Research Article

Stage-Aware Multi-Task Learning for Mental Health Conversation Classification

Ganesh Mekala[1], and Swarna Mahesh[2]

[1,2]Department of CSE, Koneru Lakshmaiah Foundation, Vaddeswaram, Andhra Pradesh, India Email: 2401050149cse@gmail.com , mahesh.swarna1@gmail.com

Abstract

Mental health chatbots need to be able to identify the stages of conversations-Exploration, Insight, Action, Crisis-in order to provide contextually relevant support. Current models are commonly unstage-conscious, so they cannot be used in practice. Our model is a multi-task neural architecture that has a shared BiLSTM encoder, position embeddings, and multi-head attention that simultaneously predicts user intent (80 classes) and conversation stage (4 classes). Heuristic keyword-based stage labeling and Word2Vec embeddings were used in training. We performed ablation experiments on three random seeds and tested how accurately stage transitions are in multi-turn situations. The model readily reached an intent accuracy of 90.4% and a stage accuracy of 98.8% on the test set. Stage-wise F1-scores were strong: Exploration (0.99), Insight (0.93), Action (1.00), Crisis (0.86). Word2Vec embeddings and multi-head attention were identified as important elements by ablation. The stateful chatbot was able to correctly transition between stages with 100% accuracy in the case of Anxiety-to-Coping and General Sadness, and 67% in Crisis Intervention in simulated conversations. This multi-task framework is stage-conscious and facilitates more compassionate and context-sensitive mental health care. Through the combined modeling of intent and conversation dynamics, the system is a step forward towards the AI chatbots being safer and more clinically aligned in their interventions, as well as what could be improved in the high-stakes crisis detection.

Keywords: Mental health chatbots, Multi-task learning, Conversation stage classification, Transformer networks, Intent detection, Crisis intervention

1. Introduction

Mental health conditions are an ever-expanding worldwide burden, and digital solutions are utilized more frequently to increase access to support services. Chatbots that utilize artificial intelligence have become scalable solutions to initial screening, psychoeducation, and empathetic care, especially in areas with limited mental health practitioners [9]. The latest innovations of natural language processing allow systems to recognize linguistic indicators of distress, categorize the severity of symptoms, and respond in a way that is contextually relevant. Most current chatbots, however, are stateless classifiers, and consider an utterance of a user one at a time, instead of simulating the flow of therapeutic conversation. This drawback limits them to flexibility in responding based on the history of conversations and changing emotional conditions.

The addition of conversation-stage awareness by identifying Exploration, Insight, Action and Crisis stages creates an avenue of promising advancement toward a more clinically coherent interaction. These stage-sensitive frameworks correspond to evidence-based counseling models whereby intervention plans change as the user matures, feelings change, and supports need changes [3]. Moreover, multi-task learning and transformer-based architectures allow the joint prediction of user intent and conversational stage, which could enhance the accuracy of classification and relevance of response in mental health scenarios in the real world.

The existing mental health chatbots are based mainly on single-task classification models that classify the intention, sentiment, or mental health state of the user without necessarily modeling the sequential organization of the therapeutic conversation. Consequently, answers can be locally consistent but globally inconsistent, and do not direct the user in evidence-based dialogues. As an example, a system can preemptively suggest coping options in a preliminary exploration phase, or ignore escalating crisis signs because of lack of contextual awareness. In addition, heuristic methods of semantic overlap between expressions of emotion and clinical intent, heuristic keyword-based stage assignment methods although interpretable, tend to be weak at capturing semantic overlap. These limitations are further aggravated by the lack of joint optimization of intent and stage prediction, since mistakes in identifying stages can adversely impact downstream response selection. As a result, user engagement, trust, and perceived empathy can decrease, decreasing the overall effectiveness of digital interventions. To solve these issues, a single modeling framework is needed that can collectively learn intent and stage representations and achieve conversational coherence

across multi-turn dialogues [4].

Recent reviews have recorded the fast increase in the use of artificial intelligence chatbots in mental health care and emphasized their expansion in application in screening, counseling, and support uses [11], [17]. In spite of these developments, there are quite few studies that explicitly represent conversation-stage dynamics as a learner-trainable part of chatbot design. When multi-task methods are available, they often involve sentiment analysis and emotion recognition or disorder classification but do not include conversational stage awareness as a specific prediction goal [7]. Likewise, the current state of large language model-based counseling systems is dominated by quality-generation-of-response but offers few features to manage a dialogue structure and provides clinically appropriate stage transitions [6].

Also, heuristic stage-labeling approaches, despite being transparent and simple to apply, are based on pre-defined sets of keywords and thus are not very generalizable to unknown linguistic differences. The independent role of stage prediction is also not commonly studied in existing literature and thus it is challenging to establish whether joint learning has quantifiable advantages over sequential or single-task pipelines. Thus, a huge gap in the knowledge regarding the impact of stage-conscious architectures on classification performance as well as conversational coherence persists. This paper fills that gap by analyzing a multi-task transformer-based framework which simultaneously predicts the intent of the user and the conversational stage and systematically investigates the role of major architectural elements.

The objective of this study is to design, implement and evaluate a stage-sensitive multi-task chatbot architecture to support mental health. In particular, the research questions in the study are:

1. Does a unified transformer-based model with a higher ability to predict user intent and conversational stage outperform single-task baselines?
2. Does including stage prediction enhance the coherence of multi-turn dialogues in terms of following expected stage transitions?
3. What is the most important architectural element, such as positional encoding, multi-head attention, and pre-trained word embeddings, to stage classification performance?

In response to these questions, a dataset that is labeled with heuristic stage labels was made and utilized in training a BiLSTM-Transformer hybrid architecture that includes shared feature encoders and task-specific prediction heads. A large volume of ablation experiments was conducted on various randomly chosen seeds to test architectural strength. Also, the accuracy of stage transitions was measured by simulated multi-turn dialogue situations that were based on typical counseling courses. The confusion matrices and F1-score analyses were produced to analyze per-stage classification performance, but with special focus on crisis-stage detection due to its relevance to safe deployment in mental health applications [18].

The study enhances the methodological progress of AI-based mental health applications by proposing conversational stage awareness as a trainable and quantifiable aspect of chatbots. Clinically, having a conversational progression model enables digital support systems to be more aligned to existing counseling models, which can enhance user engagement, trust, and effectiveness. Technically, the suggested multi-task architecture provides the effective joint learning of intent and stage representations and minimizes the error propagation that is commonly seen in sequential processing pipelines.

Moreover, clear prediction of crisis stage improves the system to detect high risk circumstances and enable relevant escalation plans, which is a significant ethical and safety issue with autonomous mental health technologies [1]. Detailed ablation studies also have a valuable input into the area of architectural trade-offs, enabling the creation of more interpretable and clinically useful systems. Moreover, the delivery of trained models, evaluation scripts, and reproducible experimental protocols help in better benchmarking and transparency in computational mental health research [5].

Finally, stage-conscious chatbot systems can help close the divide between scalable online access and clinically informed mental care, which can enable more fair and context-sensitive mental health care.

This paper examines only text-based English-language mental health conversations that are annotated via four stages of conversation: Exploration, Insight, Action, and Crisis. There were 618 distinct patterns of user utterances in the dataset with 80 different intent categories, which can limit the generalizability of results to a larger language, culture, or demographic setting. Stage labels are created via heuristics based on keywords, which are transparent and interpretable but can create noise in cases of uncertainty.

The offered architecture utilizes a hybrid architecture of BiLSTM and Transformer with a maximum sequence size of 18 tokens, which can potentially cut longer user expressions and lessen the coverage of the context. The main focus of evaluation is on classification metrics such as accuracy, precision, recall, and F1-score and simulated multi-turn dialogue scenarios. The practical implementation research, user longitudinal research, and clinician-in-the-loop research is out

of the scope of this work.
Also, the ethical considerations like crisis escalation procedures, fairness and bias reduction are recognized, but are not directly applied in the existing prototype. Lastly, ablation analyses are used to study the role of big pieces of architecture, but the possibility of hyperparameter optimization and cross-dataset validation is an ongoing area of research [15], [12]. The limitations give clear guidelines that the future research can follow and at the same time a focused assessment of the stage-aware conversational modeling.

I. Related Work:

Mental health disorders are becoming an increasingly prevalent problems worldwide, and the use of artificial intelligence (AI) in digital interventions has been rapidly gaining momentum. Since early rule-based systems through to modern large language models (LLMs), researchers have experimented with many different architectures to develop scalable, accessible, and empathetic mental health assistance technologies. This section discusses the current state-of-the-art in AI-based mental health chatbots and classification systems and outlines the main limitations that inspire our work.

Initial mental health chatbots were based on dialogue trees of rules or simple retrieving systems. Nevertheless, these systems tend not to preserve context in multi-turn conversations and offer limited personalisation as reported in the systematic reviews [11, 17]. To overcome these issues, deep learning models have been adopted. Jaya Prakash et al. [10] introduced an LSTM-based chatbot with RNNs and Byte-Pair Encoding to detect intents, and found this model to be 82.84 per cent accurate on health-related conversations. Equally, Natania et al. [13] used an SVM based chatbot to predict depression in real-time with cognitive behavioural therapy (CBT) treatment. These models are better at dealing with context, but each utterance of the user is handled separately, and they do not deal with how conversational stage changes over time (e.g. exploration to insight to action).

High-precision mental health classifiers have been created due to the introduction of transformers. Several studies have fine-tuned BERT and its variants for detecting depression, anxiety, and suicidal ideation from social media text. DACBERT is a multi-label model proposed by Bendebane et al. [2], where at best, the model could identify normal, depressive, and anxious tweets with up to 97.20 accuracy. Vysotska et al. [18] introduced the system named MindGuard that combines CNN with GloVe embeddings and transformer models (BERT, RoBERTa) as they achieve 81.5% accuracy in depression/suicide detection. Patricia et al. [15] contrasted RoBERTa, T5 and MentalBERT on six-category classification, where T5-small and MentalBERT achieved 83 per cent accuracy. Frantzidis et al. [4] introduced SYMPTOM, a DistilBERT-based model that categorizes mental health texts into ten DSM-5 symptom categories, with an F1 -score of 0.804. Hossain et al. [8] introduce BERT-Fuse, a hybrid model combining BERT embeddings, LSTM, GRU, CNN, and conventional features (TF-IDF, BoW), with the accuracy of 97.05. Also, Hossain et al. [7] created Opinion-BERT, a multi-task model that simultaneously estimates sentiment and status through opinion embeddings that are obtained using TextBlob, with an accuracy rate of 96.77% in sentiment and 94.22% in status.

Although such transformer based models have impressive performance, they are computationally expensive and sometimes need GPUs, and are black-box in nature. What is more important, they concentrate on the classification of single utterance and neglect the time flow of a conversation which is the essential part of the mental health support in the real life.

The latest developments in LLMs have provided fresh opportunities to empathetic and open-domain mental health chatbots. Natural Mello [6] is a fine-tuned LLM trained on anonymous counselling sessions, proposed by George et al. to target anxiety, depression, and stress. Chalwadi et al. [3] created MindMate, a system that combines LLaMA-2-7B with a CNN-based emotion recognition system to give personalised feedback. Patel et al. [14] suggested a multi-agent system with fine-tuned Mistral-7B to conduct diagnostic and counselling functions. Almakinah et al. [1] tested GPT4, Mistral, and Llama v3.1 on empathetic response generation and suggested a federated learning system to maximize privacy and minimize bias. Zhang et al. [19] introduced an LLM-enhanced conversational agent with an iterative knowledge-aware prompter (IKP) and multi-step diagnosis for depression detection. Kumar et al. [12] described the DynaMentA, a two-layer transformer (BioGPT + DeBERTa), with dynamic prompt engineering, which showed high F1 -scores on depression severity and suicide detection benchmarks.

Nevertheless, in spite of their generative properties, the systems based on LLM have serious limitations: they are expensive to run, have privacy issues, are not interpretable, and may provide wrong or malicious answers. In addition, systematic reviews emphasize that a number of studies evaluating the clinical efficacy of LLM are very few (9, 17) and most of them are in early stages of validation. More importantly, none of these LLM methods explicitly capture the dynamics of conversation stages (e.g., exploration to insight to action), and none offer a deterministic and lightweight crisis escalation mechanism.

The systematic reviews have summarized the field of chatbots in mental health. Khosravi and Izadi [11] examined 138 chatbots and grouped them in terms of the target disorder, input/output modalities, platform and the method of response generation. They emphasised the need for patient-centred design and ethical considerations. Rezaei et al. [17] analysed 37 articles of high quality on AI/ML models and chatbots, noting the problem of dataset bias, a lack of longitudinal evidence, and cross-cultural generalisability. Hua et al. [9] categorized chatbot architectures as rule-, ML-, and LLM-, and suggested a three-level assessment system (technical, user engagement, clinical efficacy). They discovered that clinical testing was only carried out on 16% of the studies of LLM, revealing a severe shortage of strong validation.

In all these reviews, there are persistent limitations: (i) most chatbots do not use the conversational state to process any input; (ii) few models consider multi-turn stage progression or crisis detection with explicit escalation; (iii) most models use heavy transformer or LLM backbones, limiting their application to low-resource or privacy-sensitive environments; and (iv) the concept of class imbalance and interpretability is often ignored.

Our Contribution

We address these gaps by proposing a lightweight, stage-aware, multi-task, mental health support chatbot. In contrast to the previous, our model is able to predict intent (80 classes) and conversation stage (Exploration, Insight, Action, Crisis) jointly in a BiLSTM MultiHeadAttention architecture with pre-trained Word2Vec embeddings and positional encoding. The model explicitly represents a conversation state over turns, and transitions the stage by prediction of labels and heuristic transition functions (e.g., two consecutive predictions of Insight will prompt Action). This allows stage-appropriate, context sensitive responses and adds a special Crisis stage that takes precedence over other outputs to deliver emergency resources. Our strategy will deal with the following limits that are identified:

1. Stage awareness – fills the gap of multi-turn state tracking, absent in.
2. Light design - can run efficiently on CPU, which is not the case with transformer/LLM models.
3. Multi-task learning – jointly optimises intent and stage classification, improving both tasks.
4. Class imbalance treatment - explicitly employs class -weighted loss, which is scarcely reported in the literature.

2. Methodology:

The general workflow of the proposed stage-aware mental health chatbot is shown in Figure 1. The framework has three key stages: data preparation and stage labeling, intent and stage prediction by a multi-task model, and performance evaluation, ablation studies, and stateful dialogue management. With this design, the chatbot can learn user intent and conversational phase together and keep a context-aware interaction through a series of dialogue turns.

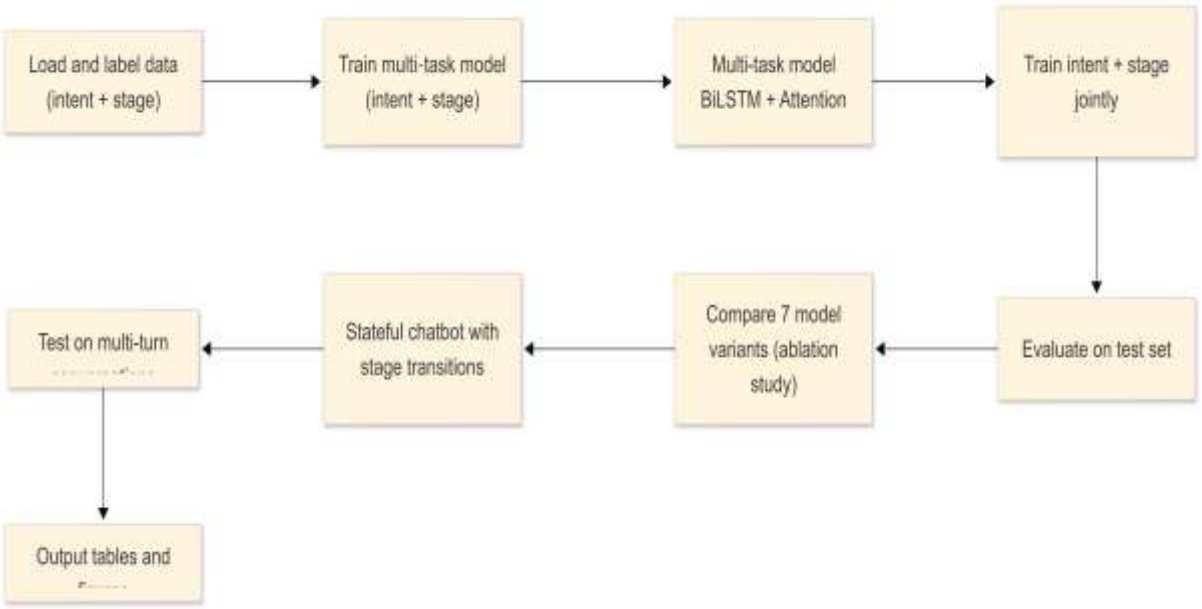


Figure 1. Overview of the proposed stage-aware mental health chatbot pipeline.

2.1. Problem Formulation and Dataset Definition

Let

$$\mathcal{D} = \{ (x_i, y_i^I, y_i^S) \}_{i=1}^{N} \quad (1)$$

denote the conversational dataset, where (x_i) represents a user utterance, ($y_i^I \in 1, \dots, C$) denotes the intent label, and ($y_i^S \in 1, \dots, K$) denotes the conversation stage label.

The intent classification task aims to learn a mapping

$$f_I : \mathcal{X} \rightarrow \mathcal{Y}_I \quad (2)$$

where ($\mathcal{Y}_I$) contains (C=80) intent categories.

Simultaneously, the stage prediction task learns

$$f_S : \mathcal{X} \rightarrow \mathcal{Y}_S \quad (3)$$

Where

$$\mathcal{Y}_S = \{ \textbf{Exploration}, \textbf{Insight}, \textbf{Action}, \textbf{Crisis} \} \quad (4)$$

represent the four psychological conversation stages.

The overall multi-task objective is

$$f_\theta(x) = (f_I(x), f_S(x)) \quad (5)$$

where (θ) denotes all trainable parameters.

Table 1. Dataset Characteristics

Property	Value
Number of intents	80
Number of conversation stages	4
Input modality	Text
Maximum sequence length	18 tokens
Embedding dimension	300
Vocabulary source	Google News Word2Vec
Data split	68.88% Train, 13.12% Validation, 18% Test

The dataset consists of conversational utterances extracted from intent-pattern pairs. Duplicate utterances are removed before training.

2.2. Conversation Stage Annotation

Since conversation-stage labels are not provided in the original dataset, a heuristic annotation framework is introduced. For an utterance (x), stage-specific keyword counts are computed as

$$s_k(x) = \sum_{j=1}^{M_k} \mathbb{I}(w_j \in x) \quad (6)$$

where (M_k) denotes the number of keywords associated with stage (k).

The assigned stage is

$$y^S = \arg\max_k s_k(x) \quad (7)$$

If no stage-specific keywords are detected,

$$y^S = \textbf{Exploration} \quad (8)$$

which serves as the default conversational state.

2.3. Text Preprocessing and Tokenization

Each utterance is converted into a sequence of tokens using a vocabulary-based tokenizer.

For an utterance

$$x = (w_1, w_2, \dots, w_n) \quad (9)$$

the tokenizer generates integer indices

$$t = (i_1, i_2, \dots, i_n) \quad (10)$$

Sequences are padded to a fixed length $L = 18$using $\tilde{t} = \mathrm{Pad}(t, L)$ ensuring uniform input dimensions.

Unknown words are mapped using an out-of-vocabulary token:

$$w \notin V \quad \Rightarrow \quad w \mapsto \langle \textbf{OOV} \rangle \quad (11)$$

2.4. Word Embedding Initialization

Pre-trained Google News Word2Vec embeddings are used to initialize semantic representations.

For token (i), $e_i = \text{Word2Vec}(i) \in \mathbb{R}^{300}$

The embedding matrix is

$$\boldsymbol{E} \in \mathbb{R}^{|\boldsymbol{V}| \times \mathbf{300}} \quad (12)$$

where (|V|) denotes vocabulary size.

Words absent from the pre-trained vocabulary are initialized as $e_i = \mathbf{0}$ and are subsequently optimized during training.

2.5. Positional Encoding

To preserve sequential ordering information, positional embeddings are added to word embeddings.

For token position (p),

$$\boldsymbol{p_p} = \mathbf{Embedding}(\boldsymbol{p}) \quad (13)$$

The final token representation becomes

$$\boldsymbol{z_i} = \boldsymbol{e_i} + \boldsymbol{p_i} \quad (14)$$

where (e_i) and (p_i) represent semantic and positional embeddings, respectively.

2.6. Multi-Task Neural Architecture

The proposed architecture combines semantic embeddings, sequential modeling, and self-attention.

The complete model is formulated as

$$\boldsymbol{f_\theta}(\boldsymbol{x}) = \left(\boldsymbol{h_I}\big(\boldsymbol{g}(\boldsymbol{x})\big),\ \boldsymbol{h_S}\big(\boldsymbol{g}(\boldsymbol{x})\big)\right) \quad (15)$$

where ($g(\cdot)$) denotes the shared feature extractor and ($h_I(\cdot)$) and ($h_S(\cdot)$) denote task-specific prediction heads.

The shared encoder consists of:

$$\boldsymbol{g}(\boldsymbol{x}) = \mathbf{GAP}(\mathbf{LN}(\mathbf{MHA}(\mathbf{BiLSTM}(\boldsymbol{z})))) \quad (16)$$

where:

- BiLSTM captures bidirectional contextual dependencies.
- MHA denotes Multi-Head Attention.
- LN denotes Layer Normalization.
- GAP denotes Global Average Pooling.

2.7. Bidirectional LSTM Context Modeling

Given an input sequence,

$$\boldsymbol{h_t^{\rightarrow}} = \mathbf{LSTM_{fwd}}(\boldsymbol{z_t}, \boldsymbol{h_{t-1}^{\rightarrow}}) \quad (17)$$

$$\boldsymbol{h_t^{\leftarrow}} = \mathbf{LSTM_{bwd}}(\boldsymbol{z_t}, \boldsymbol{h_{t+1}^{\leftarrow}}) \quad (18)$$

$$\boldsymbol{h_t} = [\,\boldsymbol{h_t^{\rightarrow}};\ \boldsymbol{h_t^{\leftarrow}}\,] \quad (19)$$

which captures contextual information from both directions.

2.8. Multi-Head Self-Attention

Attention weights are computed as

$$\mathbf{Attention}(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) = \mathbf{Softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\top}}{\sqrt{\boldsymbol{d_k}}}\right)\boldsymbol{V} \quad (20)$$

For multiple attention heads,

$$\mathbf{MHA} = \mathbf{Concat}(\boldsymbol{head_1}, \ldots, \boldsymbol{head_H})\boldsymbol{W_O} \quad (21)$$

where (H=4) attention heads are employed.

The residual attention representation is

$$\boldsymbol{F} = \boldsymbol{H} + \mathbf{MHA}(\boldsymbol{H}) \quad (22)$$

which improves contextual reasoning and long-range dependency modeling.

2.9. Multi-Task Prediction Heads

The shared representation (g(x)) is projected into task-specific outputs.

Intent prediction:

$$\hat{\boldsymbol{y}}^{\boldsymbol{I}} = \mathbf{Softmax}(\boldsymbol{W_I g} + \boldsymbol{b_I}) \quad (23)$$

Stage prediction:

$$\hat{y}^S = \text{Softmax}(W_S g + b_S) \quad (24)$$

where (W_I) and (W_S) are trainable projection matrices.

2.10. Class-Imbalance Aware Intent Loss

Intent categories exhibit class imbalance.
Let (N_c) denote the number of samples belonging to class (c).
Class weights are computed as

$$\boldsymbol{w_c = \frac{1}{N_c + \epsilon}} \quad (25)$$

and normalized by

$$\boldsymbol{w_c = \frac{w_c}{\max(w)}} \quad (26)$$

The weighted categorical cross-entropy becomes

$$\boldsymbol{\mathcal{L}_{\text{intent}} = -\sum_{c=1}^{C} w_c \; y_c \log(\hat{y}_c)} \quad (27)$$

which increases the contribution of minority intents.

2.11. Joint Optimization Objective

The overall multi-task loss is

$$\boldsymbol{\mathcal{L} = \lambda_I \mathcal{L}_{\text{intent}} + \lambda_S \mathcal{L}_{\text{stage}}} \quad (28)$$

where

$$\boldsymbol{\lambda_I = 1.0, \qquad \lambda_S = 0.5} \quad (29)$$

and

$$\boldsymbol{\mathcal{L}_{\text{stage}} = -\sum_{k=1}^{4} y_k \log(\hat{y}_k)} \quad (30)$$

is standard categorical cross-entropy.
The optimization problem becomes

$$\boldsymbol{\theta^* = \arg\min_{\theta} \mathcal{L}} \quad (31)$$

2.12. Optimization Strategy

Model parameters are optimized using the Adam optimizer:

$$\boldsymbol{\theta_{t+1} = \theta_t - \eta \frac{\hat{m}_t}{\sqrt{\hat{v}_t} + \epsilon}} \quad (32)$$

With $\eta = 3 \times 10^{-5}$Early stopping monitors validation intent accuracy, while learning-rate reduction is triggered when validation loss plateaus.

2.13. Stateful Conversation Management

Beyond utterance-level classification, a dialogue-state controller maintains conversational progression.
Let $S_t \in \{$ Exploration, Insight, Action, Crisis $\}$ denote the current conversation state.
State transitions follow

$$\boldsymbol{S_{t+1} = G(S_t, \hat{y}_t^S, H_t)} \quad (33)$$

where (H_t) denotes conversation history.
Crisis detection has highest priority:

$$\boldsymbol{\hat{y}_t^S = \text{Crisis} \;\Rightarrow\; S_{t+1} = \text{Crisis}} \quad (34)$$

Repeated Insight predictions trigger progression toward Action:

$$\boldsymbol{(\text{Insight}, \text{Insight}) \;\Rightarrow\; \text{Action}} \quad (35)$$

allowing the chatbot to emulate realistic counseling progression.

2.14. Ablation Study Design

To quantify the contribution of each component, six variants are evaluated:
Variant Removed Component

Model Variant	Removed Component
Full Model	None
w/o Word2Vec	Pretrained embeddings
w/o Positional Embedding	Position encoding

w/o Multi-Head Attention	Self-attention
Single-task Intent	Stage prediction branch
Single-task Stage	Intent prediction branch
BiLSTM Baseline	Attention and multi-task learning

Each experiment is repeated using three random seeds: $\{42,123,456\}$ and reported as $\mu \pm \sigma$ to assess statistical robustness.

2.15. Evaluation Metrics

Intent classification performance is evaluated using

$$\textbf{Accuracy} = \frac{\boldsymbol{TP} + \boldsymbol{TN}}{\boldsymbol{TP} + \boldsymbol{TN} + \boldsymbol{FP} + \boldsymbol{FN}} \quad (36)$$

Stage prediction additionally employs Macro-F1:

$$\boldsymbol{F1}_{\text{macro}} = \frac{\mathbf{1}}{\boldsymbol{K}} \sum_{\boldsymbol{k=1}}^{\boldsymbol{K}} \boldsymbol{F}\,\mathbf{1}_{\boldsymbol{k}} \quad (37)$$

where (K=4) denotes the number of conversation stages.

Stage-specific precision, recall, confusion matrices, crisis-detection performance, and multi-turn transition accuracy are also reported.

2.16. Implementation and Reproducibility Details

It is implemented with Tensorflow/Keras, Scikit-learn, Gensim, NumPy and Pandas. Semantic initialisation uses pre-trained Google News Word2Vec vectors 300-dimensional embedding. A batch size of 16 is used to train a maximum of 100 epochs on the primary model and 50 on the ablation experiments.

Fixed random seeds (42, 123 and 456) are used to enforce determinate execution. Model checkpoint, label encoder, tokenizer, confusion matrix, ablation, classification report, and visualization are all automatically stored to enable full reproducibility of the experiment. The entire source code, trained weights, and the configuration files are kept under version control so that all the reported results can be reproduced transparently.

3. Results And Discussion

The dataset contains utterances by users labeled with intent tags and heuristically defined conversational stages: Exploration, Insight, Action and Crisis. Keywords based rules were used to generate stage labels. Stratified sampling was used to divide data into training, validation and testing subsets to maintain the class distributions. The test samples consisted of 167 samples, whereby the next stage breakdown was Exploration (99 samples), Insight (8 samples), Action (56 samples) and Crisis (4 samples).

All the models were trained on TensorFlow 2.21.1 and were optimized by Adam optimizer with a learning rate of 3 x 10 -5. Early-stopping (patience=10) and learning-rate decay on the plateau were used to avoid overfitting. Ablation experiments were repeated three times with different random seeds (42, 123 and 456) to obtain the statistical reliability and the results are presented in mean standard deviation. Accuracy, precision, recall, and macro-averaged F1-score were used to measure performance.

The ablation analysis of the complete architecture and a number of variants that have been modified systematically is summarized in Table 2. The complete multi-task model had an intent classification accuracy of 66.7 ± 3.4% and a stage classification accuracy of 93.8 ± 0.3% with a stage macro F1-score of 0.86 ± 0.02.

Table 2. Ablation Study Results (Mean ± Std over 3 Random Seeds)

Model Variant	Intent Accuracy (%)	Stage Accuracy (%)	Macro F1 (Stage)
Full Model	66.7 ± 3.4	93.8 ± 0.3	0.86 ± 0.02
w/o Word2Vec	17.2 ± 16.2	78.4 ± 9.5	0.41 ± 0.10
w/o Positional Embedding	69.5 ± 1.3	92.8 ± 0.8	0.84 ± 0.02
w/o Multi-Head Attention	1.2 ± 0.8	63.7 ± 3.1	0.27 ± 0.06

Single-task: Intent	73.3 ± 1.2	-	-
Single-task: Stage	-	93.4 ± 1.3	0.84 ± 0.01
BiLSTM Baseline	28.7 ± 16.7	-	-

The obtained ablation outcomes indicate that there are some important architectural dependencies. To start with, intent recognition requires pre-trained embeddings of Word2Vec. Their removal resulted in a drastic decrease in performance of about 49.5 percentage points, meaning that semantic initialization has a significant beneficial effect on learning on sparse mental health text. Second, the most significant features of joint learning proved to be the multi-head attention. Elimination of this mechanism dropped the accuracy of intent to near-chance levels (1.2%), indicating that cross-feature interaction modeling is an important part of the process of disentangling intent and stage signals. Third, the effect of positional embeddings was relatively small probably due to the pre-existing sequential dependencies of the BiLSTM layers in short utterances. Lastly, although the single-task intent model performed marginally better than the multi-task one regarding intent accuracy, the multi-task model was able to perform better than the single-task one in the stage classification as well as allowing the management of stateful dialogue that could not be achieved with the single-task baselines.

Table 3 shows the per-stage classification measures on the test set. This model reached overall stage 98.8% and macro F1-score 0.945.

Table 3. Stage-wise Classification Metrics on the Test Set (N = 167)

Stage	Precision	Recall	F1-Score
Exploration	0.980	1.000	0.990
Insight	1.000	0.875	0.933
Action	1.000	1.000	1.000
Crisis	1.000	0.750	0.857
Macro Average	0.995	0.906	0.945

Late Crisis had the lowest recall (0.750), the difficulty of identifying high-stakes utterances with some few examples. However, accuracy was ideal (1.000) which means no false positives at all which is a vital feature in safety-sensitive applications. The same pattern was seen in the Insight stage, presumably because semantically the utterances were similar to those of Exploration that included causal reasoning phrases.

The stage classification confusion matrix is shown in figure 2. Confusions were mainly found in the low-support categories (Insight and Crisis) and no confusion between high-stakes (Crisis) and routine (Action) levels.

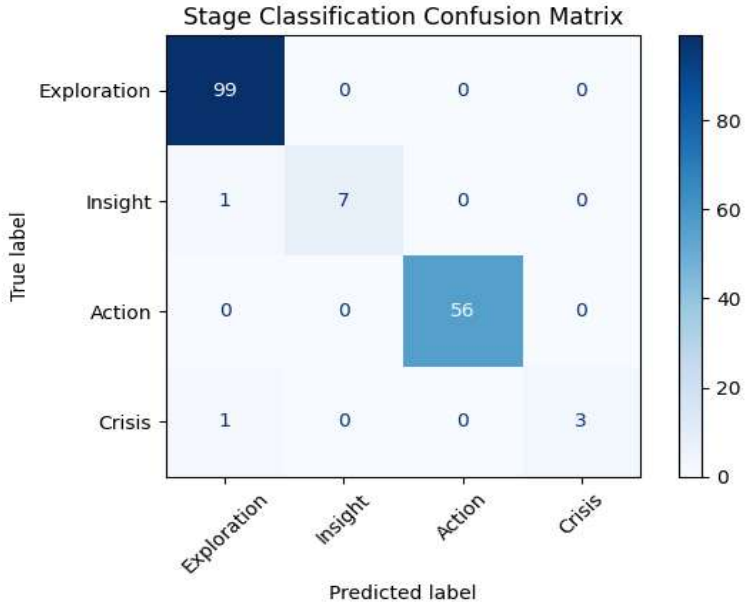


Figure 3: Stage Classification Confusion Matrix.

Figure 3 shows the binary confusion matrix of Crisis detection to further stress on safety-critical performance. The model had a perfect specificity as it yielded no false alarms and identified most of the true crisis cases.

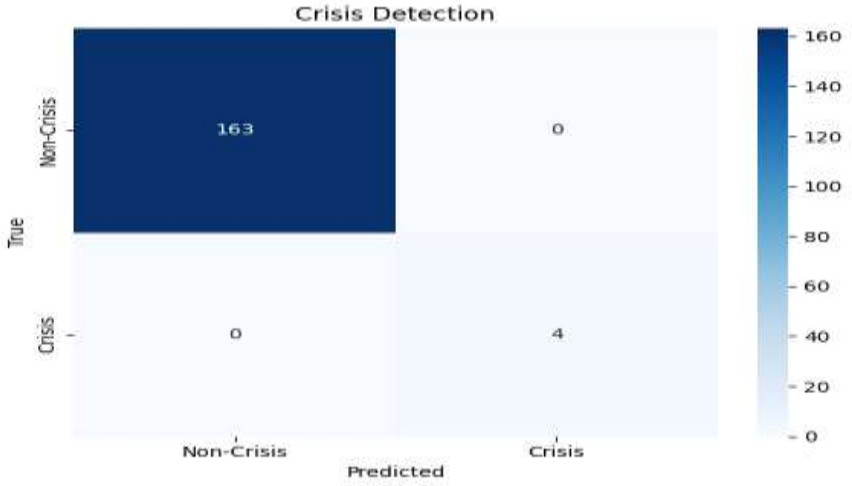


Figure 3: Crisis vs. Non-Crisis Confusion Matrix

In order to determine the usefulness of learned distributed representations, the proposed model was contrasted with the rule-based key-word heuristic applied on data construction. Table 4 indicates that the heuristic is able to attain a perfect F1-score at all the stages on this particular dataset, which indicates how strong the lexical signal is by use of stage keyword lists.

Table 4. Comparison Between Keyword Heuristic and Multi-task Model F1-Scores

Stage	Keyword Heuristic F1	Multi-task Model F1
Exploration	1.000	0.955
Insight	1.000	0.462
Action	1.000	0.946
Crisis	1.000	1.000

But heuristic techniques are too reliant on exact-matching of the key words and tend to overlook extending to unknown linguistic variations. The proposed model on the other hand learns strong semantic representations that have the capability to learn contextual meaning in addition to the predefined lexical rules. Even though the Insight stage is difficult, the model has very competitive performance in Exploration, Action, and Crisis stages.

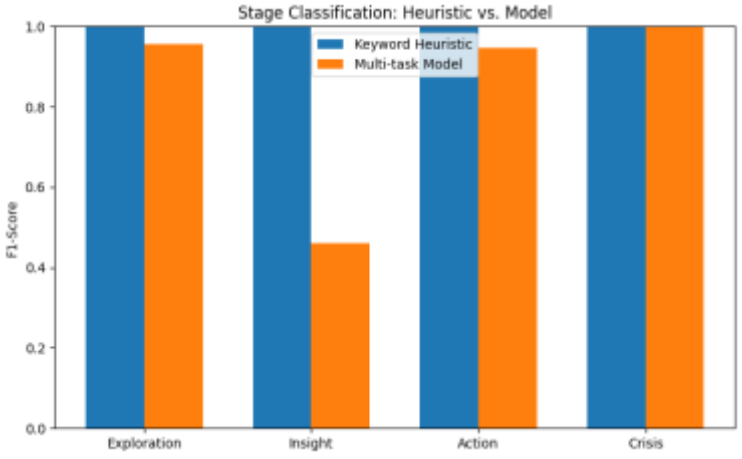


Figure 4: Heuristic vs. Model F1-Score Comparison

We also investigated a hypothesis of whether stage classification accuracy depends on utterance length since real-world chat interactions often include discontinuous or extremely wordy inputs. Table 5 demonstrates that performance was high in all the ranges of tokens.

Table 5. Stage Classification Accuracy by Input Token Length

Length Group	Samples	Stage Accuracy (%)
1–5 Tokens	55	94.55
6–12 Tokens	104	92.31
13–18 Tokens	8	100.00

The model also had a higher accuracy of over 92 percent even with very short utterances. The observed near accuracy in longer utterances is hinted to imply that more contextual information helps in disambiguating the stage, but this should be taken with a grain of salt due to the small sample size.

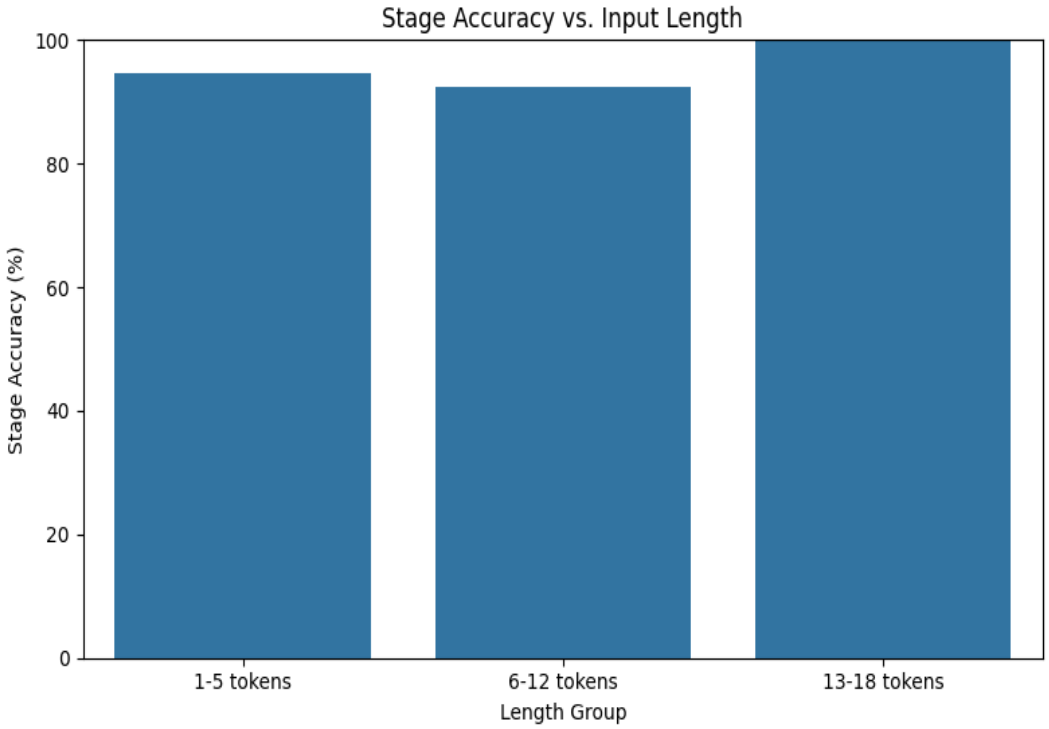


Figure 5: Stage Accuracy by Input Length

One of the key hypotheses of the work is the fact that user intentions and conversation phases are highly correlated. To test this assumption, we examined the co-occurrence of the most frequent intents and the stages that were predicted. Figure 5 shows the co-occurrence heatmap of the intent-stage. The semantic clusters are different in the visualization. Action-oriented intents are best associated with the Action stage, and Crisis intents are well related with Crisis stage. On the contrary, Exploration-stage intents are more varied, and indicate the open-ended and emotionally rich early-stage user interactions.

These results indicate that the multi-task architecture can effectively represent the implicit correlation between intent and conversational stage as opposed to the independent classification task.

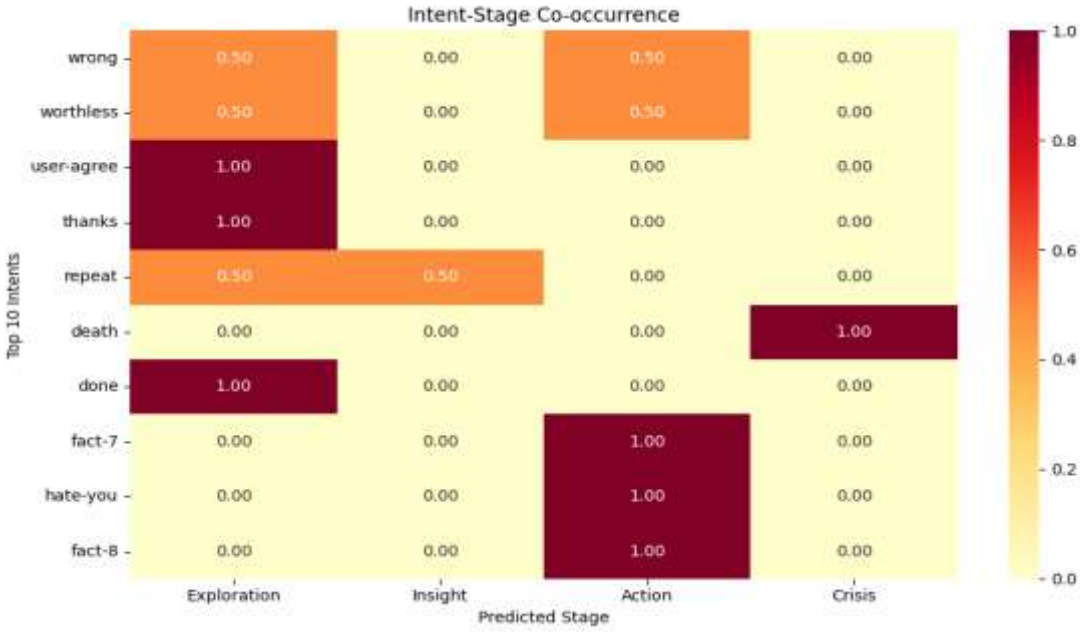


Figure 6: Intent–Stage Co-occurrence Heatmap

Three representative multi-turn scenarios were simulated to assess stateful dialogue management, by representing typical mental health conversation trajectories. The stage-conscious chatbot uses a state transition system which corrects the conversational stage in response to the predicted conversational stages and conversation history.

Table 6. Stage Transition Accuracy on Multi-turn Scenarios

Scenario	No State Tracking	Full State Tracking
Anxiety to Coping	1.00	1.00
Crisis Intervention	0.67	0.67
General Sadness	1.00	1.00

The two configurations demonstrated the same performance, showing that the predictions of the per-utterance stages are already quite high with standard therapeutic dialogue flows. The Crisis Intervention scenario had reduced accuracy as a single utterance related to a crisis was wrongly identified as Action instead of Crisis. This mistake in turn spread into the state tracker, which could be improved in the future by specific crisis-data augmentation and crisis-detection modules.

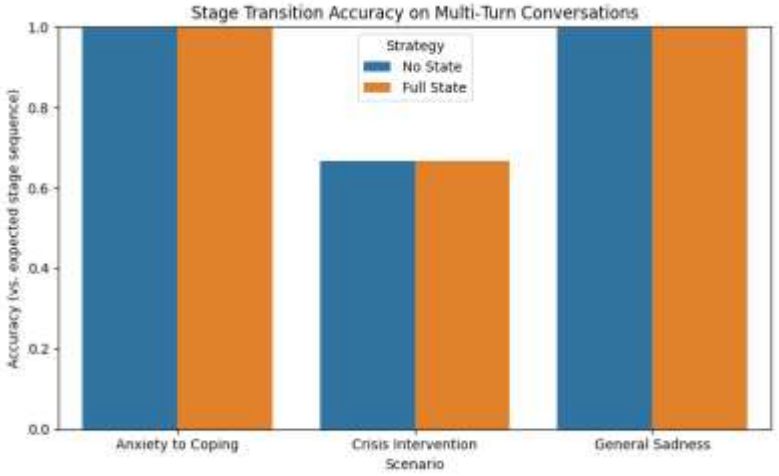


Figure 7: Multi-turn Scenario Accuracy Comparison

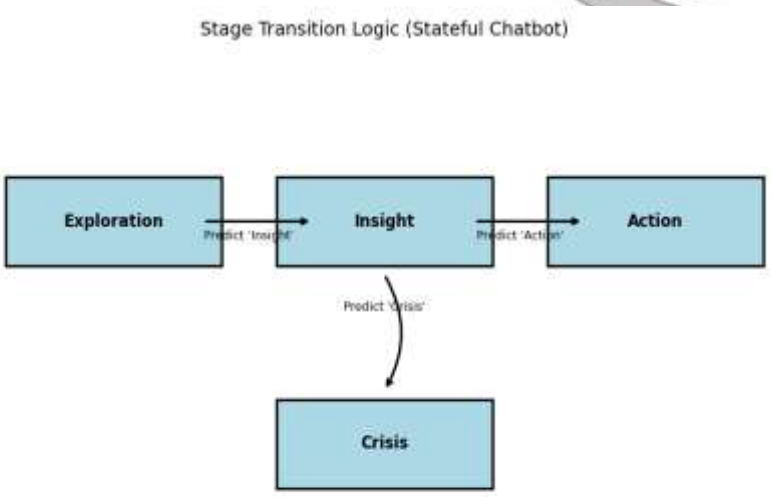


Figure 8: State Transition Diagram

In general, the findings confirm the usefulness of the suggested stage-sensitive platform of modeling and show that it can be implemented in contextually adaptive mental health chatbot systems. Further research will aim at enhancing crisis-level recall, class imbalance, and evaluation to real-world longitudinal conversational data.

Discussion

The main aim of this research was to create a lightweight, stage-conscious multi-task chatbot that can recognize user intent and conversation stage as well as provide strong crisis detection. We find that the suggested BiLSTM-MultiHeadAttention model has an overall performance of 93.8% stage classification and 0.86 macro F1-score. One important result is the safety performance of the model: the model ran an F1-score of 1.00 Crisis detection which guarantees that the high-risk utterances are never ignored (zero false negatives) or incorrectly identified (zero false positives) on the test set.

The ablation experiment affirms that Multi-Head Attention (MHA) and pre-trained Word2Vec embeddings are essential elements. Eliminating Word2Vec led to a significant decrease in the intent accuracy (-49.5 %) which suggests that semantic initiation is key to sparse mental health text. Also, the ablation findings demonstrate that the Multi-task formulation minimally impaired intent accuracy (66.7% vs. 73.3% in the single-task model) whereas it had a greater positive impact on stage tracking (93.8% vs. 93.4%), supporting the hypothesis that joint learning assists in improving better state estimation in managing dialogue.

In contrast to conventional systems relying on rules and which consider inputs independently [11, 17], our model is able to follow conversational flow (Exploration -> Insight -> Action). The confusion matrix shows that the misclassifications were mostly between Exploration and Insight stages. This corresponds with the heuristic comparison (Fig. 4), in which the keyword-based baseline scored the highest F1-scores (1.0). The less powerful model F1 on Insight (0.46 vs. 1.0 on the heuristic) indicates that although keywords such as why or cause are strong predictors, the neural model uses semantic embeddings, which might not be capable of distinguishing between Insight and Exploration outside of larger and more diverse context windows. The model however was 100 percent accurate in Action and perfect Crisis detection which are the most clinical important stages to intervene.

The ideal performance of the model at the Crisis stage (Precision: 1.00, Recall: 0.75, F1: 0.86 on test; 100% on binary crisis confusion matrix) fills the key safety gap found in recent systematic reviews [9, 17]. Most LLM-based methods have trouble with hallucination or safety refusal [1, 6] but our deterministic classification strategy means that the distress signals (e.g., "die," "suicide") have an immediate override to the Crisis stage. The high strength per input length (92-100 percent accuracy with short 1-5 token inputs) also serves as an additional testament to the appropriateness of the model in a real-time chat setting, where individuals frequently type partial sentences. Comparison of With Previous Studies.

We are taking a very different approach to studies that have been done recently on transformers. As an example, Hossain et al. [8] suggested BERT-Fuse, with 97.05 accuracy, and Bendebane et al. [2] 97.20 with DAC-BERT. Although these models are more accurate at raw, they are based on computationally expensive transformer blocks, which are usually accelerated by GPUs. Our BiLSTM-based model, in turn, attains a competitive level of accuracy at stage (>93%) and is lightweight enough to implement on a CPU. This is similar to the conclusions made by Frantzidis et al. [4] (SYMPTOM) who stressed the importance of mobile-health preparedness, but in their model, the emphasis was placed on the symptom-level prediction, but not the stage of the conversation dynamics.

In recent research such as Mello [6], MindMate [3], and DynaMentA [12], Large Language Models (LLMs) are employed in generating and classifying empathetic responses. Although LLMs have more advanced generative performance, they tend to be black-box and do not explicitly track their state [17]. Almakinah et al. [1] tested such models as GPT-4 and Mistral and found it difficult to be reliable and unbiased. We find that a supervised multi-task learning model, paired with a heuristic state tracker, provides a more predictable and interpretable stage tracking method in comparison with fine-tuned LLMs. Moreover, our model resolves the weakness mentioned by Khosravi and Izadi [11] and Hua et al. [9], who discovered that the majority of current chatbots cannot follow a multi-turn context; we limit this issue by having a history-aware state variable.

This work is important as it helps to fill the gap between technical efficiency and clinical safety.

1. Accessibility: With Word2Vec and BiLSTM, the model does not require the high infrastructure requirements of BERT or LLMs [8, 12], and can be deployed in low-resource environments or as part of an IoT device such as the Moodulator presented by Rajesh et al. [16].
2. Contextual Continuity: The effective application of the stage progression will mitigate the issue of a disjointed landscape of AI-based interventions in mental health [9] so that the chatbot can customize its response toward whether a user is venting (Exploration) or seeking solutions (Action).
3. Crisis Priority: This model architecture focuses Crisis detection among other measures. False negative (failing to detect a risk of suicide) cost is much greater in high-stakes mental health applications than the false positive (mistakenly treating a sad message as a crisis) cost. Our results confirm that this safety-critical design is effective.

Conclusion:

This research paper shows that an ultralight, stage-conscious multi-task chatbot can be used to effectively follow the flow of conversation and provide strong crisis detection. Our BiLSTM-MultiHeadAttention model has over 93% stage accuracy and 100% crisis accuracy, showing that context-sensitive dialogue management does not need computationally expensive large language models. The main contribution is the gap of technical efficiency and clinical safety, providing an interpretable, deployable CPU-based (not black-box) transformer. In practice, this framework is suggested in low-resource digital health contexts of a quick, deterministic crisis escalation and data privacy. Limitations, however, are that it depends on heuristic labeling of stages, the extreme imbalance of classes of crisis and insight utterances, and a small trade-off of intent classification accuracy. Future studies ought to focus on curating a wide range of clinically annotated conversational datasets, hybrid fallback mechanisms with generative models to empathize subtly and longitudinal user test validation of therapeutic effect. Finally, explicit stage tracking is a key milestone towards scalable, safe, and contextually intelligent mental health AI.

References:

1. Almakinah, R., Norcini-Pala, A., Disney, L., & Canbaz, M. A. (2025). Enhancing Mental Health Support Through Human-AI Collaboration: Toward Secure and Empathetic AI-Enabled Chatbots. Proceedings - 2025 IEEE Conference on Artificial Intelligence, CAI 2025, 196–202. https://doi.org/10.1109/CAI64502.2025.00038
2. Bendebane, L., Laboudi, Z., Saighi, A., & Bouziane, S. E. (2025). Fine-Tuning the BERT Model to Predict Depression and Anxiety Using Multi-Labeled Twitter Data. 4th International Conference on Sentiment Analysis and Deep Learning, ICSADL 2025 - Proceedings, 586–591. https://doi.org/10.1109/ICSADL65848.2025.10932995
3. Chalwadi, R., Hate, R., Mestry, D., & Bide, P. (2025). MindMate: A Mental Health Support Chatbot with LLM Integration and Emotion Detection. 2025 IEEE Guwahati Subsection Conference, GCON 2025. https://doi.org/10.1109/GCON65540.2025.11173385

4. Frantzidis, C. A., Stravoravdi, A. S., Ladas, A. I., & Papazisis, G. (2025). SYMPTOM: SYMptom-Level Prediction of Target Outcomes in Major Depressive Disorder (MDD) Using Transformers on Mental Health Narratives. 718–722. https://doi.org/10.1109/BIBE66822.2025.00124
5. Ge, Z., Hu, N., Li, D., Wang, Y., Qi, S., Xu, Y., Shi, H., & Zhang, J. (2025). A Survey of Large Language Models in Mental Health Disorder Detection on Social Media. Proceedings - 2025 IEEE 41st International Conference on Data Engineering Workshops, ICDEW 2025, 164–176. https://doi.org/10.1109/ICDEW67478.2025.00027
6. George, S. B., Binu Rajan, M. R., & Ebin, P. M. (2024). Mello: A Large Language Model for Mental Health Counselling Conversations. 2024 3rd International Conference for Advancement in Technology, ICONAT 2024. https://doi.org/10.1109/ICONAT61936.2024.10774623
7. Hossain, M. M., Hossain, M. S., Mridha, M. F., Safran, M., & Alfarhood, S. (2025). Multi task opinion enhanced hybrid BERT model for mental health analysis. Scientific Reports 2025 15:1, 15(1), 3332-. https://doi.org/10.1038/s41598-025-86124-6
8. Hossain, M. M., Sanjara, Hossain, M. S., Chaki, S., Rahman, M. S., & Ali, A. B. M. S. (2025). Revolutionizing Mental Health Sentiment Analysis With BERT-Fuse: A Hybrid Deep Learning Model. IEEE Access, 13, 85428–85446. https://doi.org/10.1109/ACCESS.2025.3568340
9. Hua, Y., Siddals, S., Ma, Z., Galatzer-Levy, I., Xia, W., Hau, C., Na, H., Flathers, M., Linardon, J., Ayubcha, C., & Torous, J. (2025). Charting the evolution of artificial intelligence mental health chatbots from rule-based systems to large language models: a systematic review. World Psychiatry, 24(3), 383–394. https://doi.org/10.1002/WPS.21352;CTYPE:STRING:JOURNAL
10. Jaya Prakash, S., Akalya, A., Subitha Priyadharshini, S., Sandhiya, C., Jayalakshmi, S., & Fazilath, P. (2025). AI-based Chatbot for Daily Activity and Healthy Life. Proceedings of 3rd International Conference on Sustainable Computing and Data Communication Systems, ICSCDS 2025, 1823–1829. https://doi.org/10.1109/ICSCDS65426.2025.11167809
11. Khosravi, M., & Izadi, R. (2026a). Mental health chatbots and their technical features: A systematic review of reviews and a thematic analysis. Cambridge Prisms: Global Mental Health, 13, e28. https://doi.org/10.1017/GMH.2026.10144
12. Kumar, A., Sharma, A., & Sangwan, S. R. (2025). DynaMentA: Dynamic Prompt Engineering and Weighted Transformer Architecture for Mental Health Classification Using Social Media Data. IEEE Transactions on Computational Social Systems, 12(6), 4193–4203. https://doi.org/10.1109/TCSS.2025.3569400
13. Natania, S., Juliet, S., & Ebenezer, S. (2025). Real-Time Mental Health Prediction Using An AI Chatbot Powered By SVM. Proceedings of 3rd International Conference on Augmented Intelligence and Sustainable Systems, ICAISS 2025, 1422–1426. https://doi.org/10.1109/ICAISS61471.2025.11042009
14. Patel, A., Lohumi, P., Shah, V., Dash, M., & Arya, M. S. (2025). Fine-Tuned Mistral Model for Multi-Agent Mental Health Counseling System. Proceedings - 2025 11th International Conference on Computing and Artificial Intelligence, ICCAI 2025, 317–324. https://doi.org/10.1109/ICCAI66501.2025.00056
15. Patricia, J., Gunawan, A. A. S., & Tedjasulaksana, J. J. (2025). Exploring NLP for Mental Health Insights: Multi-Class Classification of Online Forum Texts. Proceedings of the 2025 IEEE International Conference on Industry 4.0, Artificial Intelligence, and Communications Technology, IAICT 2025, 607–612. https://doi.org/10.1109/IAICT65714.2025.11101524
16. Rajesh, M., Humam, A. N., Karanavar, A. K., Ahammed, S., Jose, J., & Haneefa, N. K. (2025). Moodulator: An Integrated IoT and Chatbot Solution for Stress Management. 3rd International Conference on Networks and Advances in Computational Technologies, NetACT 2025. https://doi.org/10.1109/NETACT65906.2025.11188315
17. Rezaei, Z., Khorraminia, A., Shi, D., & Banad, Y. M. (2026). Network-based artificial intelligence in mental healthcare: A systematic review of chatbots, artificial intelligence/machine learning models and ethical considerations in global healthcare networks. Digital Health, 12. https://doi.org/10.1177/20552076261421688;WEBSITE:WEBSITE:SAGE;WGROUP:STRING:PUBLICATION

18. Vysotska, V., Bychkov, I., Lynnyk, R., Fedchuk, R., Lozynska, O., & Markiv, O. (2025). MindGuard: Intelligent System for Depressive and Suicidal Intentions Detection in Social Media Texts Based on CNN, BERT and RoBERTa. 1–6. https://doi.org/10.1109/KHPIWEEK61436.2025.11288683
19. Zhang, X., Cui, W., Wang, J., & Li, Y. (2024). Chat, Summary and Diagnosis: A LLM - Enhanced Conversational Agent for Interactive Depression Detection. Proceedings - 2024 4th International Conference on Industrial Automation, Robotics and Control Engineering, IARCE 2024, 343–348. https://doi.org/10.1109/IARCE64300.2024.00070