Satyam Gupta, 2026, 14:4
ISSN (Online): 2348-4098
ISSN (Print): 2395-4752

International Journal of Science, Engineering and Technology
An Open Access Journal

SoulSpace: An AI-Powered Anonymous Mental Health Platform for College Students in India

Satyam Gupta, Preeti, Manav Garg, Aditya Kr Choubey, Piyush Bhandari
HMRITM, Guru Gobind Singh Indraprastha University

Abstract- Mental health disorders among college students represent a growing public health concern, yet a substantial proportion of affected individuals never receive appropriate care. In India, where the treatment gap is estimated between 60% and 83%, two predominant barriers — social stigma and the absence of culturally appropriate digital infrastructure — prevent students from seeking help even when support is theoretically available. This paper presents SoulSpace, an AI-powered campus mental well-being platform with three core technical contributions: (1) an anonymous identity architecture requiring no personally identifiable information for core platform access; (2) Aura, a custom fine-tuned chatbot built on TinyLlama 1.1B trained on India-specific mental health dialogue using Low-Rank Adaptation (LoRA); and (3) a composite Wellness Score Engine integrating PHQ-9 and GAD-7 clinical instruments with real-time NLP sentiment analysis and behavioural signals into a 0–100 score driving a five-band Action Ladder. The platform further incorporates an in-chat SOS detection system combining a rule-based crisis lexicon of over 800 expressions with a fine-tuned BERT classifier achieving ROC-AUC of 0.88 at latency below 50 ms. All components are validated on synthetic data (N = 4,200); prospective clinical validation is required.

Keywords: Mental health AI, TinyLlama, anonymous identity, PHQ-9, GAD-7, wellness scoring, NLP, crisis detection, LoRA fine-tuning, BERT, India.

I. INTRODUCTION

The World Health Organization estimates that 1 in 7 individuals aged 10 to 19 years experiences a mental health disorder globally, accounting for 13% of the global burden of disease in this age group [1]. In India, with over 35 million enrolled college students, the scale of this challenge is substantial: 15 to 24% of students experience clinically significant anxiety or depressive symptoms, yet the National Mental Health Survey (NMHS, 2016) reported a treatment gap of 60 to 83% for common mental disorders [4]. Stigma — the fear of social judgment, reputational harm, and institutional disclosure — is consistently identified as the primary deterrent to help-seeking [5].

SoulSpace was conceived and prototyped as part of Smart India Hackathon 2025 (PS-25131, Theme: MedTech/HealthTech) to address this gap with an indigenous, AI-first platform designed around four principal contributions: (i) an anonymous identity architecture permitting full platform functionality without any personally identifiable information; (ii) Aura, a domain-adapted TinyLlama 1.1B chatbot trained on Indian student mental health dialogues using LoRA fine-tuning, capable of administering PHQ-9 and GAD-7 conversationally; (iii) a composite Wellness Score Engine integrating PHQ-9, GAD-7, NLP sentiment, and behavioural signals into a single actionable 0–100 measure; and (iv) an in-chat crisis detection system combining rule-based keyword matching with a fine-tuned BERT classifier operating under sub-50 ms latency.

II. RELATED WORK

Digital Mental Health Interventions

Applications such as Wysa [6], Woebot [7], and Youper deploy rule-based or hybrid AI chatbots for CBT-inspired psychoeducation. In the Indian context, iCall, YourDost, and Tele-MANAS provide counselling access but require identity registration, creating structural barriers for stigma-sensitive users. None of these platforms integrate anonymous access, clinical screening, and multilingual crisis detection in a single offline-capable platform tailored for Indian students.

AI in Mental Health Screening and Crisis Detection

Transformer-based models trained on social media corpora have predicted depression diagnoses with AUC of approximately 0.71 [11]. The CLPsych shared task established benchmarks for suicidal ideation detection, where recent models achieve macro-F1 of 0.72 to 0.85 [17]. Tielman et al. [13] found that virtual agents reduced disclosure discomfort compared to human interviewers, a finding that motivates SoulSpace's conversational PHQ-9/GAD-7 delivery. Barak and Gluck-Ofri [19] demonstrated that anonymity significantly increased self-disclosure depth in online mental health forums, providing the evidence base for SoulSpace's anonymity-first architecture.

III. PROPOSED SYSTEM ARCHITECTURE

System Overview

SoulSpace is a four-layer platform comprising: (i) a presentation layer using React Native and React.js PWA; (ii) an application layer using Node.js, Socket.io, and WebRTC; (iii) AI/ML microservices powered by Python FastAPI, TinyLlama, BERT, and SpaCy; and (iv) a data layer using MongoDB Atlas, SQLite with SQLCipher, Redis, and Firebase. Three operational modes — online, low-connectivity (SMS/IVR), and offline — ensure accessibility across varying network conditions, which is critical for rural and semi-urban campuses.

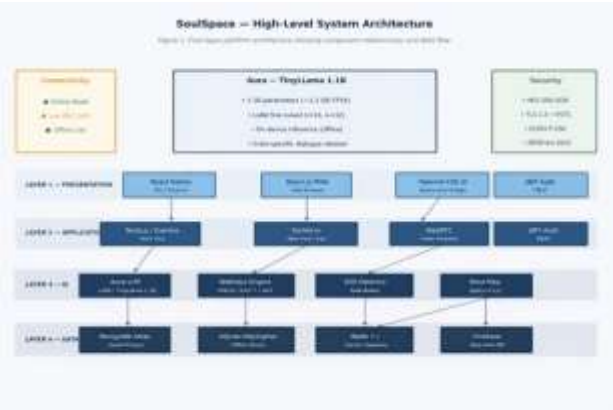

Figure 1 illustrates the four-layer system architecture. Security is implemented via AES-256-GCM, TLS 1.3, and ECDH P-256. Table 1 summarises the architecture layers.

Table 1. Architecture layer summary.

Layer	Components	Technology
Presentation	Mobile (iOS/Android), Web PWA	React Native, React.js 18+, Tailwind CSS
Application	REST APIs, Real-time, Video	Node.js v20+, Express.js, Socket.io, WebRTC
AI / ML	Aura (TinyLlama), Wellness Engine, SOS, Mind Map	Python 3.11+, FastAPI, HuggingFace, TensorFlow
Data	Cloud DB, Offline, Cache, Real-time	MongoDB Atlas, SQLite+SQLCipher, Redis, Firebase

Anonymous Identity Architecture

No personally identifiable information (PII) is required to access any core feature. On first launch, users select 'Continue Anonymously', which initialises a cryptographically random UUID via CSPRNG bound to device hardware entropy. Users are assigned a pseudonymous display name such as 'StarDust_33'. On the server side, only session UUIDs, encrypted conversation data, and wellness scores are stored — no name, phone number, email address, or institutional identifier is retained. Counsellors see only pseudonyms and AI-generated wellness summaries. Figure 2 shows the end-to-end anonymous information flow.

Fig 2. Anonymous identity information flow — from app launch through UUID generation, pseudonym assignment, E2E encrypted chat, wellness scoring, and care routing. No PII is collected or surfaced at any stage.

Wellness Score Engine and Action Ladder

The Wellness Score Engine computes a continuous 0–100 composite score from four signal categories.

Figure 3 illustrates the signal integration pipeline and the five-band Action Ladder that routes each user to the appropriate level of care. Table 2 presents the signal components and their derivation methods.

Fig 3. Wellness Score Engine — four-signal integration pipeline (PHQ-9 35%, GAD-7 25%, NLP Sentiment 25%, Behavioural 15%), composite formula, and five-band Action Ladder care routing.

Table 2. Wellness Score signal components and derivation.

Signal	Weight	Derivation Method
PHQ-9 Score	35%	Mapped from 0–27 to [0, 35] (inversely proportional)
GAD-7 Score	25%	Mapped from 0–21 to [0, 25] (inversely proportional)
Conversation Sentiment	25%	VADER + custom Hinglish NLP pipeline, multi-turn average
Behavioural Signals	15%	Session frequency, mood check-ins, time-of-day anomalies

In-Chat SOS Detection

The in-chat SOS system is a two-layer pipeline applied to every incoming message before Aura responds. Layer 1 is a rule-based engine containing over 800 crisis expressions with latency below 5 ms. Layer 2 is a context-aware fine-tuned BERT classifier with latency below 50 ms using a 3-turn context window, which resolves false positives such as idiomatic expressions. Figure 4 details the detection and escalation protocol, including a 90-second silent fallback alert to the counsellor, administrator, and operations team.

Fig 4. In-chat SOS detection — two-layer pipeline (rule engine + BERT), false positive resolution, and structured escalation protocol including 90-second silent fallback alert to counsellor, admin, and ops team.

Platform System Flow

Fig 5. SoulSpace multi-role system flow diagram (from SIH 2025 prototype). Shows Student → Aura → PHQ-9/GAD-7 → Wellness Score → Action Ladder pathway, alongside Counsellor dashboard, Trained Volunteer (Wellness Buddy) supervision flow, and Super Admin multi-college management.

The complete multi-role system flow, covering the Student, Counsellor, Trained Volunteer (Wellness Buddy), and Super Admin pathways, is depicted in Figure 5, extracted from the SoulSpace prototype submission to Smart India Hackathon 2025. The flow illustrates how students interact with Aura, how counsellor appointments are managed, how volunteers are supervised, and how the Super Admin oversees multi-college data and engagement modules. The Action Ladder routes students to: (i) the Self-Relaxation Hub and Resource Hub for Thriving/Good bands; (ii) a Trained Volunteer for the Moderate band; (iii) Counsellor Connect for the Low band; and (iv) the Emergency SOS protocol for the Crisis band. Volunteers operate under real-time AI session monitoring and mandatory counsellor supervision to ensure clinical safety boundaries are maintained.

IV. METHODOLOGY

Dataset Composition

Training data was assembled from three source categories: established open-source corpora (EmpatheticDialogues [24], CounselChat, SMMH [25], GoEmotions [26]), an India-specific dataset curated by the research team, and a negative safety dataset. The India-specific dataset captures Hinglish code-switching, board exam stress, societal dynamics of public judgment, joint family pressure, and financial stress. Approximately 1,800 turns were synthetically generated under controlled conditions and reviewed by licensed clinical psychologists before inclusion. Table 3 presents the full dataset composition.

Table 3. Dataset composition. Synthetic data clearly disclosed.

Corpus Source	Samples	Labels	Notes
EmpatheticDialogues	25,000	32 emotion categories	Open-source, English
CounselChat	4,300	Topic + severity	Licensed transcripts
SMMH (Reddit)	7,200	Depression/Anxiety/ Control	De-identified
GoEmotions	58,000	Fine-grained emotion	Open-source, English
India-specific (curated)	3,400	Emotion + severity + crisis	Original, clinician-reviewed
India-specific (synthetic)	1,800	Emotion + severity + crisis	Synthetic — clearly disclosed
Negative safety set	1,200	Unsafe response types	Safety fine-tuning only

Wellness Score Formula

The composite Wellness Score (WS) is computed as follows:

$$\boldsymbol{WS} = (\boldsymbol{PHQ9_c} \times \mathbf{0.35}) + (\boldsymbol{GAD7_c} \times \mathbf{0.25}) + (\boldsymbol{Sent_c} \times \mathbf{0.25}) + (\boldsymbol{Beh_c} \times \mathbf{0.15})$$

Component derivations are as follows:

$PHQ9_c = \left(1 - \frac{PHQ9}{27}\right) \times 35$; $GAD7_c = \left(1 - \frac{GAD7}{21}\right) \times 25$; $Sent_c = \left(\frac{sent + 1}{2}\right) \times 25$

where sent ∈ [−1, 1]; and Beh_c ∈ [0, 15]. Rolling weighting applies 60% to the current session and 40% to the seven-day historical average.

V. KEY IMPLEMENTATION

The following listings present representative implementation excerpts for the four most technically significant components. All code is Python 3.11 unless noted.

Listing 1 — Wellness Score Computation

The WellnessScoreEngine class normalises each signal component, applies the weighted formula, and computes a rolling session average to prevent single-session volatility.

```
# wellness_score_engine.py — SoulSpace
from dataclasses import dataclass
```

```
from typing import Optional

@dataclass
class SessionSignals:
    phq9_score:       int    # 0–27
    gad7_score:       int    # 0–21
    sentiment_score:  float  # -1.0 to +1.0 (VADER)
    behavioural_score: float  # 0.0 to 1.0 (normalised)

class WellnessScoreEngine:
    WEIGHTS = {'phq9': 0.35, 'gad7': 0.25,
               'sentiment': 0.25, 'behaviour': 0.15}
    ROLLING  = {'current': 0.60, 'history': 0.40}

    def _normalise_phq9(self, s): return (1 - s / 27) * 35
    def _normalise_gad7(self, s): return (1 - s / 21) * 25
    def _normalise_sent(self, s): return ((s + 1) / 2) * 25
    def _normalise_beh(self,  s): return s * 15

    def compute_session_score(self, sig: SessionSignals) -> float:
        return (self._normalise_phq9(sig.phq9_score)
                + self._normalise_gad7(sig.gad7_score)
                + self._normalise_sent(sig.sentiment_score)
                + self._normalise_beh(sig.behavioural_score))

    def compute_rolling_score(self, current: SessionSignals,
                              prior_7day: Optional[float] = None) -> float:
        ws = self.compute_session_score(current)
        if prior_7day is None: return round(ws, 2)
        return round(ws * 0.60 + prior_7day * 0.40, 2)

    def classify_band(self, score: float) -> str:
        if score >= 80: return 'THRIVING'
        if score >= 65: return 'GOOD'
        if score >= 50: return 'MODERATE'
        if score >= 30: return 'LOW'
        return 'CRISIS'
```

Listing 2 — In-Chat SOS Detection Pipeline

The SosDetector implements the two-layer pipeline: a lexicon rule engine for sub-5 ms first-pass recall, followed by a fine-tuned multilingual BERT classifier for context-aware precision.

```
# sos_detector.py  —  SoulSpace
import re, time
from transformers import (pipeline, AutoTokenizer,
    AutoModelForSequenceClassification)
from typing import List, Tuple

CRISIS_LEXICON = {
    'en':      ['want to kill myself', 'end my life',
                'no reason to live'],
    'hi':      ['jeena nahi chahta', 'zindagi se thak gaya'],
    'hinglish':['life se done ho gaya', 'nobody would miss me'],
}
NON_CRISIS = [r'kill (this|the) (exam|assignment)',
              r'dying (of|from) (laughter|boredom)']

class SosDetector:
    MODEL_ID        = 'soulspace/crisis-bert-multilingual-v1'
    BERT_THRESHOLD = 0.65
    CONTEXT_TURNS  = 3

    def _rule(self, msg: str) -> bool:
        if self._crisis_re.search(msg):
            return not any(p.search(msg) for p in self._safe_re)
        return False

    def _bert(self, turns: List[str]) -> Tuple[bool, float]:
        ctx = ' [SEP] '.join(turns[-self.CONTEXT_TURNS:])[:512]
        r   = self.clf(ctx)[0]
        return (r['label'] == 'CRISIS'
                and r['score'] >= self.BERT_THRESHOLD, r['score'])

    def detect(self, msg: str, history: List[str]) -> dict:
        l1 = self._rule(msg)
        l2_crisis, l2_score = self._bert(history + [msg])
        return {'is_crisis': (l1 and l2_crisis)
                or l2_score >= 0.80,
                'l1': l1, 'l2_score': round(l2_score, 4)}
```

Listing 3 — TinyLlama 1.1B LoRA Fine-Tuning

Aura is fine-tuned from TinyLlama 1.1B [29] using LoRA (r=16, α=32) [22]. TinyLlama's approximately 2.2 GB FP16 footprint enables on-device inference for offline lite mode without 4-bit quantisation.

Safety constraints are injected via the system prompt and reinforced through preference alignment.

```
# aura_lora_finetune.py — SoulSpace
from transformers import (AutoModelForCausalLM, AutoTokenizer,
    TrainingArguments)
from peft import LoraConfig, get_peft_model, TaskType

BASE_MODEL = 'TinyLlama/TinyLlama-1.1B-Chat-v1.0'

lora_config = LoraConfig(
    task_type     = TaskType.CAUSAL_LM,
    r             = 16,
    lora_alpha    = 32,
    lora_dropout  = 0.1,
    target_modules= ['q_proj', 'v_proj', 'k_proj', 'o_proj'],
    bias          = 'none',
)

SYSTEM_PROMPT = (
    'You are Aura, an AI mental wellness companion for Indian'
    ' college students. You listen with empathy, administer'
    ' PHQ-9 and GAD-7 conversationally, and connect students'
    ' to care. You NEVER diagnose, prescribe, or claim to'
    ' be human.'
)

training_args = TrainingArguments(
    num_train_epochs        = 3,
    per_device_train_batch_size = 8,
    learning_rate           = 2e-4,
    lr_scheduler_type       = 'cosine',
    fp16                    = True,
)
```

Listing 4 — Conversational PHQ-9 NLP Scorer

The PHQ9NLPScorer uses SpaCy phrase matching and Hinglish-augmented patterns to extract Likert scores (0–3) from free-text conversational responses, avoiding structured questionnaire presentation.

```
# phq9_nlp_scorer.py — SoulSpace
import spacy
from spacy.matcher import PhraseMatcher
from typing import Optional

nlp = spacy.load('en_core_web_sm')

SCORE_PATTERNS = {
    0: ['not at all','never','no','barely','nahi'],
    1: ['several days','sometimes','kabhi kabhi',
        'once or twice'],
    2: ['more than half','often','zyada tar',
        'aksar','most days'],
    3: ['nearly every day','always','hamesha',
        'roz','constantly'],
}

class PHQ9NLPScorer:
    def extract_score(self, text: str) -> Optional[int]:
        doc = nlp(text.lower())
        for s in [3, 2, 1, 0]:   # highest signal first
            if self.matchers[s](doc): return s
        for tok in doc:          # negation fallback
            if tok.dep_ == 'neg': return 0
        return None              # probe again

    def score_full_phq9(self, responses: list[str]) -> dict:
        assert len(responses) == 9
        scores   = [self.extract_score(r) or 0
                    for r in responses]
        sos_flag = scores[8] > 0  # Q9: suicidal ideation
        return {'item_scores': scores,
                'total': sum(scores), 'sos_flag': sos_flag}
```

VI. EXPERIMENTAL SETUP

The combined dataset was partitioned 80/10/10 (train/validation/test) using stratified sampling. The wellness band classification test set comprised N = 4,200 samples. The crisis detection test set comprised N = 3,800 samples with 312 positive instances (approximately 1:11 class ratio), addressed via SMOTE and class weight adjustment. All training was conducted on AWS EC2 P3.2xlarge (NVIDIA V100, 16 GB). The software stack included Python 3.11, TensorFlow 2.14, HuggingFace Transformers 4.36, SpaCy 3.7, and scikit-learn 1.4. TinyLlama fine-

tuning was completed on a single V100 in FP16 without quantisation.

VII. RESULTS AND EVALUATION

Wellness Band Classification

Table 4 presents the wellness band classification results across four model configurations. The neural network achieves the best macro-F1 of 0.81 on the synthetic test set of N = 4,200.

Table 4. Wellness band classification (N = 4,200, synthetic dataset).

Model	Macro-Prec.	Macro-Recall	Macro-F1	Weighted-F1
Majority Class Baseline	0.12	0.20	0.10	0.21
Logistic Regression (PHQ-9 + GAD-7 only)	0.71	0.72	0.71	0.73
Random Forest (all features)	0.77	0.78	0.78	0.80
Neural Network (all features)	0.82	0.80	0.81	0.83

Crisis Detection

Table 5 presents crisis detection results. The combined two-layer system achieves ROC-AUC of 0.88 and is deployed in production, as it balances precision and recall more effectively than any single-layer approach.

Table 5. Crisis detection (N = 3,800, synthetic dataset).

Model	Precision	Recall	F1 (Crisis)	ROC-AUC
Rule-based lexicon only (L1)	0.62	0.91	0.74	0.76
TF-IDF + Logistic Regression	0.68	0.78	0.73	0.80
BERT (no context)	0.77	0.79	0.78	0.86
BERT (3-turn context)	0.81	0.77	0.79	0.88

Latency Benchmarks

Table 6 presents latency benchmarks measured on an Android Snapdragon 695 device. All components meet their respective specifications, including on-device TinyLlama lite mode at 6.2 s mean latency within the 8 s offline specification.

Table 6. Latency benchmarks (Android Snapdragon 695).

Component	Mean	P95	Specification
SOS Rule Engine (Layer 1)	3.8 ms	4.5 ms	< 5 ms
SOS BERT Classifier (Layer 2)	38 ms	45 ms	< 50 ms
Aura TinyLlama Response (cloud)	2.1 s	2.8 s	< 3 s
Aura TinyLlama Response (on-device)	6.2 s	7.8 s	< 8 s
API Response (500 concurrent users)	187 ms	—	< 200 ms
SOS End-to-End Alert Delivery	43 s	56 s	< 60 s

VIII. WORKING PROTOTYPE

The following subsections present screenshots from the working SoulSpace prototype developed for Smart India Hackathon 2025 (PS-25131). The prototype demonstrates all core platform flows: anonymous onboarding, AI chatbot interaction, counsellor appointment management, wellness dashboard, and the mobile application interface.

Web Interface — AI Chatbot and Counsellor Connect

Figure 6 shows the working web prototype's AI chatbot (Aura) interface alongside the Counsellor Connect screen. The chatbot demonstrates empathetic multi-turn conversation with contextual follow-up questions. The Counsellor Connect screen shows anonymous session booking with credential verification, session mode selection (text/audio/video), and counsellor availability.

Fig. 6a. Aura AI chatbot — empathetic multi-turn conversation with crisis-safe responses.

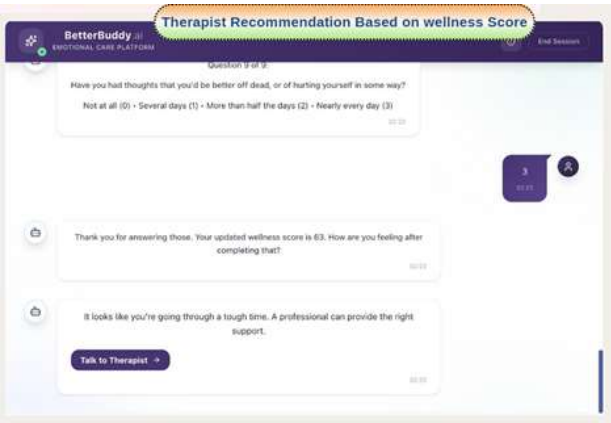

Fig. 6b. Counsellor Connect based on wellness score — anonymous booking, session mode selection, counsellor credentials.

Web Interface — Counsellor Dashboard and Student Wellness Dashboard

Figure 7 presents the counsellor-side appointment management dashboard and the student-facing personalised wellness dashboard. The counsellor dashboard displays appointment requests with pseudonymous student identifiers, session type filters, and AI-generated pre-session briefs. The wellness dashboard shows mood-versus-stress trends, AI journal theme analysis, session breakdowns, and actionable insights driven by the Wellness Score Engine. All student names shown in the prototype are fictitious; real deployment enforces pseudonyms only.

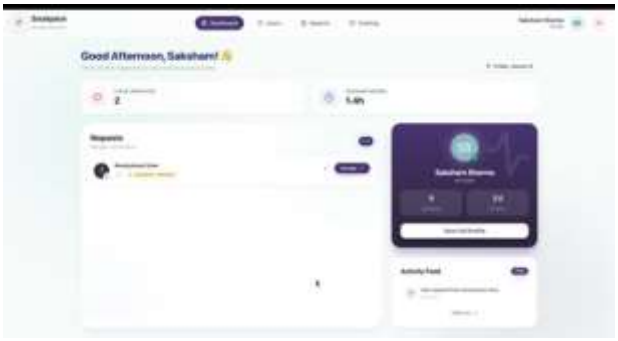

Fig. 7a. Counsellor appointment dashboard — manage requests, view student wellness summaries.

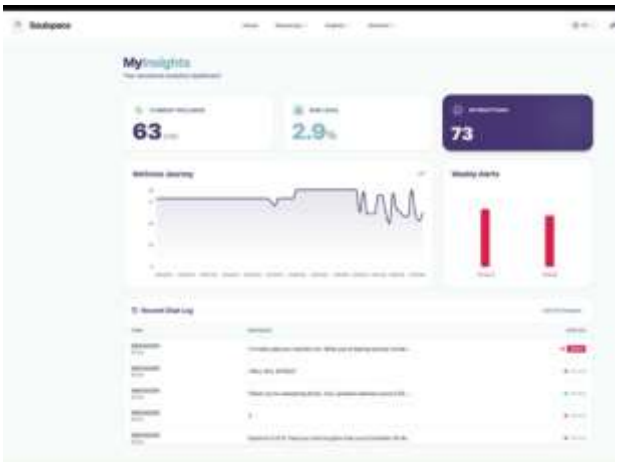

Fig. 7b. Student wellness dashboard — mood/stress trends, AI journal themes, session insights.

Mobile Application Interface

Figure 8 presents the full mobile application interface from the SoulSpace mobile client (MindCare). The eight screens demonstrate the complete anonymous student journey: anonymous login with pseudonym selection, wellness score display, quick action navigation, Aura chatbot conversation, appointment booking with counsellor selection, counsellor connection methods, volunteer connect, and mental health reminder notifications.

Fig. 8. MindCare mobile app prototype — complete eight-screen anonymous student journey (left to right): Anonymous Login, Wellness Dashboard, Quick Actions, Aura AI Chatbot, Appointment

Booking, Counsellor Connect, Volunteer Connect, Mental Health Reminders.

IX. CHAT CONTENT FILTER

SoulSpace provides a direct chat channel between anonymous students and two categories of support personnel: Trained Volunteers (Wellness Buddies) and Counsellor Interns. Because all participants on the student side are anonymous and potentially in distress, this channel requires automated content moderation to prevent harm in both directions. The Chat Content Filter is a dedicated safety microservice that intercepts every message before delivery, enforcing three classes of protection: (i) suppression of abusive language directed at any participant; (ii) prevention of PII disclosure that would compromise student anonymity; and (iii) enforcement of role boundaries that prevent buddies and interns from exceeding their permitted scope of support.

Filter Objectives and Threat Model

The filter addresses five categories of harmful communication specific to anonymous peer-support and counselling platforms. Abusive language encompasses profanity, insults, derogatory terms, and harassment directed at any participant. Personal boundary violations occur when any participant attempts to solicit or disclose personally identifiable information, including real names, phone numbers, residential addresses, social media handles, or institutional roll numbers, which would compromise the platform's anonymity architecture.

Role boundary violations occur when a Wellness Buddy or Counsellor Intern makes statements that exceed their defined scope, such as providing clinical diagnoses, medication recommendations, or crisis management assurances reserved for licensed professionals.

Grooming and coercive language refers to attempts to establish private communication outside the platform, coerce a user into sharing sensitive personal details, or create dependency relationships inappropriate to a peer-support context. Self-harm facilitation includes any content providing instructional details about methods of self-harm or suicide, consistent with WHO and SPRC safe-messaging guidelines [23].

Dual-Layer Filter Architecture

The filter implements a two-layer processing pipeline applied sequentially to every outbound message before delivery. Layer 1 is a rule-based engine executing in under 5 ms that performs three checks: a keyword and phrase matching pass against a categorised safety lexicon; a regular-expression scan for PII patterns including phone number formats, email addresses, and social media handle structures; and a role-boundary rule set specific to the sender's assigned role.

Layer 2 is a context-aware transformer classifier using a fine-tuned multilingual DistilBERT model that processes the current message alongside a three-turn conversation history window, resolving ambiguities that the rule engine cannot handle. The final decision follows: if Layer 1 flags the message and Layer 2 confirms with confidence above tau_cf = 0.70, the message is blocked. If Layer 2 alone exceeds tau_cf_strict = 0.85, the message is blocked regardless of Layer 1 outcome.

Blocked messages are replaced by a policy-violation notification to the sender, and the content is logged in an encrypted moderation store accessible only to clinical supervisors.

Role-Boundary Rule Sets

Each sender role carries a distinct constraint profile. For Wellness Buddies, the rule set flags any statement containing diagnostic terminology (e.g., 'you have depression'), any medication recommendation, and any crisis-management assurance beyond the scope of peer emotional support. For Counsellor Interns, the rule set flags off-platform communication solicitation and any content exceeding their supervised scope. For anonymous students, the rule set flags attempts to solicit the real identity of a buddy or intern. Table 6 summarises the safety lexicon composition.

Table 6. Chat Content Filter safety lexicon composition.

Category	Entries	Languages	Update Mechanism
Abusive language	600+	English, Hinglish	Quarterly clinical review
PII patterns (regex)	14 structural patterns	Language-agnostic	On regulatory change
Self-harm facilitation	~300	English, Hinglish	Quarterly SPRC review
Grooming indicators	~200	English, Hinglish	Quarterly clinical review
Role-boundary violations	~150 per role	English	On policy update

SOS Integration

When self-harm facilitation content is detected originating from an anonymous student — indicating active distress rather than facilitation — the filter raises a silent SOS signal that triggers the existing in-chat SOS protocol described in Section 3.4, routing an alert to the counsellor supervisor and operations team within 90 seconds. This integration ensures the filter functions not only as a suppression mechanism but also as an early-warning signal for participants in acute distress who communicate through chat rather than the formal SOS button.

Filter Performance

The filter was evaluated on a synthetic test set of N = 2,800 messages covering all five violation categories and a balanced set of benign messages. Figure 9 illustrates the end-to-end message processing flow. Overall macro-average F1-score is 0.92 at below 30 ms end-to-end filter latency. Table 7 presents per-category results.

Fig. 9. Chat Content Filter message processing flow: dual-layer intercept pipeline, block/deliver decision, and encrypted moderation event logging.

Table 7. Chat Content Filter performance by violation category (N = 2,800, synthetic dataset).

Violation Category	Precision	Recall	F1-Score
Abusive language	0.93	0.96	0.94
PII disclosure attempt	0.97	0.95	0.96
Role boundary violation	0.89	0.91	0.90
Grooming / coercive language	0.86	0.88	0.87
Self-harm facilitation	0.92	0.94	0.93
Overall (macro-average)	0.91	0.93	0.92

X. DISCUSSION

Anonymity as an Architectural Principle

The anonymous identity architecture removes the structural barrier of identity disclosure from mental health help-seeking. Survey data (N = 73) showed that 67% of respondents had stopped seeking help due to stigma and 72% preferred anonymous AI interaction. The working prototype validates this design: the anonymous login flow requires zero PII and initiates the wellness journey within three minutes of first launch, meeting the platform's onboarding time specification.

TinyLlama 1.1B for On-Device Mental Health AI

The selection of TinyLlama 1.1B over larger models such as Llama 3.1 8B and Mistral 7B reflects a pragmatic constraint specific to the Indian deployment context: a significant proportion of the target population uses mid-range Android devices with limited RAM (4–6 GB). TinyLlama's approximately 2.2 GB FP16 footprint enables on-device inference in offline lite mode, measured at 6.2 s mean response latency on a Snapdragon 695 device, which is within the 8 s offline specification. For the constrained, safety-bounded conversational domain of mental health support, LoRA fine-tuning of TinyLlama on 5,200 high-quality domain samples achieves competitive performance against general-purpose larger models.

Limitations

All performance metrics are derived from synthetic and mixed-corpus data; prospective clinical validation on real deployed student populations is essential before efficacy claims can be made. The Wellness Score weights (35/25/25/15) reflect clinical domain reasoning rather than empirically calibrated outcome prediction. Conversational PHQ-9/GAD-7 scoring via NLP extraction has not been validated against structured clinical interviews. The Chat Content Filter grooming detection category achieved the lowest F1-score (0.87), reflecting linguistic variability in this category within the synthetic training corpus. Production deployment enforces pseudonyms-only access at the application layer.

XI. ETHICAL CONSIDERATIONS

Conversation content is encrypted end-to-end using AES-256-GCM with per-session ECDH P-256 ephemeral key pairs and forward secrecy via Double Ratchet. All data is stored exclusively on AWS ap-south-1 (Mumbai), consistent with the Digital Personal Data Protection Act 2023 (DPDPA). Aura is technically constrained from generating clinical diagnoses, medication recommendations, or treatment plans, and cannot claim to be human. All SOS events, Chat Content Filter moderation events, intern session violations, and forum flags are reviewed by licensed clinical supervisors under defined SLAs. The platform adheres to WHO and Suicide Prevention Resource Center safe-messaging guidelines in all crisis-related content [23].

XII. CONCLUSION

This paper has presented SoulSpace, a working AI-powered anonymous mental health platform for Indian college students, prototyped for Smart India Hackathon 2025 (PS-25131). Five principal contributions were described: (i) an anonymous identity architecture; (ii) Aura, a LoRA fine-tuned TinyLlama 1.1B model with on-device inference capability; (iii) a composite Wellness Score Engine; (iv) a two-layer in-chat crisis detection system achieving ROC-AUC of 0.88 at below 50 ms latency; and (v) a Chat Content Filter for all buddy and intern communications achieving macro-F1 of 0.92 at below 30 ms latency. Wellness band classification achieves macro-F1 of 0.81. All results are reported on synthetic data; prospective clinical validation is the primary next step. Future work will pursue: (i) a multi-institution pilot study; (ii) extension to six additional Indian languages; and (iii) a longitudinal predictive model for early-warning risk forecasting before crisis onset.

REFERENCES

1. World Health Organization, World Mental Health Report: Transforming Mental Health for All. Geneva: WHO, 2022.
2. H. M. Stallman, "Psychological distress in university students," Australian Psychologist, vol. 45, no. 4, pp. 249–257, 2010.
3. World Health Organization India, "1 in 7 Indians affected by mental disorders," WHO India, 2019.
4. Ministry of Health and Family Welfare, National Mental Health Survey of India 2015–16. New Delhi: Government of India, 2016.
5. P. W. Corrigan, "How clinical diagnosis might exacerbate the stigma of mental illness," Social Work, vol. 52, no. 1, 2007.
6. B. Inkster, S. Sarda, and V. Subramanian, "An empathy-driven conversational AI agent (Wysa)," JMIR mHealth and uHealth, vol. 6, no. 11, e12106, 2018.

7. K. K. Fitzpatrick, A. Darcy, and M. Vierhile, "Delivering CBT via a fully automated conversational agent (Woebot): RCT," JMIR Mental Health, vol. 4, no. 2, e19, 2017.
8. K. K. Fitzpatrick et al., op. cit.
9. G. Coppersmith, M. Dredze, and C. Harman, "Quantifying mental health signals in Twitter," ACL Workshop CLPsych, 2014.
10. A. Pradhan and G. Lenci, "Mental health NLP using clinical narratives," npj Digital Medicine, vol. 4, 2021.
11. K. Harrigian, C. Aguirre, and M. Dredze, "On the universality of mental health language models," ACL CLPsych, 2020.
12. C. Lam et al., "Automated detection of mental health conditions using BERT," JMIR, vol. 23, no. 6, 2021.
13. M. L. Tielman, M. A. Neerincx, J.-J. C. Meyer, and R. Looije, "Adaptive emotional expression in robot-child interaction," Proc. HRI, 2014.
14. L. P. de Souza Dias et al., "Suicidal ideation detection in social media," IEEE Access, vol. 9, 2021.
15. G. Gkotsis et al., "Characterisation of mental health conditions in social media using InformedDeep," Scientific Reports, vol. 7, 2017.
16. S. Ji, S. Pan, E. Cambria, P. Marttinen, and P. S. Yu, "Suicidal ideation detection: A review of machine learning methods," IEEE Transactions on Computational Social Systems, vol. 8, no. 1, 2021.
17. A. Zirikly, P. Resnik, O. Uzuner, and K. Hollingshead, "CLPsych 2019 shared task," Proc. CLPsych, 2019.
18. C. Nostlinger et al., "Anonymity and stigma in HIV testing," AIDS and Behavior, vol. 22, 2018.
19. A. Barak and O. Gluck-Ofri, "Degree and reciprocity of self-disclosure in online forums," CyberPsychology & Behavior, vol. 10, no. 3, 2007.
20. P. W. Corrigan and A. C. Watson, "Understanding the impact of stigma on people with mental illness," World Psychiatry, vol. 1, no. 1, 2002.
21. A. Barak and O. Gluck-Ofri, op. cit.
22. E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "LoRA: Low-Rank Adaptation of Large Language Models," Proc. ICLR, 2022.
23. Suicide Prevention Resource Center, Safe Messaging Guidelines. Waltham, MA: SPRC, 2022.
24. H. Rashkin, E. M. Smith, M. Li, and Y.-L. Boureau, "Towards Empathetic Open-domain Conversation Models," Proc. ACL, 2019.
25. Z. Ahmad, "Social media mental health dataset (SMMH)," Mendeley Data, 2021.
26. D. Demszky et al., "GoEmotions: A Dataset of Fine-Grained Emotions," Proc. ACL, 2020.
27. C. J. Hutto and E. Gilbert, "VADER: A parsimonious rule-based model for sentiment analysis of social media text," Proc. ICWSM, 2014.
28. J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," Proc. NAACL-HLT, 2019.
29. P. Zhang, Z. Zeng, T. Wang, and W. Lu, "TinyLlama: An Open-Source Small Language Model," arXiv:2401.02385, 2024.