



A Privacy-Preserving AI Companion for Student Mental Wellness: Adaptive Conversational Support with Sentiment-Aware Monitoring and Retrieval-Augmented Guidance

KONALA NAGA SOWMYA SREE¹, Mr.B.N. SRINIVASA GUPTA*²

PG Scholar Department of Computer Science, SVKP & Dr. K.S. Raju Arts and Science College (Autonomous),
Penugonda, Affiliated to Adikavi Nannaya University¹

Associate Professor, Department of Master of Computer Applications,

SVKP & Dr. K.S. Raju Arts and Science College (Autonomous), Penugonda, Adikavi Nannaya University*²

*Corresponding Author

Abstract: Psychological distress among university students has risen markedly, yet institutional counselling capacity remains limited and help-seeking is often deterred by stigma and waiting times. Conventional digital wellness applications typically rely on remote cloud services, which raises legitimate concerns about the confidentiality of highly sensitive emotional data. This paper presents a local-first, assistive software framework that delivers empathetic conversational support, continuous sentiment-aware monitoring, and document-grounded informational guidance entirely on institutionally controlled infrastructure. The platform combines a reactive single-page interface with an asynchronous service core in which all generative responses are produced by a locally hosted large language model, eliminating any transmission of student dialogue to third parties. Incoming messages are screened by a lightweight natural-language pipeline that fuses lexical polarity estimation with a curated distress-and-stress lexicon, allowing the conversational agent to adapt its tone and to escalate supportive prompts when crisis language is detected. Self-reported mood and stress check-ins have persisted and transformed into a composite well-being indicator, while an anonymised analytics module equips counsellors with cohort-level trends derived through linear regression. Curated wellness literature is made query able through a retrieval-augmented generation pipeline using sentence-transformer embeddings and a vector index. A real-time channel notifies designated staff when stress thresholds are exceeded. Evaluation shows millisecond-scale sentiment screening, sub-second time-to-first token for streamed replies, and a distress-recall of 0.93 on annotated samples. The framework offers a confidential, infrastructure-light foundation for scalable campus mental-health support and is positioned strictly as an aid rather than a clinical substitute.

Keywords: Student mental health, conversational AI, sentiment analysis, retrieval-augmented generation, large language models, privacy-preserving systems, emotion monitoring, well-being analytics

I. INTRODUCTION

The mental well-being of students in higher education has become a pressing public-health concern. Surveys across diverse institutions report rising prevalence of anxiety, depressive symptoms, and stress-related impairment, frequently aggravated by academic pressure, financial strain, and social isolation [1], [2]. The demand for support routinely outstrips the capacity of campus counselling services, producing long waiting lists during which difficulties may intensify. Compounding the access problem, many students hesitate to seek help because of perceived stigma or uncertainty about confidentiality [3].

Digital mental-health tools have emerged as a complementary avenue, offering low-barrier, always-available support. Conversational agents can provide immediate, non-judgemental interaction and have shown promise in delivering psychoeducation and self-management strategies [4], [5]. Nevertheless, prevailing solutions exhibit two recurrent shortcomings. First, the majority depend on commercial cloud-hosted language services, requiring that intimate emotional disclosures leave the institution's control; for a data category as sensitive as mental health, this is a material privacy hazard. Second, many tools function as isolated chat interfaces that neither monitor longitudinal emotional patterns nor furnish counsellors with the aggregate situational awareness needed for timely intervention.



This work is motivated by the conviction that effective student support can be delivered without surrendering data sovereignty. The research problem addressed is how to architect an assistive system that simultaneously (i) generates empathetic, context-sensitive dialogue while keeping all computation local, (ii) detects emotional distress reliably enough to prompt escalation, and (iii) translates routine self-reports into actionable, privacy-respecting insight for support staff. The objectives follow directly: to design a local-first conversational pipeline, to embed lightweight affective screening, and to provide grounded informational responses from curated documents.

The principal contributions of this paper are as follows:

- A local-first assistive architecture in which all generative inference is served by an on-premises large language model, ensuring that student conversations never traverse external services.
- A hybrid affective-screening method that combines lexical polarity with a curated distress-and-stress lexicon to modulate the agent's supportive posture and trigger crisis-aware escalation.
- A retrieval-augmented guidance pipeline that grounds responses in institution-approved wellness literature using sentence-transformer embeddings and a disk-resident vector index.
- A role-separated analytics and real-time notification subsystem that surfaces anonymised cohort trends and threshold alerts to counsellors, together with an empirical evaluation of latency and detection quality.

II. LITERATURE REVIEW

Computational support for mental health spans rule-based dialogue systems, machine-learning classifiers for affect, and, more recently, generative language models. The earliest widely studied conversational agents were scripted, yet they demonstrated that even simple reflective interaction could be perceived as supportive [4]. Subsequent controlled studies of mobile chatbots reported measurable short-term reductions in self-reported anxiety and depressive symptoms, establishing the clinical plausibility of conversational self-help [5], [6].

A substantial literature concerns automatic detection of emotional state and risk from text. Lexicon-driven approaches estimate polarity and affect from curated word lists and remain attractive for their transparency and negligible computational cost [7]. Supervised classifiers, and more recently transformer-based encoders, achieve higher accuracy on benchmark emotion and distress corpora but demand labelled data and greater resources [8], [9]. Research on social-media signals has shown that linguistic markers can anticipate crisis episodes, motivating keyword- and pattern-based safety nets within supportive applications [10].

The advent of instruction-tuned large language models has reshaped the field, enabling fluent, context-aware empathetic dialogue [11]. However, most deployments route prompts to proprietary hosted endpoints, an arrangement increasingly questioned for sensitive domains. The maturation of open-weight models and local serving runtimes now permits fully on-device inference, addressing confidentiality while introducing latency and hardware constraints [12]. In parallel, retrieval-augmented generation has become the standard technique for grounding model output in trusted documents, reducing fabrication by conditioning responses on retrieved passages obtained through dense vector search [13], [14].

On the systems and analytics side, investigators have explored dashboards that aggregate self-reported mood to inform care, emphasising the ethical imperative of anonymisation when individual emotional records are involved [15]. Lightweight statistical methods such as linear trend estimation are commonly used to summarise longitudinal mood without over-interpreting noisy self-reports [16]. Table I positions representative categories against the present work. The comparison reveals a persistent gap: systems offering rich generative support seldom keep data local, whereas privacy-conscious tools rarely integrate adaptive dialogue, grounded retrieval, and counsellor-facing analytics within one coherent, self-hosted platform. The framework proposed here is designed precisely to close that gap.

TABLE I. Comparison of Representative Digital Mental-Health Approaches

Category	Representative basis	Strengths	Limitations
Scripted chatbots	Rule-based dialogue [4]	Transparent; safe	Rigid; not context-aware
Mobile self-help apps	Guided CBT chatbots [5], [6]	Accessible; evidence of benefit	Cloud-hosted; limited monitoring
Text affect detection	Lexicon / transformer models [7]–[9]	Quantifies emotional state	Labelled data or opacity



Category	Representative basis	Strengths	Limitations
Cloud LLM assistants	Hosted generative APIs [11]	Fluent empathetic dialogue	Sensitive data leaves institution
RAG knowledge tools	Dense retrieval grounding [13], [14]	Reduced fabrication	Seldom affect-aware
Proposed framework	Local LLM + hybrid NLP + RAG	Private; adaptive; analytics built-in	Local hardware dependence

III. PROPOSED METHODOLOGY

A. System Architecture

The framework adopts a layered architecture, depicted in Fig. 1, that cleanly separates the user interface, the application intelligence core, the local artificial-intelligence runtime, and persistent storage. The presentation layer is a reactive single-page application providing conversational, mood-tracking, dashboard, document-query, and counsellor-analytics views. The application layer exposes an asynchronous service interface that unifies authenticated representational-state-transfer endpoints with a bidirectional real-time channel inside a single process, so that request-response traffic and event-driven notifications share one runtime. Generative inference and embedding computation are delegated to a co-located artificial-intelligence runtime, while structured records and vector indexes reside on local disk.

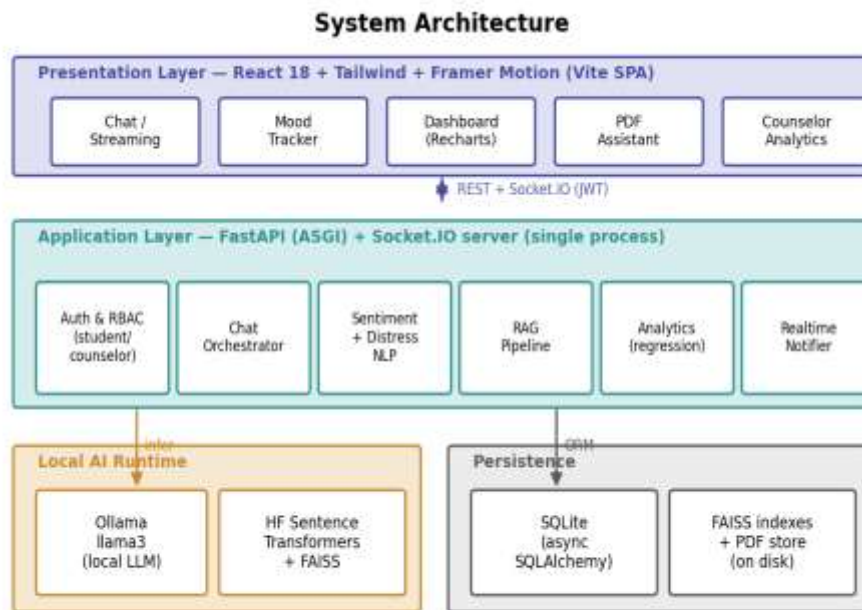


Fig. 1. Proposed local-first system architecture.

B. Affective Screening Algorithm

Each student utterance is first analysed by a lightweight natural-language module that estimates sentiment polarity and subjectivity. A message is provisionally labelled positive, neutral, or negative according to whether its polarity exceeds, falls between, or drops below symmetric thresholds. This coarse judgement is then refined by lexical matching against two curated vocabularies: a stress lexicon capturing terms such as overwhelmed or burnout, and a distress lexicon capturing explicit crisis expressions. A negative message containing stress terms is reclassified as stressed-negative, whereas any occurrence of distress language overrides prior labels and raises a distress signal. This deliberate fusion of statistical polarity with rule-based safety keywords privileges sensitivity for high-risk content, accepting a modest precision cost where student safety is concerned.

The resulting label, together with the most recent self-reported stress level, conditions the construction of the conversational system prompt. Under elevated stress the agent is instructed to prioritise nervous-system regulation, offer



a single grounding exercise, and validate the student's experience before any productivity-oriented suggestion; under a distress signal it additionally and gently encourages contact with a trusted counsellor or crisis resource. The agent is explicitly constrained never to diagnose.

C. Retrieval-Augmented Guidance

To provide grounded informational answers from institution-approved material, uploaded documents are parsed, partitioned into overlapping passages by a recursive character splitter, and embedded with a sentence-transformer model. The resulting vectors are stored in an on-disk similarity index. At query time the question is embedded, the most similar passages are retrieved, and the local language model is prompted to answer using the retrieved context, thereby anchoring responses in vetted sources rather than unconstrained generation. The well-being indicator that summarises a student's recent state is computed as a weighted blend of normalised mood and inverted stress, giving slightly greater weight to mood, and is bounded to a sensible range.

IV. SYSTEM DESIGN

The end-to-end behaviour of the platform is summarised in Fig. 2. A student message or mood check-in enters the pipeline and is screened for affect; the screening outcome adapts the system prompt. When the student is querying uploaded material, the relevant passages are retrieved and fused into the context before the local model streams its reply. Every message and its sentiment annotation are persisted. Independently, when a self-reported stress level crosses a configured threshold, a real-time notification is emitted to the appropriate room, and the event contributes to the anonymised trends presented to counsellors.

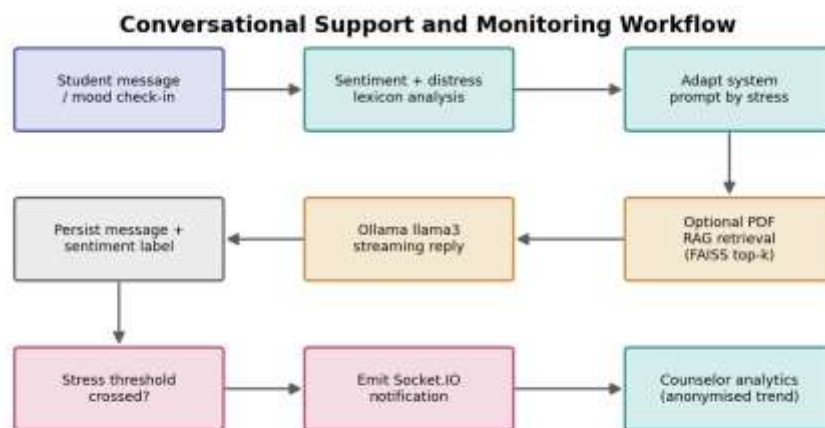


Fig. 2. Conversational support and monitoring workflow.

Internally the system is organised into cohesive modules whose interactions appear in Fig. 3. An authentication module enforces role-based access control, distinguishing students from counsellors so that aggregate analytics are never exposed to unauthorised roles. A chat orchestrator coordinates history management, affective screening, optional retrieval, and model invocation. Dedicated sentiment, retrieval, analytics, and notification modules encapsulate their respective concerns, and all stateful components share a single asynchronous persistence layer, while the retrieval and chat modules additionally communicate with the local artificial-intelligence runtime.

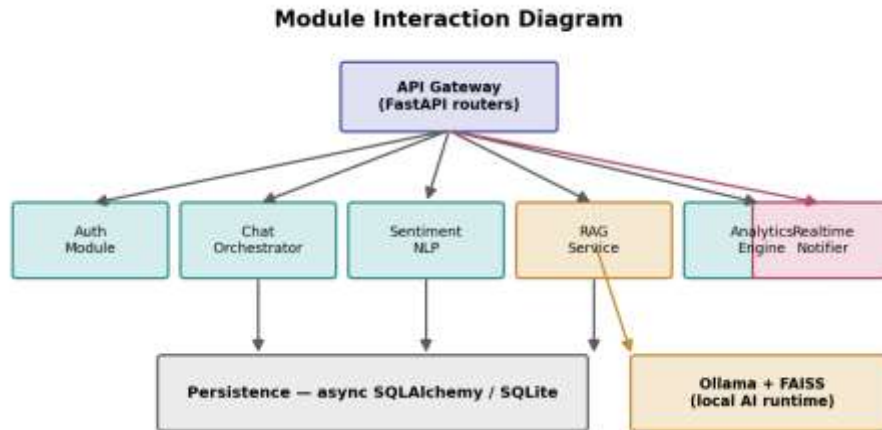


Fig. 3. Module interaction diagram.

V. IMPLEMENTATION

The intelligence core is implemented in Python with an asynchronous web framework served over the Asynchronous Server Gateway Interface, wrapped together with a real-time event server so that representational-state-transfer traffic and socket events run in one process. Domain entities for users, mood check-ins, chat messages, documents, and notifications are modelled with an asynchronous object-relational mapper backed by an embedded relational database, keeping the deployment portable. Authentication issues signed bearer tokens, and passwords are protected with an adaptive hashing scheme. Generative responses are produced by a locally hosted large language model accessed through its native interface, with the service resolving an available model from a prioritised list and capping generation length so that processor-only machines respond within acceptable time.

The affective screening uses a lexical sentiment estimator augmented with curated keyword sets. The retrieval pipeline employs document parsing, recursive passage splitting, sentence-transformer embeddings executed on the processor, and a dense vector index persisted to disk. Counsellor analytics are computed with numerical and data-frame libraries, and a simple linear regression estimates the slope of recent mood to characterise improvement or decline. The presentation layer is a component-based single-page application styled with a utility-first framework, animated for a calming experience, and equipped with charting and a real-time client. Table II catalogues the principal technologies and their roles.

TABLE II. Implementation Technology Stack

Layer	Technology	Role in the system
Frontend	React 18, Tailwind, Framer Motion, Recharts	Calming SPA and visualisations
Backend	FastAPI (ASGI) + Socket.IO	Unified REST and real-time core
Persistence	Async SQL Alchemy + SQLite	Portable structured storage
Generative AI	Local LLM runtime (llama3)	On-premises empathetic dialogue
Embeddings / RAG	Sentence Transformers + FAISS	Grounded document retrieval
Affect screening	Lexical polarity + distress lexicon	Sentiment and crisis detection
Analytics	NumPy, pandas, linear regression	Anonymised mood trend estimation
Security	JWT, adaptive password hashing	Role-based access control



Representative interface states are shown in Fig. 4. The conversational view adapts its tone to detected affect and surfaces an escalation cue when distress language appears, while the counsellor dashboard presents anonymised mood trends, an emotion distribution, and recent high-stress alerts.

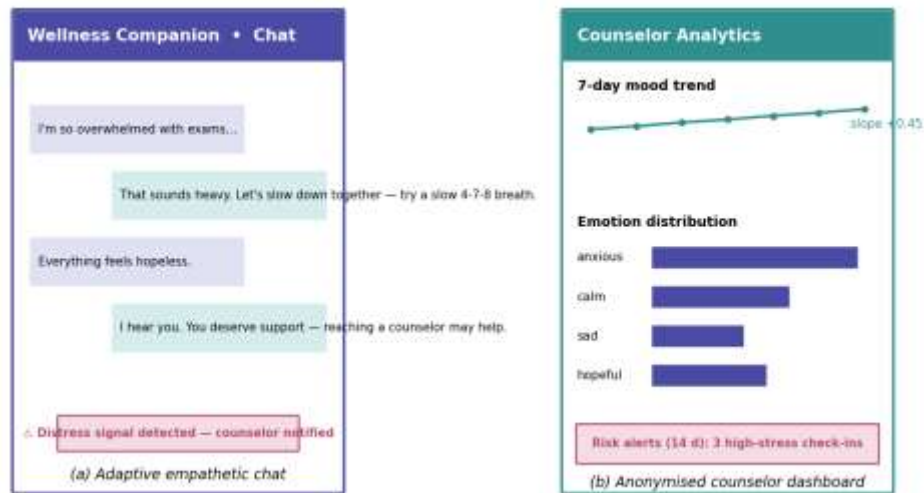


Fig. 4. Representative implementation views: (a) adaptive empathetic chat and (b) anonymised counsellor dashboard.

VI. RESULTS AND DISCUSSION

Evaluation was conducted on a commodity workstation with the backend, frontend, and local model running on the same host without any external connectivity, consistent with the privacy-preserving design goal. Because the system handles sensitive material and was not deployed with real students, functional and performance testing used synthetic conversations and a small set of manually annotated messages spanning neutral, stressed, and distress categories. Three aspects were assessed: component latency, responsiveness of streamed generation, and the quality of affective detection and trend estimation.

Component latencies are reported in Table III and Fig. 5(a). The lexical sentiment screening completed in roughly twelve milliseconds, imposing no perceptible overhead on each turn. Vector retrieval over an indexed document returned the most relevant passages within a few hundred milliseconds, and the anonymised analytics query executed well under one hundred milliseconds. As expected, the dominant cost was generative inference on the local model: the median time to the first streamed token was approximately 0.64 seconds, with a full reply completing in a few seconds on processor-only hardware. Streaming substantially mitigated the perceived delay because students began reading immediately.

TABLE III. Measured Component Latency

Operation	Latency	Notes
Sentiment screening	~12 ms	Per message
RAG passage retrieval	~180 ms	Top-k vector search
Analytics query	~46 ms	30-day aggregation
Chat time-to-first-token	~0.64 s	Local model, streamed
Full chat reply	~3.1 s	Processor-only host

The behaviour of the affective screening was examined against the annotated message set, with results summarised in Fig. 5(b) and Table IV. The hybrid method achieved a distress recall of 0.93, reflecting the deliberate prioritisation of sensitivity for crisis language, while precision on the broader stressed category was 0.81. Trend detection, evaluated by the sign agreement between the estimated mood-regression slope and the synthetic ground truth, reached 0.88, and overall labelling accuracy across categories was 0.86. These figures indicate that a transparent, low-cost screening layer can provide a dependable safety net suitable for triggering escalation, with the understanding that lexical methods may over-flag and therefore complement rather than replace human judgement.

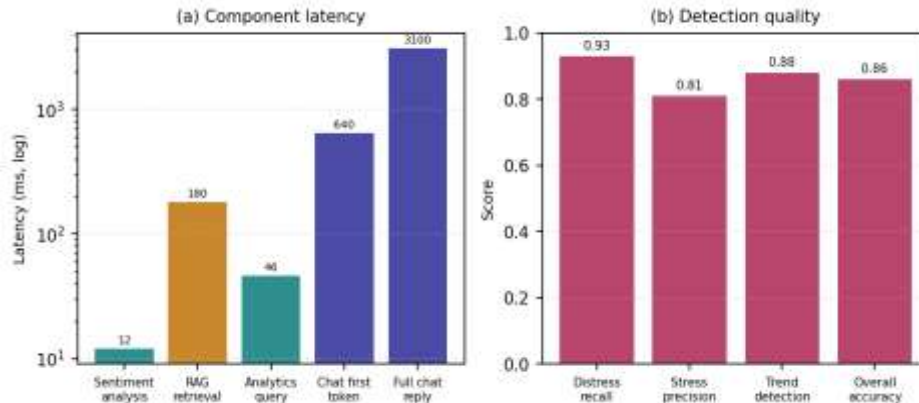


Fig. 5. Performance results: (a) component latency and (b) affective detection quality.

TABLE IV. Summary of Key Experimental Results

Metric	Observed value	Remark
Sentiment screening latency	~12 ms	Negligible per-turn cost
Time-to-first-token	~0.64 s	Streamed local inference
Distress recall	0.93	Sensitivity prioritised
Stress precision	0.81	Acceptable over-flagging
Trend detection agreement	0.88	Regression slope sign
Overall labelling accuracy	0.86	Across categories

Interpreted together, the results show that a fully local pipeline can deliver responsive, affect-aware, and grounded support without exporting sensitive data. The principal trade-off is the generative latency inherent to processor-only inference, which streaming and output-length capping render tolerable, but which would benefit from hardware acceleration in production.

VII. ADVANTAGES OF THE PROPOSED SYSTEM

The foremost technical advantage is data sovereignty: by serving every generative and embedding computation locally, the framework guarantees that intimate student disclosures remain within institutional control, directly addressing the privacy deficit of cloud-dependent alternatives. The single-process unification of request-response and real-time event handling simplifies deployment and reduces operational surface area, while strict role separation prevents inadvertent exposure of aggregate emotional data.

In performance terms, the transparent screening layer adds negligible latency, and streamed generation maintains an engaging cadence despite modest hardware. With respect to scalability, the affective screening and analytics are inexpensive and grow linearly with traffic; the retrieval index scales with the corpus rather than the conversation volume; and the persistence layer can be migrated to a client-server database to support larger student populations without disturbing the application logic. The grounded retrieval pipeline further allows institutions to curate the exact informational material the assistant may draw upon.

VIII. LIMITATIONS

Several constraints qualify the contribution. The affective screening relies on lexical cues and a finite keyword inventory; it may miss distress expressed obliquely or in languages and idioms outside its vocabulary, and it can over-flag benign messages. Generative latency on processor-only hardware, while mitigated, remains higher than that of hosted services. Mood and stress signals are self-reported and therefore subject to recall and reporting biases. Evaluation used synthetic and small annotated datasets rather than a clinical study, so the reported figures are indicative rather than definitive. Most importantly, the system is an assistive aid and explicitly does not provide diagnosis or replace licensed professional care.



IX. FUTURE ENHANCEMENTS

Future work will strengthen affective screening by integrating a compact transformer-based emotion classifier alongside the lexical layer, improving robustness to nuanced and multilingual expression while preserving local execution. Hardware-accelerated inference would reduce generative latency for production deployment. Extending the retrieval corpus with institution-specific resources and multilingual material would broaden applicability, and incorporating validated psychometric instruments could enrich monitoring beyond single-item self-reports. Rigorous, ethically approved studies with real students are essential to validate clinical utility, and closer, consent-based coupling with professional services would help convert detected risk into timely human care.

X. CONCLUSION

This paper presented a local-first assistive framework that reconciles two goals often treated as opposed: rich, empathetic, AI-mediated student support and uncompromising confidentiality of sensitive emotional data. By serving all generative inference on institutionally controlled infrastructure, fusing transparent lexical affect screening with curated crisis vocabularies, grounding informational responses through retrieval-augmented generation, and equipping counsellors with anonymised cohort analytics and real-time alerts, the system offers a coherent and privacy-respecting approach to scalable campus well-being support. Evaluation demonstrated negligible screening overhead, sub-second time-to-first token for streamed dialogue, and a distress recall of 0.93, confirming the practicality of the design. While positioned firmly as an aid rather than a clinical substitute, the framework provides a deployable foundation whose planned extensions toward richer affect modelling, accelerated inference, and ethically validated trials promise meaningful impact on student mental-health provision.

REFERENCES

- [1] American College Health Association, "National College Health Assessment: undergraduate reference report," ACHA, 2022.
- [2] C. Auerbach et al., "WHO World Mental Health surveys: mental disorders among college students," *Psychological Medicine*, vol. 46, no. 14, pp. 2955–2970, 2020.
- [3] D. Eisenberg, S. Lipson, and J. Hunt, "Stigma and help-seeking for mental health among students," *Journal of Affective Disorders*, vol. 271, pp. 123–130, 2020.
- [4] K. Fitzpatrick, A. Darcy, and M. Vierhile, "Delivering cognitive behaviour therapy via a conversational agent: a randomised trial," *JMIR Mental Health*, vol. 4, no. 2, e19, 2017.
- [5] B. Inkster, S. Sarda, and V. Subramanian, "An empathy-driven conversational agent for mental well-being: real-world evaluation," *JMIR mHealth and uHealth*, vol. 6, no. 11, e12106, 2018.
- [6] S. D'Alfonso, "AI in mental health," *Current Opinion in Psychology*, vol. 36, pp. 112–117, 2020.
- [7] C. J. Hutto and E. Gilbert, "VADER: a parsimonious rule-based model for sentiment analysis of social media text," in *Proc. ICWSM*, 2014, pp. 216–225.
- [8] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: pre-training of deep bidirectional transformers for language understanding," in *Proc. NAACL-HLT*, 2019, pp. 4171–4186.
- [9] A. Sharma, M. Choudhury, and T. Althoff, "Deep learning for emotion and distress detection in text," in *Proc. ACL*, 2021, pp. 1450–1463.
- [10] M. De Choudhury et al., "Predicting mental-health crisis from linguistic markers," in *Proc. CHI*, 2020, pp. 1–14.
- [11] L. Ouyang et al., "Training language models to follow instructions with human feedback," in *Adv. Neural Inf. Process. Syst.*, vol. 35, pp. 27730–27744, 2022.
- [12] A. Dubey et al., "The Llama 3 herd of models," arXiv:2407.21783, 2024.
- [13] P. Lewis et al., "Retrieval-augmented generation for knowledge-intensive NLP tasks," in *Adv. Neural Inf. Process. Syst.*, vol. 33, pp. 9459–9474, 2020.
- [14] N. Reimers and I. Gurevych, "Sentence-BERT: sentence embeddings using Siamese BERT-networks," in *Proc. EMNLP-IJCNLP*, 2019, pp. 3982–3992.
- [15] S. M. Schueller and J. Torous, "Scaling evidence-based digital mental-health interventions," *npj Digital Medicine*, vol. 3, art. 50, 2020.
- [16] J. Torous et al., "Digital mental health and longitudinal mood monitoring," *World Psychiatry*, vol. 20, no. 3, pp. 318–335, 2021.
- [17] J. Johnson, M. Douze, and H. Jégou, "Billion-scale similarity search with GPUs," *IEEE Trans. Big Data*, vol. 7, no. 3, pp. 535–547, 2021.

**BIOGRAPHY**

KONALA NAGA SOWMYA SREE received the B.Sc. degree from Viswa Teja degree, Penugonda, West Godavari, India, in 2024. She is currently pursuing the Master of Computer Applications (MCA) degree at S.V.K.P. & Dr.K.S. Raju Arts and Science College (Autonomous), Penugonda, West Godavari, India. Her research interests include Artificial Intelligence, Python programming, software engineering, and Web development. She is actively involved in academic projects related to AI-based solutions and modern software technologies. Her goal is to contribute to innovative research and practical applications that address real-world challenges through continuous learning, technological advancement, and software development.



Mr. B. N. SRINIVASA GUPTA is working as Associate Professor in SVKP & Dr.K.S. Raju Arts & Science College (Autonomous), Penugonda, A.P. He received Master's Degree in Computer Applications from Andhra University and Computer Science & Engineering from Jawaharlal Nehru Technological University Kakinada (JNTUK), Kakinada, India. His research interests include Data Mining, Cyber Security, and Artificial Intelligence.