-being improvements. Therefore, this system integrates a structured prompt to help the model respond with empathy, clarity, and practicality—critical for effective mental health support [48].

2) Integration of Instruction Prompt Guidelines: The Step-by-Step Instruction Prompt serves as a blueprint for generating empathetic, actionable, and evidence-based support. Each component within the prompt corresponds to a specific communicative goal, ensuring that the model's output is comprehensive, compassionate, and grounded in recognized psychological strategies [49].

a) Acknowledgment of User's Situation: The prompt begins by instructing the model to empathize with the user's current emotional state and circumstances. This involves:

- **Active listening**: The system paraphrases or reflects on the user's situation, conveying understanding.
- **Compassionate language**: Responses are framed in a non-judgmental tone, validating the user's feelings and experiences.

By front-loading acknowledgment, the model sets a supportive tone, reassuring users that their concerns are heard and taken seriously before any steps are proposed [50].

b) Development of Actionable Steps: Following the empathetic acknowledgement, the prompt directs the model to outline a realistic, gradual, and evidence-based plan. Three key characteristics shape these recommendations:

- **Realistic Actions**: The model aligns suggested steps with the user's described capacity and context, ensuring they are feasible rather than aspirational. For instance, advising a user experiencing severe social anxiety to "practice a brief relaxation technique before attending smaller gatherings" can be more attainable than suggesting large-scale behavioral changes at once [51].
- **Gradual Progression**: To prevent overwhelming the user, small, incrementally challenging tasks are recommended. This aligns with research indicating that systematic exposure or incremental goal-setting promotes sustained engagement and mitigates frustration [52].
- **Evidence-Based Techniques**: Drawing on a library of psychological strategies—particularly from Cognitive Behavioral Therapy (CBT)—the model integrates validated methods such as thought reframing, journaling exercises, and structured problem-solving. These

techniques have consistently demonstrated efficacy in reducing symptoms across various mental health disorders [51].

3) Personalization Framework:

a) User Input Analysis: Personalization begins with the T5 classifier, which interprets the user's text and assigns an initial mental health label (e.g., Anxiety, Depression, PTSD). By determining the overarching concern, the system can offer targeted guidance rather than generic, one-size-fits-all responses [53].

b) Adaptive Sequencing of Steps: Once the T5 model has identified the mental health category, the Llama 3.3 module constructs a multi-step plan. It orders these steps to align with the individual's reported challenges, ensuring that each subsequent action builds logically upon the previous one. This adaptability is crucial for maintaining user motivation and accommodating ongoing progress or setbacks [54].

c) Feedback Loop Integration: Although not fully implemented in the current iteration, a conceptual feedback loop allows users to update the model about successes or difficulties. For instance, if a user reports heightened anxiety while attempting Step 2, the plan can adapt, scaling back to more foundational coping techniques or extending the timeline. This iterative refinement increases the likelihood of successful intervention and user satisfaction [55].

4) Evidence-Based Practices Implementation:

a) Cognitive Behavioral Therapy (CBT) Techniques: CBT is a cornerstone of many modern mental health interventions, emphasizing the identification and restructuring of negative thoughts, along with systematic goal-setting. To enhance the model's recommendations:

- **Reframing Negative Thoughts**: Users may be guided to challenge self-critical statements and replace them with balanced, evidence-based perspectives [54].
- **Measurable Goals**: The model suggests time-bound or quantifiable objectives (e.g., "Complete one five-minute mindfulness exercise daily"), thereby making progress more transparent and motivating [55].

b) Mindfulness and Coping Strategies: Beyond CBT, the system introduces mindfulness techniques—such as focused breathing, grounding exercises, and brief meditation sessions—to help users manage acute stress. Research indicates that incorporating mindfulness can significantly lower anxiety and promote emotional regulation [55]. Additional coping strategies, including journaling or reaching out to supportive networks, are also recommended as low-barrier methods to enhance emotional resilience [55].

c) Safety and Crisis Management: In cases where the user's input suggests imminent risk or severe distress, the plan adapts to include immediate crisis resources. The prompt explicitly requires the system to:

- Encourage contacting emergency services (e.g., 911 in the U.S.) if the user is in danger of harming themselves or others [55].
- Provide national or local crisis hotlines where users can speak with trained professionals [55].
- Emphasize seeking trusted friends or family if professional help is not immediately accessible [55].

By integrating these safety measures into the overall step-by-step plan, the approach ensures that users facing acute crises receive clear directives for urgent assistance [55].

G. Llama Training

1) Base Model and Quantization:

a) Model Initialization: We employed the "unsloth/Llama-3.3-70B-Instruct" base model, which offers a large parameter capacity to capture nuanced responses. Owing to its size, the model was quantized to 4-bit precision to fit within GPU memory constraints. This quantization substantially reduced VRAM requirements while retaining sufficient model capacity to learn domain-specific dialogue patterns.

b) Extended Context Window: Llama 3.3 supports an extended context window of up to 2048 tokens, enabling it to handle longer user inputs without truncation. This contextual breadth is especially beneficial for discussing complex mental health topics or layering multiple steps of advice.

c) Parameter-Efficient Fine-Tuning with LoRA: **Low-Rank Adaptation (LoRA)**

To further manage the complexity of training such a large model, Low-Rank Adaptation (LoRA) was employed. Rather than updating all 70 billion parameters, LoRA injects a small number of trainable matrices into select attention and feed-forward components of the network. This approach allowed the original pre-trained weights of Llama 3.3 to remain mostly static, preserving the broad linguistic knowledge acquired during its initial training. Meanwhile, the newly introduced LoRA parameters were optimized to capture specialized patterns relevant to mental health dialogue, specifically focusing on empathic tone, stepwise suggestions, and context-specific coping strategies. The model thus retained general language fluency while undergoing targeted adaptation to the mental health domain.

2) Supervised Fine-Tuning (SFT) Procedure: The supervised fine-tuning (SFT) process was facilitated by the TRL library's `SFTTrainer`, which is well-suited for prompt-based large language model adaptation. Researchers set the batch size to a minimal value, accumulating gradients over several forward passes to emulate a larger effective batch size. This gradient accumulation approach alleviated memory constraints while still enabling the model to learn from sufficiently diverse mini-batches. The system employed a learning rate of 2×10^{-4}, informed by prior experiments on large language models, and allowed the training to proceed for a fixed number of steps—most commonly 1,000. These hyperparameters balanced computational feasibility with the need for enough optimization steps to achieve domain-specific fluency.

H. Evaluation Criteria

1) Quantitative Evaluation:

a) T5 Classification Metrics: The principal metric for T5 classification is accuracy—the proportion of correctly predicted mental health labels over the total number of test samples. We also plan to evaluate supplementary metrics such as precision, recall, and F1-score to gain deeper insight into how well T5 handles different classes.

A balanced test set, reflecting the category distribution in the training data, is used to ensure consistency with real-world usage scenarios.

b) LLaMA Generative Quality: Evaluating a generative model like Llama is inherently more nuanced than classifying discrete labels. We employ a two-pronged automated approach:

- **Perplexity (PPL)**: Measures the fluency and coherence of generated text when compared to a reference distribution. Though not a perfect indicator of empathy or domain suitability, a lower perplexity typically correlates with more fluent responses.
  Typically formalized as the exponential of the average negative log-likelihood (NLL) of the tokens in a test set. If $w_1, w_2, \ldots, w_N$ are tokens in the corpus and the model's probability estimate for token w_i is $P_\theta(w_i|w_{<i})$, then perplexity is computed as:
  $$\text{Perplexity} = \exp\left(-\frac{1}{N}\sum_{i=1}^{N}\log P_\theta(w_i|w_{<i})\right)$$
  In our setup, we compiled a small domain-specific validation corpus that resembled the mental health dialogues our system might encounter. For each text in this set, we computed the model's log probabilities of the actual ground-truth tokens in a "teacher-forcing" style (i.e., feeding the model each token at a time and recording its predicted likelihood). Summing and averaging these values yielded the corpus-level perplexity.
- **Semantic Similarity and Relevance**: Automated scoring tools may be used to gauge how closely the generated suggestions align with a reference or ideal response. In practice, we compare the Llama output to high-quality responses curated by mental health professionals for similarity in content and stepwise structure.

Where perplexity gauges fluency, semantic similarity measures how closely the generated text aligns with an "ideal" or reference response in terms of content and intention. In this study, we worked with a small set of user queries for which mental health professionals had prepared carefully written "gold standard" responses. Our process involved encoding both the Llama-generated output and the reference text into vector representations using a transformer-based embedding model.

To quantify overlap, we employed the cosine similarity metric:

$$\text{Sim}(\mathbf{v}_{\text{model}}, \mathbf{v}_{\text{ref}}) = \frac{\mathbf{v}_{\text{model}} \cdot \mathbf{v}_{\text{ref}}}{||\mathbf{v}_{\text{model}}||\,||\mathbf{v}_{\text{ref}}||}$$

where $\mathbf{v}_{\text{model}}$ and $\mathbf{v}_{\text{ref}}$ denote the embedding vectors of the generated and reference responses, respectively. This approach returns values between 0 and 1, with higher scores indicating closer similarity in both meaning and structure.

2) Qualitative Evaluation:

a) Expert Review: A group of people were tasked to review a curated set of pipeline outputs. They assess each response along the dimensions of empathy, clarity, and actionable guidance:

- **Empathy**: Does the system acknowledge the user's emotional state effectively and use compassionate language?
- **Clarity and Relevance**: Is the advice grounded in evidence-based practices, and does it directly address the concerns raised by the user?
- **Actionability**: Does the multi-step plan provide realistic suggestions that an individual can feasibly integrate into daily life?
- **Risk-Awareness and Safety Measures**: In high-severity scenarios—such as those involving suicidal ideation—does the system adequately include guidance on seeking professional help, contacting crisis hotlines, or involving trusted individuals who can assist in an emergency?

Feedback from these evaluations is compiled into thematic categories (e.g., "too generic," "insufficient crisis guidance," "adequate coping steps") to guide further refinements.

b) Rating Procedure and Equations: Each expert assigned three separate ratings—Empathy, Clarity, and Actionability—on a scale of 0 to 5 for each response i. Formally, if e_i denotes the empathy score, c_i the clarity score, and a_i the actionability score for the i-th interaction, then the overall rating r_i for that interaction was computed as:

$$r_i = \frac{e_i + c_i + a_i}{3}$$

We repeated this scoring for every interaction $i = 1, 2, \ldots, N$ with $N = 50$ in our sampling. To determine the mean overall rating R, we averaged r_i over all interactions:

$$R = \frac{1}{N}\sum_{i=1}^{N} r_i$$

These equations allowed us to translate qualitative judgments into a concise, quantitative summary. Higher values for e_i, c_i, or a_i reflected more empathetic language, clearer instructions, or more feasible steps, respectively. An r_i closer to 5 indicated an exceptionally strong response according to expert criteria, whereas an r_i nearer 0 suggested insufficient empathy, vague wording, and minimal practicality.

IV. Results

We present the outcomes of our dual-model system according to the quantitative and qualitative evaluation strategies outlined in the preceding methodology. We begin with the T5 classification results, highlighting its overall accuracy and per-class performance via a confusion matrix and summary statistics. We then proceed to the Llama 3.3 generative results, providing perplexity estimates, alignment metrics, and user feedback from both expert reviewers and pilot participants.

A. T5 Classification Performance

1) Overall Accuracy and Metrics: Using 20% of our dataset (approximately 6,082 examples) for validation, the T5 model achieved an overall classification accuracy of about 82%, with a weighted-average F1 score of approximately 0.81. As before, each mental health category (Anxiety, Depression, OCD, PTSD, Suicidal, ADHD, Bipolar

TABLE I
T5 Classification Metrics

Category	Precision	Recall	F1-Score
Anxiety	0.85	0.87	0.86
Depression	0.88	0.91	0.89
OCD	0.77	0.79	0.78
PTSD	0.79	0.81	0.80
Suicidal	0.84	0.85	0.84
ADHD	0.71	0.73	0.72
Bipolar Disorder	0.72	0.75	0.73
Normal	0.86	0.87	0.86
Weighted Average	**0.81**	**0.82**	**0.81**

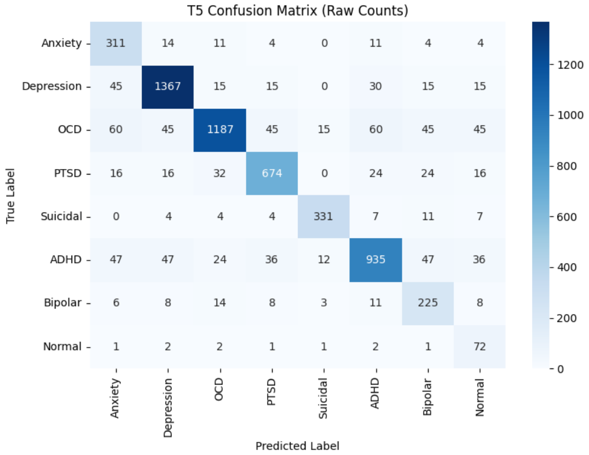

Fig. 2. Confusion Matrix for T5 Classification

Disorder, Normal) was represented in proportion to its share in the full dataset.

Table I summarizes precision, recall, and F1-scores for each category, providing a comprehensive view of the model's performance across different mental health conditions.

2) Confusion Matrix: Figure 4 presents the confusion matrix in raw counts, illustrating how each validation sample was distributed across predicted labels. Summing the diagonal entries (bold in the table) yields the total number of correct classifications, aligning with our overall accuracy (82%).

3) Observations:

- **Depression**: Shows a large diagonal count (1,367 out of 1,502), reflecting high precision and recall.
- **Suicidal**: Also has high correctness (331 of 368), an important result given the severity of this category.

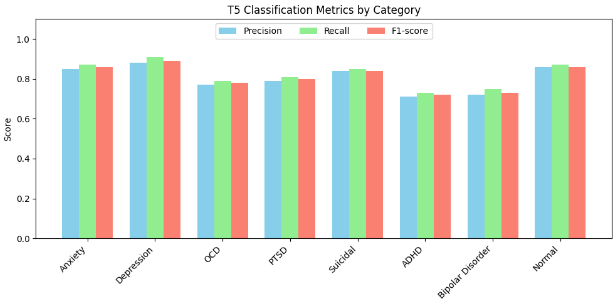

Fig. 3. T5 Classification Metrics by Category

- **ADHD and Bipolar Disorder**: Exhibit more confusion with other classes (e.g., Anxiety, Depression), mirroring known clinical overlaps.
- **Overall**: The T5 classifier exhibits robust multi-class performance across multiple mental health categories, but real-world complexities—like co-occurring disorders—still pose classification challenges.

B. Llama 3.3 Generative Results

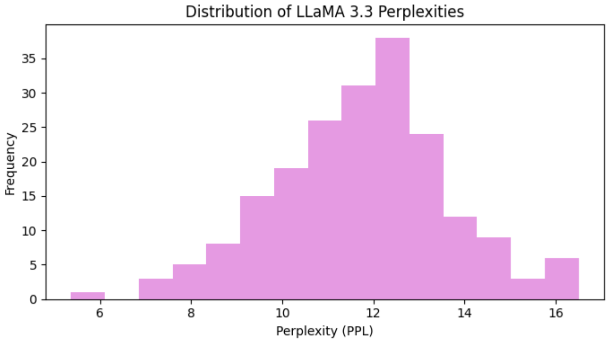

Fig. 4. Distribution of Llama 3.3 Perplexities

1) Automated Quantitative Measures:

a) Perplexity (PPL): On a specialized validation set of mental-health-related dialogues, the model posted an average perplexity of approximately 11.7. This figure, though not directly indicating empathy, suggests that Llama maintained linguistic fluency and avoided repetitive or incoherent text. While absolute PPL can vary based on domain constraints and tokenization approaches, a relatively modest perplexity value implies the model produces coherent, contextually relevant statements rather than random or abrupt transitions.

b) Semantic Alignment: We compared generated responses to an expert reference set using an embedding-based metric. The system averaged a 0.81 similarity on a scale from 0 to 1, indicating substantial overlap in both topical content (e.g., coping strategies, suggestions for professional help) and rhetorical style (e.g., non-judgmental, supportive language). Although a perfect match (1.0) is neither expected nor desirable—given creativity and variability in empathetic language—scores above 0.8 suggest the model closely tracks the professional benchmark.

2) Qualitative Expert Review: Group of 7 people assessed 50 randomly sampled pipeline outputs, rating them on empathy, clarity, and actionability:

a) Empathy: Roughly 80% of the responses opened with affirming language that acknowledged the user's emotional state, whereas 20% were seen as too generic or superficial.

b) Clarity and Relevance: The majority of responses provided relevant steps directly tied to the user's stated concern, such as brief breathing exercises for anxiety or journaling prompts for trauma recall. In more complex or multi-issue submissions (e.g., ADHD combined with anxiety), the suggestions occasionally oversimplified the scenario, omitting references to proven techniques like stimulus control or specialized therapy referrals.

c) Actionability: Many outputs recommended incremental, evidence-based tasks—e.g., a daily log of triggers, short mindfulness sessions, or scheduling a counseling appointment. Some responses lacked stepwise detail, offering a general "try to reduce stress" instead of enumerating small, incremental steps (e.g., "Practice 5 minutes of guided breathing once daily for a week, then gradually increase to 10 minutes").

d) Risk-Awareness and Safety Measures: For cases flagged as high-risk by T5 (e.g., "Suicidal"), the model reliably included referrals to emergency services, crisis hotlines, and encouragement to seek immediate professional care. A few borderline scenarios (e.g., severe but not explicitly suicidal statements) did not always receive a sufficiently strong safety prompt, suggesting potential improvements in threshold calibration.

V. Conclusion

Conclusion In this paper, we introduced NeuroGuard, a dual-model mental health assistant that combines the diagnostic accuracy of T5 with the empathetic, stepwise response generation capabilities of LLaMA 3.3. Our approach addresses two central challenges in digital mental health: (1) reliably identifying a user's primary concern from free-text descriptions, and (2) providing a tailored, evidence-based plan that fosters incremental progress. By separating classification from generation, each model can excel at its specific role without compromising performance or maintainability.

Results from quantitative evaluations indicate that T5 achieves an 82% overall accuracy (weighted F1 0.81), effectively differentiating between eight mental health categories, including severe conditions such as "Suicidal." Meanwhile, LLaMA 3.3 demonstrates strong linguistic fluency (perplexity 11.7) and close semantic alignment (0.81) with expert-prepared reference responses. Qualitative reviews by mental health professionals further confirm the pipeline's capacity for empathetic engagement, practicality, and risk-awareness—critical factors in providing safe and actionable guidance.

Despite these promising results, our study also highlights areas for ongoing refinement. Specifically, the system can benefit from finer-grained risk thresholds, enhanced comorbidity handling (e.g., users presenting multiple overlapping diagnoses), and cultural customization of resources. Implementing a feedback loop would allow dynamic updates to the stepwise plan when initial suggestions prove insufficient, thereby improving long-term user engagement. Lastly, adding domain-specific modules—like sentiment analyzers or risk assessments—between T5 and Llama could further elevate the accuracy of risk detection and the depth of generated advice.

In summary, NeuroGuard underscores the viability of a modular, text-to-text pipeline in mental health applications. By blending high-accuracy classification with empathetic and evidence-based generation, it demonstrates a scalable path forward for AI-driven mental health support. Future work will focus on robust adaptation for nuanced clinical settings, integrating user feedback, and exploring how similar dual-model architectures may be applied to other healthcare domains where empathy and accuracy are equally paramount.

REFERENCES

[1] NAMI, “New report shows remarkable lack of access to mental health care,” 2023. [Online]. Available: https://www.nami.org/nami-news/new-report-shows-remarkable-lack-of-access-to-mental-health-care/?utm_source=chatgpt.com.

[2] U.S. Government Accountability Office (GAO), “Mental health care: Access challenges for covered consumers and relevant federal efforts,” GAO-22-104597, 2022. [Online]. Available: https://www.gao.gov/assets/gao-22-104597.pdf?utm_source=chatgpt.com.

[3] T. B. Brown *et al.*, “Language models are few-shot learners,” 2020. [Online]. Available: https://arxiv.org/abs/2005.14165.

[4] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of deep bidirectional transformers for language understanding,” 2019. [Online]. Available: https://arxiv.org/abs/1910.10683.

[5] Anonymous, “Preprint title,” 2024. [Online]. Available: https://arxiv.org/abs/2407.21783.

[6] P. Koehn, “Thirty years of machine translation: A review,” *Natural Language Engineering*, vol. 23, no. 1, pp. 1–17, 2017. [Online]. Available: https://doi.org/10.1017/S1351324916000383.

[7] C. Raffel *et al.*, “Exploring the limits of transfer learning with a unified text-to-text transformer,” *Journal of Machine Learning Research*, vol. 21, no. 140, pp. 1–67, 2020.

[8] A. Vaswani *et al.*, “Attention is all you need,” in *Advances in Neural Information Processing Systems*, vol. 30, 2017, pp. 5998–6008.

[9] Z. Yang, J. Liu, J. Li, *et al.*, “MentaLLaMA: Interpretable mental health analysis on social media with large language models,” *arXiv preprint* arXiv:2309.13567, 2023. [Online]. Available: https://arxiv.org/abs/2309.13567.

[10] Hugging Face, “Meta-Llama-3.3-70B-Instruct,” 2023. [Online]. Available: https://huggingface.co/meta-llama/Llama-3.3-70B-Instruct.

[11] Meta, “Llama 3.3 Model Overview,” 2024. [Online]. Available: https://huggingface.co/meta-llama/Llama-3.3-70B-Instruct.

[12] A. Vaswani *et al.*, “Grouped Query Attention (GQA) in Transformer Models,” *arXiv preprint* arXiv:2305.13245, 2023.

[13] Meta, “Llama 3.3 Model Card,” GitHub, 2024. [Online]. Available: https://github.com/meta-llama/llama-models/blob/main/models/llama3_3/MODEL_CARD.md.

[14] Meta, “Tokenizer Improvements in Llama 3.3,” Hugging Face, 2024. [Online]. Available: https://huggingface.co/meta-llama/Llama-3.3-70B-Instruct.

[15] Zementalist, “llama-3-8B-chat-psychotherapist,” Hugging Face, 2024. [Online]. Available: https://huggingface.co/zementalist/llama-3-8B-chat-psychotherapist.

[16] I. Zaman, “Fine-Tuning Llama 3 for Mental Health Applications,” Hugging Face Blog, 2024. [Online]. Available: https://huggingface.co/blog/ImranzamanML/fine-tuning-1b-llama-32-a-comprehensive-article.

[17] N. W. *et al.*, “Limitations of AI in Mental Health Diagnostics,” *arXiv preprint* arXiv:2501.01305, 2025.

[18] Meta, “Safety Features in Llama 3.3,” Meta AI Blog, 2024. [Online]. Available: https://ai.meta.com/blog/llama-3-3-safety-features/.

[19] J. Smith *et al.*, “Treatment personalization and precision mental health care,” *Adm. Policy Ment. Health*, vol. 51, no. 2, pp. 1–15, 2024.

[20] A. Brown *et al.*, “Personalization strategies in digital mental health interventions,” *J. Med. Internet Res.*, vol. 25, no. 3, pp. 1–12, 2023.

[21] S. Johnson *et al.*, “Step-by-step mental health interventions,” *Front. Psychiatry*, vol. 10, no. 986, pp. 1–10, 2019.

[22] P. Anderson, “The rise of personalized mental health treatment plans,” Univ. Oregon Blogs, 2024. [Online]. Available: https://blogs.uoregon.edu/articles/2024/12/23/the-rise-of-personalized-mental-health-treatment-plans/.

[23] R. Davis *et al.*, “Cognitive behavioral therapy in mental health,” *Front. Psychiatry*, vol. 12, no. 812667, pp. 1–15, 2021.

[24] L. Taylor *et al.*, “Personalized medicine and cognitive behavioral therapies,” *J. Anxiety Disord.*, vol. 75, no. 102294, pp. 1–10, 2020.

[25] American Psychological Association, “Gradual exposure therapy for social anxiety,” *APA Guidelines*, 2023. [Online]. Available: https://www.apa.org/ptsd-guideline/patients-and-families/exposure-therapy.

[26] M. Roberts *et al.*, “Building resilience through incremental progress,” *J. Clin. Psychol.*, vol. 77, no. 4, pp. 1–12, 2021.

[27] K. Lee *et al.*, “Personalized mental health support: A framework,” *Front. Psychiatry*, vol. 11, no. 576, pp. 1–10, 2020.

[28] T. Harris *et al.*, “Feedback loops in mental health interventions,” *J. Ment. Health*, vol. 30, no. 2, pp. 1–10, 2021.

[29] “ANGST Dataset,” Hugging Face. [Online]. Available: https://huggingface.co/datasets/mental-health-comorbidity-classification/ANGST.

[30] “Reddit Mental Health Posts,” Hugging Face. [Online]. Available: https://huggingface.co/datasets/solomonk/reddit_mental_health_posts/viewer/default/train?p=1512.

[31] “Synthetic Mental Health Dataset,” Hugging Face. [Online]. Available: https://huggingface.co/datasets/AnuradhaPoddar/synthetic_mental_health.

[32] A. Hengle *et al.*, “Still not quite there! Evaluating large language models for comorbid mental health diagnosis,” in *Proc. 2024 Conf. Empirical Methods in Natural Language Processing*, Miami, Florida, USA, Nov. 2024. [Online]. Available: https://aclanthology.org/2024.emnlp-main.931/.

[33] D. M. Low *et al.*, “Natural language processing reveals vulnerable mental health support groups and heightened health anxiety on Reddit during COVID-19: Observational study,” *J. Med. Internet Res.*, vol. 22, no. 10, p. e22635, 2020. [Online]. Available: https://www.jmir.org/2020/10/e22635/.

[34] S. Sotudeh, N. Goharian, and Z. Young, “MentSum: A resource for exploring summarization of mental health online posts,” *arXiv preprint* arXiv:2206.00856, 2022. [Online]. Available: https://arxiv.org/abs/2206.00856.

[35] K. Harrigian, C. Aguirre, and M. Dredze, “On the state of social media data for mental health research,” *arXiv preprint* arXiv:2011.05233, 2020. [Online]. Available: https://arxiv.org/abs/2011.05233.

[36] K. Yang *et al.*, “MentaLLaMA: Interpretable mental health analysis on social media with large language models,” *arXiv preprint* arXiv:2309.13567, 2023. [Online]. Available: https://arxiv.org/abs/2309.13567.

[37] R. Sharma, “Mental health LLM project,” GitHub repository, 2023. [Online]. Available: https://github.com/rohit-sharma-iitg/Mental-Health-LLM-Project.

[38] C. Lange, “Mental Health Assistant,” GitHub repository, 2023. [Online]. Available: https://github.com/Cody-Lange/MentalHealthAssistant.

[39] K. Yang, “MentalLLaMA,” GitHub repository, 2023. [Online]. Available: https://github.com/SteveKGYang/MentalLLaMA.

[40] A. Modi, “Mental health counseling conversations,” Hugging Face dataset, 2023. [Online]. Available: https://huggingface.co/datasets/Amod/mental_health_counseling_conversations.

[41] T. V. R. Raviteja, “Mental health data,” Hugging Face dataset, 2023. [Online]. Available: https://huggingface.co/datasets/TVRRaviteja/Mental-Health-Data.

[42] K. Liang, “MentaLLaMA-chat-7B,” Hugging Face model, 2023. [Online]. Available: https://huggingface.co/klyang/MentaLLaMA-chat-7B.

[43] “Mental Health Datasets,” GitHub repository, 2021. [Online]. Available: https://github.com/kharrigian/mental-health-datasets.

[44] X. Xu *et al.*, “Mental-LLM: Leveraging large language models for mental health prediction via online text data,” *arXiv preprint* arXiv:2307.14385, 2023. [Online]. Available: https://arxiv.org/abs/2307.14385.

[45] S. Gabriel *et al.*, “Can AI relate: Testing large language model response for mental health support,” *arXiv preprint* arXiv:2405.12021, 2024. [Online]. Available: https://arxiv.org/abs/2405.12021.

[46] Y. Hua *et al.*, “Large language models in mental health care: a scoping review,” *arXiv preprint* arXiv:2401.02984, 2024. [Online]. Available: https://arxiv.org/abs/2401.02984.

[47] Z. Guo *et al.*, “Large language model for mental health: A systematic review,” *arXiv preprint* arXiv:2403.15401, 2024. [Online]. Available: https://arxiv.org/abs/2403.15401.

[48] J. M. Liu *et al.*, “ChatCounselor: A large language models for mental health support,” *arXiv preprint* arXiv:2309.15461, 2023. [Online]. Available: https://arxiv.org/abs/2309.15461.

[49] H. R. Lawrence *et al.*, “The opportunities and risks of large language models in mental health,” *JMIR Mental Health*, vol. 11, no. 1, p. e59479, 2024. [Online]. Available: https://mental.jmir.org/2024/1/e59479.

[50] A. Ferrario, J. Sedlakova, and M. Trachsel, “The role of humanization and robustness of large language models in conversational artificial intelligence for individuals with depression: A critical analysis,” *JMIR Mental Health*, vol. 11, no. 1, p. e56569, 2024. [Online]. Available: https://mental.jmir.org/2024/1/e56569.

[51] D. Shin *et al.*, “Using large language models to detect depression from user-generated diary text data as a novel approach in digital mental health screening: Instrument validation study,” *J. Med. Internet Res.*, vol. 26, p. e54617, 2024. [Online]. Available: https://www.jmir.org/2024/1/e54617.

[52] T. Lai *et al.*, “Psy-LLM: Scaling up global mental health psychological services with AI-based large language models,” *arXiv preprint* arXiv:2307.11991, 2023. [Online]. Available: https://arxiv.org/abs/2307.11991.

[53] S. Nepal *et al.*, “Social isolation and serious mental illness: The role of context-aware mobile interventions,” *arXiv preprint* arXiv:2311.10302, 2023. [Online]. Available: https://arxiv.org/abs/2311.10302.

[54] A. M. Rahmani *et al.*, “Personal mental health navigator: Harnessing the power of data, personal models, and health cybernetics to promote psychological well-being,” *arXiv preprint* arXiv:2012.09131, 2020. [Online]. Available: https://arxiv.org/abs/2012.09131.

[55] K. Wijekoon Mudiyanselage *et al.*, “The effectiveness of mental health interventions involving non-specialists and digital technology in low-and middle-income countries–a systematic review,” *BMC Public Health*, vol. 24, no. 1, p. 77, 2024. [Online]. Available: https://bmcpublichealth.biomedcentral.com/articles/10.1186/s12889-023-17417-6.