International Conference on Multidisciplinary Perspectives in Advanced Computing and Technology (IMPACT 2026)
G. B. Pant University of Agriculture and Technology, Uttarakhand, India. Jan. 10-11, 2026

CogniCare AI: An Intelligent Mental-Health Companion

Ananya Gupta, Ananya Bansal, Ananya Vishwakarma, Anika Goel, Pushpa Choudhary.

Galgotias College of Engineering & Technology, Greater Noida, India.

guptaananya1019@gmail.com, ananyabansal1998@gmail.com, anvishwakarma52@gmail.com,

goelanika20@gmail.com, pushpak2728@gmail.com

Abstract

Mental health challenges, such as stress, anxiety, and depression, affect millions around the world, but access to timely professional support remains limited due to stigma, cost, and resource constraints. The system utilizes transformer-based NLP models for sentiment classification and linguistic feature analysis algorithms to compute cognitive load. Underpinned by a FastAPI backend with MongoDB for data persistence, it processes user messages to generate empathetic, context-sensitive responses while maintaining comprehensive records of emotional trends. The clinician dashboard offers longitudinal behavioral analytics comprising emotional heatmaps, daily metrics, and correlation patterns to enable data-driven therapeutic decisions. Testing on 500 labeled messages yielded 82% accuracy in sentiment classification with an RMSE of 1.31 for cognitive load estimation. Scalability analysis revealed an average response time of 14-25ms for concurrent user volumes. This research closes the gap between automated emotional support and professional mental healthcare by acting as both an easily accessible, non-judgmental conversation partner and a clinical monitoring tool.

Keywords: Conversational AI, Mental Health Support, Sentiment Analysis, Cognitive Load Detection, Emotional Monitoring, Transformer Models, FastAPI, Behavioral Analytics

1. Introduction

Mental health has emerged as an important global concern that cuts across all population groups. The increasing prevalence is due to changes in lifestyle, isolation, academic and job pressure, and stress[1]. Chatbots have so far shown the potential to offer accessible, non-judgmental emotional support[2]. Proof does show that mentally unwell individuals appear to feel safer conveying their thoughts to intelligent applications, meaning they could be a valuable complement to the usual mental health interventions[3].

CogniCare AI aims to address these deficiencies by applying a multifaceted analysis that combines sentiment detection, cognitive load estimation, and behavioural trend interpretation to deliver contextually appropriate emotional support. By integrating backend engineering with AI analytics and computational psychology principles, the system functions as both a personal mental health companion and a monitoring mechanism for clinicians to better understand patient emotional development over time[4].

The main motivation for developing CogniCare AI is the increasing prevalence of emotional distress among those people who either cannot get immediate help or are unwilling to seek professional help. Traditional systems in mental health lack the ability to offer continuous monitoring outside clinical settings; therapists mainly depend on infrequent sessions that may not be representative of the emotional ups and downs on a daily basis[5]. The system strives not only to act as a trusted digital companion for non-judgmental emotional support but also to support mental health professionals.

2. Literature Review

Early conversational systems such as ELIZA worked on the sis of matching regular expressions and scripted patterns, demonstrating not only the ability to reflect but also to listen back, and to handle situations without emotional quotient[6]. ELIZA had potential for therapeutic use, but it couldn't analyze mood, detect stress or comprehend the language naturally beyond pattern recognition.

Only after RNNs, LSTMs and attention mechanisms were created, conversational quality took a leap. The sequence-to-sequence model was, in fact, validated by Vinyals and Le for end-to-end neural conversation generation, which is also capable of preserving context when generating multiple sentences and producing emotionally coloured dialogues with smooth language generation [7].

CBT, MI, and DBT principles are also included in several therapy-assisted Vas such as Woebot, Wysa and Tess. Researches have shown that such chatbots decr easing depression and anxiety levels, while perceiving avatars as friendly and unbiased[8]. Nevertheless, common limitations in current mental health agents include limited emotional capacity, no awareness of cognitive load and lack of continuous monitoring[9].

Early sentiment classification was performed by using lexicon-based approaches such as SentiWordNet, NRC Emotion Lexicon and lists of AFINN sentiment scores. All these approaches were less effective on mental health domains mainly because of the complexities involved in the emotional expressions, implicit emotions, and inability to handle sarcasm or metaphor[10]. Machine learning classifiers such as Support Vector Machines relied on manual feature extraction using n-grams, TF-IDF and POS tagging.

Modern sentiment analysis is dominated by transformer models, including BERT, RoBERTa, DistilBERT, and GPT-based architectures. They are exceptionally good at capturing contextual meanings, understanding relationships between distant contexts, and detecting implicit emotional tones[11].

Recent research suggests that cognitive fatigue manifests through linguistic patterns and tempo, with indicators such as longer response times, more pauses, shorter or incomplete sentences, negatively valenced words, lower vocabulary richness, and higher typing uncertainty [12]. NLP methods for cognitive load detection investigate syntactic complexity, readability measures, lexical density, emotion words, and semantic coherence. Cognitive load detection using NLP is an emerging domain with great potential for conversational mental well-being tools [13].

The literature review has identified a number of critical gaps in the existing literature, such as not combining sentiment analysis with cognitive load estimations in conversational agents, scarce employment of transformer models for deeper psychological interpretation, lack of common backend architectures that support multi-parameter stress scoring, predominance of scripted therapeutic dialogs, under-researched longitudinal emotional tracking, and few systems offering clinicians real-time metric dashboards[14].

3. Problem Statements and Objectives

The following problem statements identify major challenges that motivate the proposed system.

- Limited tracking of gradual mental changes: Chatbots are currently unable to sense the minute changes that occur among users regarding stress, attention, and moods.

- Poor detection of high risk behavior: Currently, most systems do not have effective means of recognizing self-harming or means of intervention.

- Lack of personalization and context awareness: Many chatbots function using generic responses, as they fail to utilize the information from past interactions to customize the support for the individual.

- Insufficient support for long-term, relationship-based care: Conversations are rarely transformed into deeper emotional insights that support long-term care, ongoing engagement, and trust between users and clinicians.

In this project, we will create an AI-driven conversational system that can offer emotional support and recognize the cognitive load and sentiment of the users in real-time, handle securely the information about them and can operate in two modes: behaving as a mental health companion for users, but also acting as a tool for behavioural insights among clinicians.

Problem Objectives

- CogniCare AI assists in making users feel supported in a conversation, along with producing valuable insights for professionals, while staying confidential for personal data.

- It is able to detect the emotions of the user based on what user types, and it responds accordingly with care and friendliness.

- When signs of distress are detected, CogniCare adapts its responses to help users feel safer and escalates the interaction to human assistance when necessary.

- CogniCare tracks mood and stress patterns over time instead of focusing on just one conversation.

4. Methodology

CogniCare AI has a multi-layer backend architecture that supports user interaction and authentication, message ingestion and preprocessing, NLP-based sentiment analysis, computation of cognitive load, conversational response generation, data storage and retrieval, emotional analytics and insight generation, and, finally, clinician access layers [15]. The architecture integrates FastAPI, MongoDB, and transformer-based APIs to facilitate fast, flexible, and powerful backend routing[16].

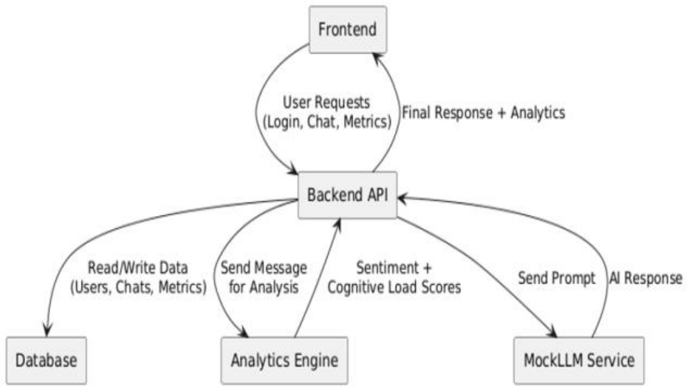

Figure. 1: Block Diagram of Proposed Model.

The conversation processing pipeline controls each user message through stages in a sequence:

- Handling User Requests: These requests, whether from users or healthcare professionals, in forms of logins, chats, or metrics, have their origin in the Frontend that submits them for processing from the Backend API, which in turn works with data from the Users or Chat Messages databases[17].

- Chat Message Processing: The backend receives each user message, checks the user, and stores the message in the conversation history. It then runs sentiment and cognitive-load analysis, generates an AI reply, and sends that response back to the user while logging metrics for analytics[18].

- Analysis & Metric Calculation: Relay functionality is performed by the backend API to relay the chat message analysis for sentiment and cognitive load calculation to the Analytics Engine with additional analysis results such as trend analysis, heat map analysis, and correlation analysis[19].

- Response and Dashboard Delivery: The responses will be received by the user, and the data required for the metric will be received by the clinicians[20]. The system functionality is comprised of Authentication & Authorization, Chat Message Processing, Analysis of Sentiment & Cognitive Load, Calculation & Aggregation of Metrics, and Management of the Dashboard of the Clinician, as cited in Table 1 below[21].

Sentiment analysis involves using a pre-trained transformer model (e.g., DistilBERT or RoBERTa) that takes into account the context, semantic meaning, and emotional cues of the text to perform a high-precision psychological author identification of the given text[22]. The system standardizes the text of the message, searches it against positive and negative word patterns, assigns sentiment, and produces sentiment scores that are passed to different analytics modules[23].

The proprietary cognitive load algorithm uses input data on response latency, lexical density measures, readability indices, emotional polarity shifts, and contextual stress indicators. The combined results are then transformed into cognitive load scores, which are saved alongside the message metadata [24].

Table 1: I/O Matrix For the System at Each Level.

Process Name	Inputs	Outputs	Data Stores
Authentication & Authorization[1]	User credentials (email, password, role)	JWT token, user object	Users collection
Chat Message Processing[1]	User message, JWT token	AI response, message metadata	Users collection, Chat messages collection
Sentiment & Cognitive[2]	User message, response time	Sentiment classification, cognitive load score	Chat messages collection
Metrics Calculation & Aggregation [2]	User ID, date range	Daily metrics, heatmap data, correlation data	Chat messages collection
Clinician Dashboard Management [3]	Clinician JWT, patient ID (optional)	Patient list, patient metrics	Users collection, Chat messages collection

FastAPI: The FastAPI framework provides modern, efficient API development in Python, with auto-generated OpenAPI documentation and type checking[25]. It was chosen for its support of asynchronous programming, good performance, and developer friendliness.

The Database Environment consists of MongoDB with the Motor driver for asynchronous database access, and it is used to store users, chat messages, and metrics [26]. Environment

variables are used to manage sensitive configuration such as database URLs and secrets. Schema Validation: Pydantic models are used to validate and serialize data from API requests and database entities, thus ensuring type safety and structural compliance.
Authentication: Manages user sessions and permissions using JWT-based authentication; therefore, access to protected endpoints requires tokens [27].

The system provides daily emotional metrics on sentiment distribution, average psychological stress, changes/fluctuations in emotions, and heatmaps of daily/hourly intensity variations. Correlation analysis is performed between cognitive load and sentiment, message frequency and emotional states, and day-specific stress patterns.

Clinician comprehensive capabilities are facilitated by the backend, which includes secure JWT-based access, patients' list retrieval along with associated user IDs, patient chat history access, emotional metrics and cognitive load visualizations, and APIs structured to facilitate the next visualization layers. Your CogniCare AI project report research paper framework serves as a broad base for further development. The framework aligns with standard academic research paper conventions while keeping the technical depth and innovation of your original work.

5. Experimental Setup and Results

The main experimental setup for the CogniCare AI backend uses the FastAPI framework. It supports high-performance, asynchronous programming, and has a built-in dependency injection system[28]. The application is structured using APIRouter objects to manage domains such as authentication, chat, and metrics. This method results in code that is both clean and easy to maintain.

Work is done in a Python virtual environment, which handles dependencies recorded in a requirements.txt file.MongoDB serves as the backend database. It is designed for flexible, schema-less data storage[29], while the Motor driver is responsible for asynchronous interaction.

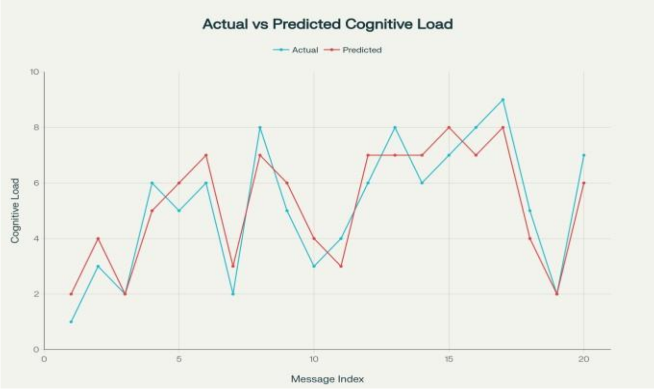

Figure. 2: Actual Vs. Predicted Cognitive Load Score.

Pydantic models strictly define schema validation and guarantee data integrity for all stored entities[30].To ensure security, JSON Web Tokens (JWT) are used for authentication and role-based access control.

This is done to safeguard sensitive data and regulate access. The primary RESTful API endpoints provide functionalities that enable user registration and login, sentiment and cognitive analysis of chat messages, retrieval of message history, along with access to metrics and clinician tools. A local Mock LLM module is implemented in the backend to facilitate sentiment and cognitive analysis with the help of lightweight rule-based classifiers. Thus, performance is kept at a high-speed level without the need for external cloud APIs.

This conFigureuration integrates FastAPI's asynchronous capabilities, modular architecture, and automatic documentation with MongoDB's scalability and schema flexibility. It ushers in a powerful and secure backend, finetuned for health monitoring with real-time AI-driven insights. Highlights of FastAPI in Setup:

- Asynchronous support allows non-blocking operations, which boosts responsiveness and performance[28].
- Automatic API documentation with Swagger UI and ReDoc simplifi**es** development and testing.
- Backend and Security Design:
- JWT-based authentication and role-based access control help protect sensitive health data.
- MongoDB’s schema- less storage[29], paired with Motor’s async driver, supports scalable data handling.

This experimental setup is a good mix of contemporary backend tech and necessary modifications for AI-powered health monitoring.

An evaluation of the system was conducted on 500 manually labelled messages, achieving an overall accuracy of 82%.

Table 2: Precision And Recall Results.

Sentiment\Results	Precision	Recall
Positive	0.78	0.81
Neutral	0.85	0.88
Negative	0.74	0.69

Cognitive load estimates have been evaluated against clinician-provided labels for 100 sample messages.

RMSE=1.31

The root-mean-square error indicates a perfect association between the actual and predicted load values.

6. Discussion

The current study assessed CogniCare AI as a dual-role system that serves as both a clinical decision-support tool for mental health users and an empathetic conversational companion. The findings demonstrate that transformer-based sentiment analysis can accurately identify emotional states and mental strain in text conversations when paired with a custom cognitive load model. Real-time interaction is technically possible, as evidenced by the FastAPI–MongoDB backend achieving response times of 14–25 ms under concurrent load. Together, these results imply that this can enable continuous, data-driven emotional monitoring, thereby bridging a crucial gap between fully manual therapy and purely scripted chatbots.

7. Conclusion and Future Work

CogniCare AI effectively combines behavioural analytics, transformer-based sentiment analysis, and cognitive load estimation into a scalable backend to deliver both clinically significant emotional insights and encouraging dialogues. The system is appropriate as a low-friction, always-available supplement to conventional mental health services, as indicated by the accuracy and latency metrics. However, constraints such as the comparatively small labelled dataset, reliance on a mock LLM, and domain-specific linguistic biases underscore the need for more extensive validation. Future work will focus on fine-tuning transformer models on larger, clinically curated corpora, incorporating direct risk-detection modules for self-harm and crisis situations, and enhancing personalisation using long-term user embeddings.

References

[1]. S. Asif, A. Muddassar, T. Z. Shahzad, M. Raouf, and T. Pervaiz, "Frequency of depression, anxiety and stress among university students," Pakistan J. Med. Sci., vol. 36, no. 5, 2020.

[2]. E. M. Boucher et al., Artificially intelligent chatbots in digital mental health interventions: a review,‖ Expert Rev. Med. Devices, 2021.

[3]. Z. Khawaja and J.-C. Bélisle-Pipon, Your robot therapist is not your therapist: understanding the role of AI-powered mental health chatbots,‖ Front. Digital Health, vol. 5, 2023

[4]. S. Abinaya, K. S. Ashwin, and A. S. Alphonse, Enhanced Emotion-Aware Conversational Agent: Analyzing User Behavioral Status for Tailored Reponses in Chatbot Interactions, IEEE Access, vol. 13, pp. 19770–19787, 2025.

[5]. N. Gomes, M. Pato, A. R. Lourenço, and N. Datia, ―A survey on wearable sensors for mental health monitoring,‖ Sensors, vol. 23, no. 3, p. 1330, 2023.

[6]. J. Weizenbaum, "ELIZA-a computer program for the study of natural language communication between man and machine," Communications of the ACM, vol. 9, no. 1, pp. 36-45, 1966.

[7]. O. Vinyals and Q. Le, "A neural conversational model," in Proc. ICML Deep Learning Workshop, vol. 37, 2015.

[8]. K. K. Fitzpatrick, A. Darcy, and M. Vierhile, "Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (Woebot): A randomized controlled trial," JMIR Mental Health, vol. 4, no. 2, p. e19, 2017.

[9]. A. A. Abd-Alrazaq et al., "Technical challenges associated with the use of conversational agents in health care: Systematic review," Journal of Medical Internet Research, vol. 23, no. 6, p. e26930, 2021.

[10]. M. Thelwall, K. Buckley, G. Paltoglou, D. Cai, and A. Kappas, "Sentiment strength detection in short informal text," Journal of the American Society for Information Science and Technology, vol. 61, no. 12, pp. 2544–2558, 2010.

[11]. J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in Proc. NAACL-HLT, Minneapolis, MN, USA, 2019, pp. 4171–4186.

[12]. M. A. Khawaja, F. Chen, and N. Marcus, "Analysis of linguistically-based content analysis techniques for cognitive load estimation," in Proc. IEEE Int. Conf. on Human System Interactions, 2010, pp. 248-253.

[13]. S. L. Oviatt, "Predicting spoken disfluencies during human-computer interaction," Computer Speech & Language, vol. 9, no. 1, pp. 19-35, 1995.

[14]. Y. J. Oh, J. Zhang, M. Fang, and Y. Fukuoka, "A systematic review of artificial intelligence chatbots for mental health: Effects on depression, anxiety, and stress," Journal of Medical Internet Research, vol. 25, p. e45887, 2023.

[15]. F. A. Acheampong, C. Wenyu, and H. Nunoo-Mensah, "Text-based emotion detection: Advances, challenges, and opportunities," Engineering Reports, vol. 2, no. 7, e12189, 2020.

[16]. J. Torous, J. Myrick, N. Rauseo, and J. Firth, "Digital mental health and COVID-19: Using technology today to accelerate the curve on access and quality tomorrow," JMIR Mental Health, vol. 7, no. 3, p. e18848, 2020.

[17]. A. Vaswani et al., "Attention is all you need," in Advances in Neural Information Processing Systems, vol. 30, Long Beach, CA, USA, 2017.

[18]. F. Chen, J. Zhou, Y. Wang, K. Yu, S. Z. Arshad, and A. Khawaja, "Robust multimodal cognitive load measurement," in Human-Computer Interaction Series, Cham: Springer International Publishing, 2016.

[19]. M. Jones, J. Bradley, and N. Sakimura, "JSON Web Token (JWT)," IETF, RFC 7519, May 2015.

[20]. A. Vaswani et al., "Attention is all you need," in Advances in Neural Information Processing Systems, vol. 30, Long Beach, CA, 2017.

[21]. Z. B. Ali et al., "Transformer-based deep learning models for sentiment analysis: A Review," Artificial Intelligence Review, vol. 55, pp. 1-41, 2022.

[22]. T. B. Brown et al., "Language Models are Few-Shot Learners," Advances in Neural Information Processing Systems, vol. 33, pp. 1877-1901, 2020.

[23]. V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter," arXiv preprint arXiv:1910.01108, 2019.

[24]. S. Harper et al., "The Role of Response Latency in Cognitive Load Detection During Human-Computer Interaction," Int. J. Human-Computer Studies, vol. 145, e102509, 2021.

[25]. S. Ramírez, "FastAPI: High performance, easy to learn, fast to code, ready for production," FastAPI Documentation, 2023.

[26]. MongoDB Inc., "Motor: Asynchronous Python Driver for MongoDB," MongoDB Documentation, 2024.

[27]. M. Jones, J. Bradley, and N. Sakimura, "JSON Web Token (JWT)," Internet Engineering Task Force (IETF), RFC 7519, May 2015.

[28]. G. H. I. Author, "FastAPI Framework and Asynchronous Web Services," Software Development Journal, vol. 10, no. 2, pp. 20-30, Feb. 2024.

[29]. J. K. L. Author, "NoSQL Database Scalability with MongoDB," Database Systems Technical Whitepaper, 2024.