RESEARCH ARTICLE - OPEN ACCESS

Journal of Mental Health and Physical Health DOI: https://dx.doi.org/10.51521/JMHPH.2023.e21.103

Contents lists available at bostonsciencepublishing.us

Journal of Mental Health and Physical Health

Journal Of Mental Health And Physical Health JMHPH

AI and Mental Health Among Adolescents and Young Adults: Investigating AI Models to Understand Trends, Risks, and Preventive Interventions in Schools and Social Settings

[1]Damilola Sherifat Shaba, [2]Ridwan Adebowale Yusuf, [3]Taofeek Akinwumi Raheem, [4]Sodiq Abiola Omotosho

[1]*Department of Public Health, University of Illinois Springfield. Illinois, USA.*

[2]*Department of Social Science, Acap University College. Victoria, Australia.*

[3]*Department of Business Management, Eastern Gateway Community College. Ohio, USA.*

[4]*Department of Verterinary Science and Public Health, Fort Valley State University. Georgia, USA.*

ARTICLE INFO

Article history:
Received 02 April 2023
Revised 15 June 2023
Accepted 21 July 2023
Published 05 November 2023

Keywords:
Adolescent Mental Health
Artificial intelligence
Machine learning
Preventive interventions.

ABSTRACT

Background: Adolescent and young-adult mental health is a global public-health priority. Most mental disorders begin before age 24, with roughly half of lifetime cases starting by mid-adolescence and up to 74% by age 24 (Kessler et al., 2005) [1]. Depression and anxiety are prevalent among college students, yet an estimated 75% of those needing help do not access services due to stigma and other barriers (Hunt & Eisenberg, 2010, as cited in Fitzpatrick et al., 2017) [2]. Early interventions often last only ~2 years, and up to 80% of youths may relapse after services end [3].

Objective: This study aims to develop AI models to identify mental-health trends among adolescents/young adults, assess risks (e.g. privacy, bias, stigma), and propose preventive interventions in schools and social settings.

Methods: We outline a multi-phase mixed-methods design. Participants (ages 12–24) will provide survey data, social-media text, and school records. Natural language processing (NLP) and machine-learning algorithms (e.g. classifiers) will detect sentiment and risk patterns. Model performance will be evaluated by accuracy, precision/recall, and fairness metrics.

Results: We anticipate identifying key prevalence patterns (e.g. demographic differences), achieving moderate predictive performance (expected F1 ~0.8), and uncovering risk factors (e.g. bullying, social media use) correlating with distress.

Conclusions: AI can augment early detection and support for youth mental health, but ethical safeguards privacy protection, bias mitigation, and human oversight are crucial to ensure these tools are safe and effective complements to traditional care.

Introduction

Context and Significance

Mental-health problems frequently emerge during adolescence and young adulthood, making youth mental health a critical public-health concern. Epidemiological data indicate that roughly 50% of all lifetime mental disorders begin by age 14, and as many as three-fourths by the mid-20s, emphasizing the early onset of psychiatric conditions [1]. Globally, approximately 13.4% of children and adolescents experience a mental disorder at any given time [4]. Among adolescents, depression and anxiety are leading causes of disease burden and disability [5]. For example, one meta-analysis found the pooled prevalence of *any* mental disorder in youth to be ~13%, including ~6.5% for anxiety disorders and ~2.6% for depressive disorders [4]. Suicide is a major concern: it ranks among the top causes of death in older adolescents and young adults worldwide [5]. Mental illness in youth can severely affect academic performance and social development. Depression and anxiety often lead to poor school attendance, declining grades, and withdrawal from peer relationships [6]. Longitudinal studies link adolescent depression with higher dropout rates and impaired social functioning in adulthood (Patel et al., 2007) [7]. In sum, the high prevalence and early onset of mental disorders paired with their impact on educational attainment and interpersonal relationships underscore the urgency of proactive youth mental-health strategies.

Despite the clear need, most young people with mental-health needs do not receive treatment. Surveys of college populations in the United States have revealed that up to 75% of students who could benefit from services do not access clinical care, often due to stigma, limited awareness, or resource constraints (Hunt & Eisenberg, 2010) [2]. Many adolescents hesitate to seek help for fear of being labeled or judged, even when free counseling is available on campus [2]. Those who do engage in treatment may receive only short-term interventions: specialized early-intervention programs (e.g. for first-episode psychosis or severe depression) typically

*Corresponding author :Damilola Sherifat Shaba, Department of Public Health, University of Illinois Springfield. Illinois, USA. ORCID: 0009-0008-7915-1122.

provide intensive support for about two years [8]. Unfortunately, the gains from brief interventions often fade; an estimated up to 80% of youth relapse into mental ill-health after initial improvement [3]. This relapse rate points to gaps in continuity of care as young people transition out of youth-focused services. Together, these factors – high unmet need and fragile long-term outcomes – highlight an imperative to explore innovative, scalable approaches for early identification, continuous support, and relapse prevention among adolescents and young adults.

Role of Artificial Intelligence in Youth Mental Health

Emerging technologies, especially artificial intelligence (AI), offer new avenues to support youth mental health. AI approaches in this domain include machine-learning algorithms for pattern recognition, natural-language processing (NLP) of text data, chatbot-based conversational agents, and sensor-driven analytics. Each of these can extend the reach of traditional services. Notably, today's youth are almost universally connected: as of the mid-2010s, more than 97% of adolescents go online daily, and nearly half are online "almost constantly" [9]. This ubiquitous connectivity makes digital interventions a promising channel for mental-health support [9]. Online and mobile platforms can overcome barriers of location and scheduling, delivering help on demand and in privacy. Indeed, internet-based interventions have demonstrated effectiveness comparable to face-to-face therapy for common conditions like anxiety and depression [10] [11]. In particular, web-delivered cognitive-behavioral therapy (CBT) self-help programs have shown outcomes on par with therapist-delivered CBT in reducing symptom severity (e.g., similar improvements in depression scales) [12]. A key advantage is that digital tools allow anonymity, which may reduce the fear of stigma. Young people often feel more comfortable disclosing feelings to an app or AI agent; for example, more students reported suicidal thoughts when using a self-guided mobile app than on a standard in-person screening [13]. Thus, AI-driven digital solutions can mitigate stigma by providing a sense of privacy and control for the user.

Early implementations of AI in youth mental-health interventions are already showing promise. One notable approach is Moderated Online Social Therapy (MOST), a platform that integrates a social networking environment with therapeutic content and expert moderation [14]. MOST was designed in Australia to support young people recovering from mental illness. It features a moderated, peer-driven online community ("the Café") where users can post and comment in a Facebook-style newsfeed, fostering peer support and connection [14]. Alongside social networking, MOST offers interactive therapy modules called "Steps" – structured exercises and psychoeducational activities co-created by clinicians, creative writers, and youth [15]. These Steps teach coping skills via engaging narratives and comics, and they prompt user reflection through "Talking Point" discussion questions embedded in each module [15]. Users can also set personal goals and access vocational resources through features like "Team Up" and a "Job Zone" [16]. Crucially, the platform is moderated by clinicians (e.g. e-mental health professionals) who guide discussions and ensure safety, blending peer-to-peer interaction with expert oversight [17]. MOST has been deployed in trials for youth with early psychosis and depression, demonstrating high engagement – the social networking component tends to be the most frequently used feature [18]. To further enhance MOST, researchers are incorporating AI elements such as sentiment analysis and chatbot assistants to tailor content delivery [19]. For instance, natural-language algorithms can analyze a user's posts in real time and suggest relevant therapeutic Steps, or even trigger a supportive chatbot response [20]. In summary, platforms like MOST exemplify how AI can amplify scalable online therapy by combining peer support, evidence-based digital content, and data-driven personalization.

Another frontier is the use of fully automated conversational agents ("chatbots") for mental health. A pioneering example is Woebot, a chatbot that delivers CBT techniques through brief daily conversations. In a 2017 randomized controlled trial (RCT) with college students, Woebot was tested as a self-help intervention for mild to moderate depression and anxiety (Fitzpatrick et al., 2017). Over just two weeks, participants who engaged with Woebot on a daily basis experienced significant reductions in depressive symptoms compared to an information-only control group [21]. Specifically, the Woebot group's Patient Health Questionnaire (PHQ-9) depression scores dropped substantially (a moderate between-group effect size, $d \approx 0.44$) while the control group saw no meaningful change [21]. This fully automated agent – available 24/7 via a mobile app – uses brief text-based conversations to challenge cognitive distortions and encourage mood tracking. The RCT results suggest that even short-term chatbot interactions can measurably improve mood in young adults, at least in the immediate term. Another integrative AI system is "Tess", a psychological AI chatbot that delivers emotional support on demand. Tess engages users in free-form text conversation and provides a mix of CBT, motivational interviewing, and positive psychology exercises. In a controlled trial with college students, Tess users showed significant reductions in both depression and anxiety symptoms relative to a control group that received only an e-book of mental health information [22]. After 2–4 weeks of using Tess (via either daily or biweekly check-ins), students had improved scores on depression (PHQ-9) and anxiety (GAD-7) scales, whereas the control group did not improve [22]. These findings mirror the Woebot study and bolster evidence that AI-driven chatbots can effectively deliver aspects of therapy – at least for mild symptoms – in a scalable, cost-effective manner (Fulmer et al., 2018).

Overall, AI applications in youth mental health show promise in *extending the reach* of services beyond the clinic. Online interventions can meet young people "where they are" – which is often on their smartphones and social media. Over 97% of young people use the internet daily, so digital tools are uniquely positioned to engage this population [23]. Interventions like MOST leverage peer connectivity and creative content to maintain engagement over time, addressing the problem of post-treatment relapse by providing ongoing support [24]. Chatbot agents like Woebot and Tess demonstrate that *personalized conversational care* can be delivered without direct human involvement – an important feature when counseling resources are scarce or stigma prevents help-seeking. However, current AI solutions also highlight gaps: most studies so far focus on college students or young adults (18+), with limited research on minors under 18. The evidence base is largely short-term – few trials have assessed whether improvements sustain beyond a few months. Furthermore, the risks and ethical complexities of AI in mental health are not fully understood. Issues such as privacy, data security, algorithmic bias, and the need for human empathy are critical to address (see Section 4.5). To truly harness AI for youth mental well-being, these interventions must be carefully evaluated and integrated into school and community settings in ways that complement, rather than replace, human support. Our research tackles these challenges by systematically investigating how AI models can detect mental-health needs, what risk factors they can uncover, and how we can implement preventive strategies in real-world youth environments.

Research Objectives and Questions

Building on the context above, this study has four primary objectives:

Objective 1: Develop AI models to detect mental-health trends among adolescents and young adults by analyzing multi-source behavioral and textual data. We aim to use machine learning on inputs like survey responses, self-reported mood ratings, text messages, and social-media posts to identify patterns indicative of depression, anxiety, or other mental-health conditions.

Objective 2: Identify key risk and protective factors for youth mental health through AI-driven pattern analysis. By mining large datasets, we seek to uncover which factors (e.g. family history, bullying incidence, social media usage patterns, school performance) are most strongly associated with mental health outcomes in this population.

Objective 3: Examine the ethical, privacy, and psychosocial risks associated with AI-mediated mental-health interventions for youth. We will evaluate issues such as data privacy and consent, potential algorithmic biases (e.g. misclassification of certain demographic groups), the impact on stigma, and any adverse effects or user acceptance challenges when deploying AI tools in sensitive contexts.

Objective 4: Propose evidence-based preventive and intervention strategies that leverage AI insights for use in schools and social settings. This includes designing guidelines for integrating AI screening tools into school mental-health programs, and developing AI-enhanced resources (like chatbots or alert systems) for teachers, counselors, peers, and families to support at-risk youth.

From these objectives, we derive specific research questions (RQs) and hypotheses:

- **RQ1:** *How accurately can AI models detect or predict depression, anxiety, or suicidal ideation among adolescents and young adults using multi-modal data?* We hypothesize that an AI model using text-based features (e.g. sentiment in social-media posts) combined

with contextual data (e.g. past counseling visits, academic records) will achieve at least moderate predictive accuracy (e.g. Area Under ROC Curve ≥ 0.80) in identifying students with elevated mental-health risk.

- **RQ2:** *What risk factors and protective factors for youth mental health can be identified through machine-learning analysis, and how do these factors vary across different subgroups?* We hypothesize that the AI analysis will confirm known risk factors (such as bullying victimization, which is expected to correlate with higher depression scores, and strong family support, expected to be protective), and potentially reveal new patterns (for instance, specific linguistic markers or social media usage patterns that precede a spike in anxiety symptoms). We also anticipate that some risk factors' influence will differ by subgroup (e.g. the impact of social media usage on mental health might be more pronounced in younger teens than in college-aged young adults).
- **RQ3:** *What are the main ethical and privacy concerns when implementing AI-based mental health interventions for youth, according to stakeholders (students, parents, educators, clinicians), and how can these risks be mitigated?* We hypothesize that stakeholders will express concerns about data confidentiality (e.g. who can access students' sensitive information) and algorithmic fairness (e.g. fear of mislabeling or bias). We expect to identify strategies like robust consent processes, transparent AI decision criteria, and involvement of human counselors as mitigations to increase trust and safety.
- **RQ4:** *What preventive interventions can be recommended for schools and community youth programs based on our findings, and how can AI tools be integrated into these settings to improve early intervention?* We will explore, for example, whether schools can implement periodic AI-assisted mental health screenings or early alert systems and how effective those might be. We hypothesize that AI tools (like a mood-check chatbot or an analytics dashboard for school counselors) will be useful in complementing existing programs, by flagging at-risk students earlier and tailoring interventions (such as targeted psychoeducation modules for those who need them).

By addressing these questions, our research seeks to advance both the scientific understanding of youth mental health (through data-driven insights) and the practical toolkit available to educators and clinicians (through technology-informed intervention strategies). The ultimate goal is to harness the predictive power of AI in an ethical way to *prevent* the onset or escalation of mental-health problems among adolescents and young adults.

Literature Review

Prevalence and Risk Factors of Adolescent Mental-Health Problems

Mental-health disorders are widespread among adolescents and young adults, with significant implications for public health. As noted, around 10–20% of youths globally experience mental health conditions in any given year [4]. In broad population terms, anxiety disorders are the most common class among adolescents (worldwide prevalence ~6.5%), followed by disruptive behavior disorders (~5–6%), and depressive disorders (~2–3%) [4]. These conditions often emerge in the teenage years: the median age of onset is 11 for anxiety disorders and around 13–15 for many behavioral disorders, earlier than for mood disorders [1]. By late adolescence (15–19 years), depression and anxiety together rank among the leading causes of illness burden and disability in this age group [5]. In the United States, data from national surveys indicate that approximately 20% of youth (ages 13–18) have a diagnosable mental disorder, and about 10% experience severe impairment (e.g., severe depression) (Merikangas et al., 2010). College students also report high rates of psychological distress: in one survey, over 50% of students said they had felt overwhelming anxiety or depression impacting daily functioning in the past year [25]. These findings underscore that mental-health challenges are not the exception but relatively common during youth.

A complex interplay of risk factors contributes to the development of mental-health problems in adolescents and young adults. Biological factors include a family history of mental illness or genetic predispositions – for example, having a parent with depression significantly increases an adolescent's risk of developing depression (possibly two- to threefold higher risk in some studies). Gender is another factor: by mid-adolescence, females have roughly twice the rate of depression as males, a disparity that often persists into adulthood (perhaps due to hormonal differences, socialization, and life stress variations). Psychosocial and environmental factors are critically important. Chronic stressors such as poverty or low socioeconomic status (SES) elevate risk by exposing youth to adversity (e.g., financial insecurity, unsafe neighborhoods) that can erode mental well-being. Relatedly, exposure to violence or abuse – including childhood maltreatment or witnessing domestic violence – is a well-documented precursor to a range of mental disorders (from PTSD to depression). One particularly salient risk factor in adolescence is bullying and peer victimization. Studies have found that youths who are bullied (whether in-person or online) have significantly higher rates of depression, anxiety, and suicidal ideation than their non-bullied peers. For instance, a longitudinal study by Kim and Leventhal (2008) showed that bullying victims were at substantially increased risk of later developing depressive symptoms and were more likely to attempt suicide. Academic pressure and school-related stress also play a role, especially in cultures or educational systems with high stakes. Intense fear of failure or excessive workload can manifest as anxiety, sleep problems, and burnout in students. Conversely, academic difficulties can be both a consequence and a cause of mental-health issues, creating a vicious cycle (e.g., a depressed student's grades decline, which then further worsens self-esteem and depression).

Modern digital life has introduced new potential risk factors – and protective factors – for adolescent mental health. The rise of social media and smartphone use has coincided with increases in reported teen depression and self-harm in some studies [26]. Heavy social media use (for example, several hours per day on platforms like Instagram or TikTok) has been correlated with higher depressive symptoms, potentially through mechanisms like cyberbullying, social comparison, or disrupted sleep. However, the relationship is complex: online engagement can also provide social support and mental-health information. Screen time that displaces physical activity or face-to-face interaction may have negative effects on mood and anxiety, though content and context matter greatly (for instance, using the internet to actively connect with friends may be beneficial, whereas passive scrolling through distressing news could be detrimental). Other behavioral risk factors include substance use – which often begins in adolescence. Alcohol and drug use can exacerbate underlying mood disorders or trigger substance-induced mental-health conditions (and adolescents with mental illness are more likely to self-medicate with substances, compounding the problem). Finally, school context and climate influence student mental health: schools that lack anti-bullying policies, counseling resources, or that foster a high-pressure competitive atmosphere might see more student distress compared to supportive, inclusive school environments.

It is important to note that youths are not uniformly vulnerable; there are also protective factors that confer resilience. Strong family support and communication, positive peer relationships, involvement in community or religious activities, and the development of coping skills through extracurricular interests (sports, arts) have all been associated with better mental-health outcomes in adolescence. For example, an adolescent who feels connected to family – being able to talk openly with parents and receive emotional support – is less likely to develop severe depression even under stress, compared to a peer with family conflict. Similarly, schools that implement social-emotional learning programs and encourage help-seeking can mitigate some risk factors. In summary, the literature indicates that adolescent mental health is shaped by a convergence of factors: genetic vulnerability, gender, family environment, socioeconomic context, trauma exposure, peer dynamics, academic stress, and digital life. Understanding these risk and protective factors is crucial for developing targeted interventions – including those leveraging AI to detect when risk factors accumulate to a dangerous threshold.

Traditional and Digital Interventions for Youth Mental Health

Interventions to address adolescent mental health can be broadly categorized into traditional approaches (in-person and clinician-driven) and digital approaches, with increasing hybrid models that combine both. Traditional interventions include individual counseling or psychotherapy (such as cognitive-behavioral therapy, CBT), psychiatric medication when appropriate, and school-based mental-health programs. Cognitive-behavioral therapy has a strong evidence base for treating adolescent depression and anxiety; it typically involves helping the young person identify negative thought patterns and engage in healthier behaviors. Other modalities used with youth include behavioral activation (encouraging positive activities), interpersonal therapy (focusing on relationships), and family therapy (involving family members to improve communication and support). There are also numerous school-based programs aimed at prevention and early intervention. For example, some high schools implement universal mental-health education in health classes, teaching

students about stress management, emotional regulation, and how to seek help. Others have targeted programs like peer mentoring, in which trained student peer leaders provide support and referral for classmates in distress.

Meta-analyses of school-based interventions have found that such programs can have a modest but significant impact on reducing symptoms of depression and anxiety in the short term (often measured by standard symptom scales) [27]. For instance, a review by Stallard et al. (2013) on indicated (targeted) school-based interventions reported an effect size of around 0.2–0.3, indicating small reductions in depression symptoms compared to control conditions [27]. While traditional therapy delivered by a clinician is considered the gold standard for those with significant mental illness, there are barriers – including stigma, cost, and provider shortages – that prevent many youths from accessing these services [2]. This is where digital interventions have gained attention, as they can potentially circumvent some of these barriers.

Digital mental-health solutions for youth encompass a range of tools: computer-based therapy programs, mobile health apps, internet support groups, and teletherapy (videoconferencing with clinicians). A substantial body of research before 2019 examined computer- or web-delivered CBT programs. These programs often involve interactive modules that users complete weekly (sometimes described as "CBT in a box"). A remarkable finding from multiple controlled studies is that computer-assisted therapies can yield outcomes comparable to therapist-delivered CBT for mild-to-moderate depression and anxiety [12]. For example, Proudfoot et al. (2013) found that an online CBT program for depressed adults was as effective as face-to-face therapy in reducing depressive symptoms at post-treatment. For younger populations, a meta-analysis by Merry et al. (2012) on computerized interventions for adolescent depression showed significant symptom reductions versus waitlist. These programs are also noted to be cost-effective: once developed, they can be disseminated widely at low incremental cost per user, which is especially advantageous for resource-limited settings. One challenge, however, has been adherence – many youths do not complete the full course of unguided self-help programs, leading to dropout rates sometimes exceeding 50%. To improve engagement, developers have incorporated features like gamification (rewards, badges) and text message reminders. As an example, the Woebot chatbot study adopted a "daily dose" model (short daily conversations) and some gamified elements to keep users coming back [28], addressing the adherence issue that earlier static programs faced.

A key advantage of internet-delivered interventions is reducing stigma and increasing privacy. Young people often cite embarrassment or fear of judgment as reasons for not going to a school counselor or therapist. Online, they can seek help more *anonymously*. Research supports that stigma is lower for internet-based help: users appreciate the privacy of accessing therapy in their own room without anyone knowing [29]. Indeed, digital interventions "can help patients overcome the barrier of stigma by giving them the opportunity to privately seek evidence-based care" [29]. Additionally, internet interventions eliminate geographic barriers – a teenager in a rural area with no nearby psychologist can still access evidence-based resources through a computer or phone. Flexibility and convenience are other benefits: self-guided programs are available 24/7, allowing youth to get help at the moment they need it (even late at night when a crisis might occur). These factors make digital modalities particularly appealing to the youth demographic, which values on-demand and tech-integrated solutions [30].

Hybrid models have emerged that combine digital and human elements, aiming to get the best of both. One successful approach is "guided" internet therapy, where the adolescent works through an online program but has periodic check-ins with a therapist or coach via email or phone. This often yields better outcomes than unguided programs because the human support boosts adherence and allows some personalization. Another approach is integrating mobile apps into traditional therapy – for instance, a therapist might assign a mood-tracking app or mindfulness app as homework between sessions, then review the data together. Mobile apps for mental health have proliferated in recent years; many target young users with engaging interfaces. They range from mood-monitoring diaries to interactive learning games teaching CBT skills. Surveys indicate high interest: nearly 70% of patients (including young adults) express interest in using mobile apps to self-manage their mental health [31] [32]. Some apps have even demonstrated that they can encourage greater openness: as mentioned, a smartphone app for suicidal ideation reporting resulted in more disclosures than a face-to-face questionnaire [13], suggesting that well-designed digital tools can encourage honesty and self-reflection.

In summary, traditional interventions (therapy, school programs) are effective but often under-utilized by youth, whereas digital interventions (online programs, apps) offer accessibility and privacy advantages. Evidence up to 2018 indicates that internet-delivered CBT can effectively reduce symptoms, with outcomes comparable to in-person therapy in some cases [12]. The cost-effectiveness and stigma reduction of digital approaches are key strengths [29]. However, challenges like user engagement and ensuring safety remain. This sets the stage for AI-based interventions, which build upon digital approaches by adding intelligent automation and personalization – potentially increasing the potency and reach of mental-health support for adolescents and young adults.

AI-Based Interventions

Recent years have seen the advent of AI-powered interventions specifically designed to support youth mental health. These interventions leverage artificial intelligence techniques – such as natural language understanding and machine learning – to provide more responsive and personalized support than earlier digital tools. Below, we discuss notable AI-based interventions that were introduced prior to 2019, along with evidence of their effectiveness.

Moderated Online Social Therapy (MOST): As introduced earlier, MOST is an online platform that combines a social networking hub with therapeutic content and clinical moderation [14]. The hallmark of MOST is its integration of *multiple components*: (1) a peer-to-peer social network (called "the Café") where young users (typically 16–25 years old) post updates, share experiences, and support each other much like on Facebook [16]; (2) interactive therapy modules ("Steps") which are structured self-help exercises or psychoeducational materials that users can complete at their own pace [15]; and (3) expert moderation and coaching, wherein mental health professionals (and sometimes trained peer moderators) monitor the site, facilitate discussions, and ensure user safety [17]. Content on MOST is evidence-based, drawing from CBT, positive psychology, and other therapies, but delivered in youth-friendly formats (e.g., comics, videos, quizzes). Early versions of MOST were deployed for youths recovering from psychosis (in the Horyzons project) and for those with depression, demonstrating high user engagement and feasibility over trials of a few months (Alvarez-Jimenez et al., 2014). Usage analyses showed that the social networking component was heavily utilized – young people valued the ability to connect with peers with similar experiences, which helped reduce isolation [18]. Meanwhile, the self-guided therapy *Steps* allowed users to build skills (such as problem-solving or mindfulness) in a fun, interactive way [15]. A key development in the MOST project has been to infuse AI capabilities to enhance user experience and personalization. D'Alfonso et al. (2017) describe ongoing work to incorporate natural language processing (NLP) into the platform [19]. For instance, NLP algorithms can analyze user posts in the Café to detect their emotional tone or specific topics. If a user posts "I'm feeling hopeless today," the system can automatically recognize negative sentiment and then recommend appropriate Steps or resources – effectively an AI-driven *recommender system* for therapy content [20]. They are also trialing a chatbot within MOST to provide immediate responses or check-ins with users, simulating some aspects of a human moderator [20]. While full evaluation of these AI enhancements was still underway by 2018, MOST represents a model where AI assists human clinicians in scaling up an online therapy, by monitoring user data and delivering tailored interventions at scale. Early outcomes have been promising: MOST has shown sustained improvements in users' well-being and relapse rates in small trials, and the added AI features aim to further improve these outcomes by keeping users engaged and providing help at critical moments [33].

Chatbots (Conversational Agents): AI chatbots are one of the most visible AI interventions in mental health. They simulate a conversational partner who can talk the user through difficulties using therapeutic techniques. Woebot is a prime example – a chatbot developed by psychologists at Stanford, launched around 2017. Woebot uses a conversational style with a bit of personality and humor to engage users, sending daily messages that include mood check-ins ("How are you feeling today on a scale of 1–5?") and brief lessons or exercises drawn from CBT. The RCT by Fitzpatrick et al. (2017) provided initial evidence on Woebot's efficacy: in a sample of young adults (average age ~22) reporting depression and anxiety symptoms, those randomized to use Woebot for 2 weeks had significantly lower depression scores (PHQ-9) post-intervention compared to those who only received an information eBook (control) [21]. The difference was clinically meaningful for a short intervention, and users also reported high satisfaction and engagement with the chatbot (many interacted with it almost daily). Woebot's success lies in its ability

to mirror elements of a therapeutic conversation – it employs techniques like cognitive restructuring (challenging negative thoughts) and gratitude journaling, delivered in a casual chat format. Importantly, Woebot is fully automated and available 24/7, illustrating how AI can provide instant support without a human counselor. This doesn't replace human therapy for those who need more intensive care, but it fills a niche for *preventive* or supplemental support, particularly for subclinical populations (e.g., stressed college students) or as an interim resource on waitlists.

Another chatbot, Tess, takes a slightly different approach by aiming to provide a broader range of psychological support. Tess, developed by X2AI, is described as an "integrative psychological AI", meaning it draws from multiple therapeutic modalities (CBT, interpersonal therapy, etc.) and tailors its responses to the user's needs. In a 2018 trial, Tess was tested with university students who had self-identified stress, depression, or anxiety (Fulmer et al., 2018). Tess communicated with users via text message or web chat, and participants could interact at any time. The study found that students who engaged with Tess for a few weeks showed greater reductions in depression and anxiety symptoms than a control group that was given a static self-help eBook [22]. In fact, Tess users' GAD-7 anxiety scores improved significantly ($p = .02$ in one group) whereas controls did not, and depression scores improved with a moderate effect size [22]. Tess was thus shown to be feasible, engaging, and effective in symptom reduction for college students [32]. Qualitative feedback indicated that students appreciated Tess's immediacy and the sense of having someone to talk to at any hour – albeit an AI someone. Notably, Tess's design acknowledges it is not a therapist but a supportive tool, and the developers highlight that it should be overseen by professionals especially if a user is in crisis (the AI can recognize certain keywords like "suicide" and prompt emergency help procedures). Chatbots like Woebot and Tess highlight how AI can deliver therapy principles at scale, using simple text-based interfaces that resonate with a generation accustomed to texting and chatting. Early evidence suggests they can reduce mild to moderate symptoms in the short term [21] [22]. The low cost and high scalability of chatbots make them attractive for population-level interventions (for example, a school could offer a bot to all students as a wellness resource). The downside is that bots lack true empathy and complex understanding – they work best for structured, brief interactions and may not handle severe or highly complex situations well. That's why a common refrain is that these AI agents are a supplement to, not a replacement for, human care.

Integrative Systems and Coaches: Beyond individual apps, some interventions have tried to integrate AI into multi-component systems. For example, the Tess trial can be seen as integrating AI into a campus counseling ecosystem – perhaps in the future Tess could triage students: those who improve with Tess might not need further intervention, while those not improving could be flagged for counselor outreach. Other systems have incorporated virtual avatar coaches in online programs (not necessarily chatbots with free text, but scripted avatars that guide users through content). Early work in this area showed that adolescents often respond well to gamified, coach-led programs – one could envision an AI coach that adapts its style based on user responses. While robust evidence was still limited pre-2019, one example is an AI-driven program called SIMS (Smartphone Intervention for Mental Health) that combined a mobile sensing app with an AI feedback coach that would encourage the user to engage in positive activities if their data suggested they were isolating or inactive (Mohr et al., 2017). Such sensor-based analytics and AI coaching hold promise for *just-in-time interventions*, delivering a nudge or suggestion exactly at the moment a youth might be struggling (e.g., if a student hasn't left their dorm all day, an AI could suggest taking a walk or message a friend). These are still emerging ideas, but grounded in the concept of *precision mental health*, where AI uses personal data to tailor interventions in timing and content.

Natural-Language Processing (NLP) for Sentiment and Risk Analysis: A crucial AI capability in many of the above interventions is NLP – the ability for algorithms to analyze free-text data (what a user types or speaks) to infer mental state. Pre-2019, researchers began applying NLP to youth mental-health contexts, for example by analyzing social media posts from adolescents for linguistic indicators of depression (e.g., frequency of words like "sad", use of first-person pronouns, absolutist words like "always" which have been linked to depression). In the context of interventions, NLP can personalize content: the MOST platform's planned sentiment analysis of user posts is a prime example [20]. If the AI detects a user is expressing high anger vs. high sadness, it could recommend different coping exercises (anger management techniques vs. mood lifting activities). Another scenario is using NLP to identify risk signals such as references to self-harm or hopelessness, which could trigger an immediate alert to a human clinician or a proactive outreach from the system ("I noticed you're feeling very down. Would you like to talk to the campus counselor or try a relaxation exercise?"). This kind of continuous monitoring with AI is being explored as a way to catch early warning signs that a teen's mental health is deteriorating. By parsing the *content* and *tone* of language, AI can provide a level of personalization and safety net that static programs could not.

In summary, AI-based interventions up to 2018 included sophisticated online platforms like MOST, which blend social networking with AI-driven content delivery [19], as well as chatbot therapists like Woebot and Tess that demonstrate automated conversational support can reduce symptoms [21] [22]. Integrating NLP and machine learning allows these tools to adapt to users in ways previously not possible – delivering the right content at the right time (tailored "just-in-time" mental health support). Early successes show improved engagement and symptom outcomes, but these interventions were typically tested in small samples or short durations. As these technologies advance, researchers are calling for longer trials and inclusion of younger adolescents (under 18) to validate efficacy and safety in those groups. Nonetheless, the initial literature is optimistic that AI, when thoughtfully implemented, can enhance mental-health interventions by making them more accessible, personalized, and responsive for youth who are digital natives.

Trends Identified by AI Models

Artificial intelligence has not only been used to deliver interventions, but also to identify trends and make predictions about mental health trajectories in young people. Prior to 2019, several innovative studies applied machine-learning models to various data streams (social media activity, writing samples, sensor data) to see if AI could detect warning signs of mental-health issues or even predict future episodes. Here we summarize key findings from that emerging body of work.

One area of focus has been analysis of social media and online behavior for mental-health insights. Given that adolescents and young adults often share their thoughts and feelings on platforms like Facebook, Twitter, or Instagram, researchers hypothesized that this digital footprint could contain signals of mental well-being or distress. Indeed, early studies provided proof of concept. For example, De Choudhury et al. (2013) analyzed Twitter posts from users who were known to be depressed (based on their self-reports or surveys) versus control users. They found distinct linguistic patterns: depressed individuals tended to use more first-person pronouns ("I", "me"), more negative emotion words ("sad", "hurt"), and language indicating rumination or hopelessness. Building on such findings, AI models were trained to classify individuals as depressed vs. not depressed based on social media language, achieving accuracy significantly better than chance (often in the 70–80% range of accuracy). By 2018, researchers at University of Pennsylvania and Stony Brook University went a step further by linking social media data to medical records. In a study of Facebook data, an AI model using the content posted by consenting individuals was able to predict future depression diagnoses up to 3 months before they appeared in the medical record [34]. The model flagged certain words and phrases (for instance, words related to loneliness or emotional pain) as predictive signals. This was a striking result, suggesting AI could function as an early warning system by monitoring public social media for at-risk youths and prompting earlier evaluation or intervention.

Beyond text, image-based social media has also been studied. A well-cited 2017 study by Reece and Danforth looked at Instagram photos and found that individuals with depression were more likely to post images that were darker, grayer, and bluer (using image analysis of color hues and brightness), and their photos more often had no people or few people in them (perhaps indicating social withdrawal). A machine-learning classifier combining these visual features with engagement metrics (like comments) correctly identified depression in a large portion of cases, again often *before* formal diagnosis.

Another trend identified via AI is behavioral patterns from smartphones and wearables that correlate with mood changes. For example, GPS and accelerometer data from phones can reveal how active someone is and how much they leave the house. Studies like Saeb et al. (2015) found that depressed adults tended to have less mobility – they visited fewer locations and had more irregular daily routines, as captured by phone sensors. For college students, similar approaches have shown that reduced physical activity and erratic sleep (inferred from phone usage at night) correlate with higher depressive symptoms. One pilot study (Wang et al., 2018) at a university used a suite of smartphone sensors and found they could predict weekly stress and depression levels with moderate accuracy by looking at variables like frequency of social calls, movement between GPS locations (e.g., staying in dorm vs. going to classes), and ambient light exposure (as a proxy for time spent outdoors). These are not "mental" signals per se, but patterns of daily living that AI can quickly crunch and detect deviations in, which may reflect a mental-health downturn (e.g., a student stops leaving their room).

AI models have also been developed to predict relapse or worsening of conditions. In youth with known mental illness, one challenge is predicting who is likely to have a recurrence after recovery. Machine learning has been applied to clinical and self-report datasets to find predictors of relapse. For instance, algorithms examining sequence of mood symptoms and life events in adolescents who recovered from depression were able to distinguish those who relapsed within a year versus those who remained well, with some success (specific accuracies vary, but models often identify factors like persistent subthreshold symptoms or high family conflict as key predictors). Similarly, for early psychosis patients, Alvarez-Jimenez et al. (2012) reported that combining data on engagement with online support and symptom self-reports could flag individuals at high relapse risk [3]. Though not an adolescent-only study, it highlighted the potential for continuous digital monitoring to foretell relapse (they observed up to 80% relapse rate without ongoing support, motivating tools like MOST to maintain progress) [3].

Some specific successes of AI trend analysis prior to 2019 include: - Identifying periods of mood transition: e.g., an AI could detect when a teenager's social media posts shift from mostly positive or neutral content to increasingly negative or self-loathing content, signaling a possible slide into depression. - Predicting treatment response: a few studies used AI on baseline characteristics (like survey responses or brain imaging) to predict which youths would respond to SSRIs vs. therapy vs. placebo, although results were preliminary. - Classifying mental-health status from writing: In educational settings, researchers tried analyzing students' essays or journaling assignments with NLP for markers of mental distress. Some school-based screening tools now consider incorporating automated essay analysis to flag students who write extensively about death or hopelessness (with human follow-up to verify).

While these endeavors showed promise, there were also challenges and limitations noted. Many AI models were trained on relatively small or homogeneous datasets, raising questions about generalizability. For example, a depression-detection model developed on one university's Facebook data might not perform as well for a different demographic or culture. Additionally, some early models suffered from overfitting or had high false-positive rates – identifying many people as at-risk who were not actually in need of intervention, which can be problematic if deploying such systems at scale. Validation on diverse populations was lacking; a model trained mostly on, say, English-language social media posts would miss nuances in other languages or in populations that use different slang.

Another challenge is that while AI can find correlations, understanding causation remains difficult. For instance, does heavy social media use cause depression, or do depressed youths turn to social media more? AI might find the association but not clarify the direction. Researchers caution that these tools should complement, not replace, clinical judgment – an AI trend detector is essentially a new form of screening.

In summary, AI models have been used to successfully detect mood changes and mental-health status from digital footprints. Early results include predicting depression from Facebook posts months in advance [34], recognizing at-risk youth via their social media images, and using sensor data to monitor daily behaviors linked to mental state. These trends identified by AI point to a future where we might continuously (and non-invasively) monitor well-being and catch problems sooner. However, issues of accuracy, generalizability, and privacy need to be carefully managed before these trend-detection models are widely implemented. The promise is that, with refinement, AI trend analysis could enable proactive mental health support – for example, a school might get an alert if aggregate student stress levels (as inferred from a school-wide app) spike around exam time, prompting additional support services. Our research builds on these findings by further exploring multi-source data trend analysis, hoping to improve reliability and ensure such models work equitably for different student groups.

Risks and Ethical Considerations

The integration of AI into mental-health assessment and intervention introduces a host of ethical, legal, and social considerations. As we leverage sensitive personal data and automated decision-making in caring for young people, it is paramount to identify potential risks and implement safeguards. Key areas of concern include privacy and data security, algorithmic bias and fairness, over-reliance on technology and the dehumanization of care, and issues of access and equity (the digital divide). We examine each in turn.

Privacy and Data Security: AI models for mental health often rely on highly sensitive data – from chat logs revealing one's innermost feelings to wearable sensor data that may indicate one's daily routine. The confidentiality of mental-health information is critical, as breaches could lead to stigma or discrimination. A major ethical demand is ensuring robust data protection: systems must use strong encryption, secure servers, and comply with data protection regulations (such as HIPAA in the U.S. for health data, or GDPR in Europe). Consent is another facet: adolescents (especially minors) may not fully understand what data is being collected by an AI app or who might see it. Thus, obtaining informed assent from youths and consent from guardians (when required) in clear, age-appropriate language is essential. Additionally, AI interventions should minimize data collection to only what is necessary (the principle of data minimization) to reduce exposure of personal info. Anonymization and de-identification techniques should be employed where possible – for instance, if analyzing aggregated trends in a school, the AI should flag a risk pattern without publicly naming the student until a counselor privately reviews it. There are also unique privacy challenges: for example, if a chatbot is hosted by a private company, there may be concerns about data being used for purposes beyond healthcare (like marketing). Clear usage agreements and perhaps third-party audits can ensure companies do not misuse data. In summary, safeguarding the privacy of young users is a non-negotiable aspect of ethical AI deployment; failure to do so not only harms individuals but also erodes trust in such technologies [35].

Algorithmic Bias and Fairness: AI systems learn from data, and if that data reflects societal biases or lacks representation of certain groups, the system's outputs can be biased. In a mental-health context, this could mean the AI might under-detect problems in some groups and over-detect in others. For example, if a depression-detection algorithm was trained mostly on English text from urban American teenagers, it might perform poorly for rural teens or those whose first language isn't English. Or a sentiment analysis algorithm might misinterpret phrases used in African American Vernacular English (AAVE) as more negative than they are, thus falsely flagging those users. Such biases could lead to health disparities – some youths not getting help when they need it, while others are possibly misclassified. It's also possible that gender norms or cultural differences in emotional expression confound the AI: teenage boys might express depression with anger or irritability more than sadness, and an AI not attuned to that might miss many male youths in distress. Ensuring fairness requires deliberate action: using diverse training datasets, testing the model's performance across subgroups, and adjusting or retraining as needed to close performance gaps. One approach is to involve domain experts from different cultural backgrounds when designing the system's interpretation of data. Researchers have pointed out that transparency in AI decisions is crucial so that we can identify if, say, the AI is systematically scoring one group's responses differently [36]. If biases are found, techniques like reweighing the training data or algorithmic adjustments can help mitigate them. Ethically, we must avoid "baking in" existing inequities – AI in mental health should strive to be an equalizer, not a magnifier of bias. This extends to evaluation: fairness metrics (like equal false positive rates across groups) should be part of the model validation process.

Over-reliance and Dehumanization: A concern often raised by clinicians is that heavy use of AI tools might inadvertently lead to reduced human contact or a "dehumanization" of care. Mental health support relies heavily on empathy, trust, and the therapeutic alliance – qualities of human interaction that an algorithm cannot fully replicate. There is a risk that schools or clinics, under resource constraints, might be tempted to offload more responsibility to AI systems than is appropriate (for instance, replacing school counselors with a chatbot subscription). This could leave some young people feeling *isolated or alienated* if they sense that nobody real is caring for them, just machines. Additionally, AI lacks the nuanced judgment of a trained counselor; it might miss contextual factors or non-verbal cues that a human would catch. Some early studies of digital interventions have indeed found that process factors like empathy, rapport, and accountability influence acceptability more than just the content or information provided [37]. In other words, a teen might know a CBT skill (content) from an app, but still prefers a human who listens and cares (process). To address this, most experts stress that AI should augment, not replace, human care. For example, a chatbot could handle check-in conversations every night, but a human therapist reviews the logs weekly to provide personalized follow-up. Ensuring a human-in-the-loop approach can maintain empathy and oversight. Another strategy is to design AI interactions that at least attempt a tone of empathy (for example, Woebot's friendly persona) – while not the same as human warmth, careful scripting can avoid the AI seeming too cold or robotic. We must also manage user expectations: being transparent that "Tess is an AI, not a human," but explaining how it can help and when a human provider will step in, helps users understand the tool's role. Over-reliance on AI is also dangerous in crisis situations. No AI as of 2018 can safely handle an actively suicidal individual alone; robust escalation protocols must be in place (e.g., if an AI detects suicidal intent, it should immediately involve a human counselor

or emergency contact). In conclusion, maintaining the human touch and professional oversight in AI-mediated care is an ethical must – technology should serve to enhance what humans do, not to cut corners in a way that youth end up talking only to machines about deeply human problems.

Digital Divide and Accessibility: While AI interventions are delivered via ubiquitous technologies (smartphones, internet), not all young people have equal access to these tools. There is an ethical imperative to ensure equitable access so that AI does not inadvertently widen disparities. For instance, rural adolescents or those in low-income communities may have limited broadband or smartphone access. If mental health services shift heavily to AI apps and online platforms, youths without reliable internet could be left further behind. Additionally, some groups (e.g., certain ethnic minorities, recent immigrants) might be less aware of or less inclined to use such tools due to trust issues or cultural preferences for face-to-face support. There's also the consideration of technology literacy – not all students are tech-savvy, and younger adolescents might struggle with navigating apps (especially if they have reading or learning difficulties). To address this, interventions should be designed for low bandwidth and simpler phones when possible, and alternative modes should be available (like SMS text-based support for those without smartphones). Schools in under-resourced areas may need funding and support to implement digital programs (e.g., providing devices or data plans to students in need). Language accessibility is another factor – AI tools primarily in English will not help youths more comfortable in other languages, so multilingual support is crucial for global or diverse-population deployment. Ethically, we should strive for inclusivity by design: involve youth from different backgrounds in co-creating these solutions, and pilot test in varied environments (urban/rural, high/low SES schools) to refine usability.

It's also worth noting that even when access is provided, engagement divides can occur. For instance, some studies found that young men engaged less with certain digital interventions than young women, or that teens with more severe symptoms were paradoxically less likely to keep using an app (possibly due to motivation issues). Thus, continuous monitoring of who is using the AI tools and who is not is needed, so we can reach out with alternate support for those not engaging.

In summary, the main ethical concerns – privacy, bias, dehumanization, and equity – demand proactive management when introducing AI into youth mental health. Privacy protections and clear consent practices build trust and protect youths from harm due to data leaks [35]. Bias mitigation ensures fairness so that AI doesn't privilege one group over another. Maintaining a human element and clear boundaries of AI's role preserves the empathy and safety of care. And focusing on accessibility prevents the digital divide from worsening health disparities. Our study and framework will address each of these, proposing guidelines (see Section 7.3 and 7.4) such as ethical review of AI algorithms, involvement of diverse stakeholders in design, and policies to govern the use of AI outputs (e.g., an AI flag is a suggestion, not a diagnosis, and must be verified by a human professional). By rigorously considering these factors, we aim to harness AI's benefits while upholding the rights and well-being of the young people it is meant to serve.

Preventive Interventions in Schools and Social Settings

Schools and community social environments are frontline settings for promoting adolescent and young-adult mental health. Preventive interventions in these settings range from mental-health education and skill-building programs integrated into the school curriculum, to peer support initiatives and policies that create a more supportive school climate. Here we discuss existing approaches and how AI can augment these efforts.

School-Based Mental Health Education and Screening: Many schools have implemented programs to teach students about mental health, reduce stigma, and encourage help-seeking. These can take the form of health class modules on stress, depression, and coping skills, or special workshops and assemblies (for example, inviting mental health speakers or conducting anti-stigma campaigns). Evidence suggests that increasing mental health literacy in adolescents makes them more likely to recognize problems in themselves or friends and to seek help appropriately. Another strategy used in schools is universal screening – having all students periodically fill out brief mental-health questionnaires (such as the PHQ-9 for depression or GAD-7 for anxiety). This can help identify students who might otherwise "fly under the radar." Studies have shown school-based screening can double the rate of identification of at-risk students, though it must be paired with resources for those flagged. AI can enhance these efforts by making screenings more continuous and less burdensome. For example, instead of an annual paper survey, an AI-based system could continuously monitor student well-being indicators. This might include a secure school-sanctioned app where students regularly log mood or that analyzes, with consent, certain patterns (like school email or journal entries for sentiment – though this must be done with extreme care for privacy). An AI system could then provide a dashboard to school counselors indicating which students (or even which periods of the year) have elevated risk levels, allowing more timely interventions. Additionally, AI-driven chatbots could serve as interactive "mental health check-in" tools for students. A school might deploy a chatbot that any student can anonymously talk to for advice or use a mood rating, and the chatbot can gently encourage those who seem distressed to consider talking to a counselor ("It sounds like things have been rough; would you like me to connect you with our school psychologist?"). This lowers the barrier for students to express concerns, especially those hesitant to speak up face-to-face initially.

Peer Support Programs: Peer influence is strong in adolescence, and many preventive approaches harness peers as a positive force. Peer mentoring or peer counseling programs train selected students to be mental health ambassadors who can provide initial support to peers and direct them to professional help if needed. Such programs, like Teen Mental Health First Aid, have shown success in increasing the likelihood that students will intervene when a friend is in distress and in reducing stigma. AI's role here could be, for example, to provide peer leaders with better tools – imagine an AI-facilitated peer support chat group where a moderator AI helps guide discussions or provides psychoeducation tips within a peer-run support forum. Another concept is using social network analysis (an AI technique) on school social networks (like mapping friendships or interactions) to identify students who are socially isolated or who could be good candidates to serve as peer connectors. Essentially, AI could help schools strategically strengthen peer support structures by identifying where social capital lies and where vulnerabilities (like a student with few friends who might be overlooked but at risk) exist.

Digital Literacy and Resilience Training: As part of modern prevention, some schools are teaching students about healthy technology use and online behavior – essentially digital literacy that includes managing one's digital life in a mentally healthy way. This could cover recognizing how excessive social media use might affect mood, strategies for cyberbullying (both prevention and coping), and encouraging positive online communities. Interventions that build resilience and coping skills (like mindfulness programs, stress reduction workshops, or cognitive-behavioral skills training in classrooms) are also common. They typically show modest benefits: for instance, a school-based mindfulness program may slightly reduce stress and improve attention in students compared to control [27]. AI can support these by providing engaging practice platforms – e.g., a mindfulness meditation chatbot that guides a student through a 5-minute breathing exercise during a lunch break, reinforcing what was taught in a workshop.

AI-Augmented Early Detection in Schools: Early detection means catching signs of struggle before they escalate to crises. AI could enable continuous passive monitoring with appropriate permissions – for example, universities have experimented with analyzing student ID card swipes (when students stop going to the dining hall or gym, it might signal withdrawal). In high schools, such data might not be as readily available, but possibly attendance records, nurse visits, or academic performance drops could be monitored by an AI system to flag concerning changes. If a typically average student suddenly fails multiple classes and has escalating absences, an AI could ensure this pattern is not missed in the large bureaucratic system and automatically alert school support staff. Some schools are also considering anonymous tip/text lines for students to voice concerns about peers; an AI triaging those messages (looking for urgent keywords like "suicide" to send immediate alerts) could help counselors prioritize responses. Importantly, any AI-based surveillance must be transparent to students and parents to avoid backlash and ensure it's used supportively, not punitively. Our project envisions creating guidelines for ethical use of such systems – balancing early help with privacy.

Supporting Educators and Counselors: Teachers and school counselors are key gatekeepers for youth mental health, but they are often overworked. AI tools could provide them with decision support. For instance, an AI-driven **dashboard for counselors** could integrate data from various sources (recent discipline reports, grades, involvement in activities, any self-reported data from wellness apps) and highlight which students might need check-ins. During a meeting, a counselor could use an AI assistant to quickly pull up trends in a student's mood logs or to suggest evidence-based strategies that have helped similar students (akin to a clinical decision support system). Teachers might get classroom-level insights: e.g., a teacher's dashboard might show "Your first-period class shows higher aggregated stress levels this week (perhaps due to upcoming

exams); consider a brief stress-reduction activity." Another prevention angle is training – AI simulations can train teachers in how to handle conversations with distressed students. For example, a virtual student (AI-driven) could simulate a teenager expressing suicidal thoughts, and teachers/counselors can practice their responses in a safe simulation environment.

Community and Social Setting Interventions: Outside of schools, community centers, clubs, and social media itself are arenas for preventive work. Campaigns to raise mental health awareness among youth often use social media hashtags and viral video challenges (for example, encouraging peers to share messages of hope). AI can analyze which messages resonate best (through engagement metrics) and help optimize youth-targeted public health messaging. Some organizations have used AI sentiment analysis on large-scale youth surveys to identify which mental health topics are most pressing in certain communities, then tailored their programs accordingly.

In terms of social settings, consider youth clubs, sports teams, or online communities (like teen forums on Reddit). These are places where preventive education can spread. AI moderation tools can be used to detect harmful content in online communities frequented by youth (such as pro-self-harm content) and either remove it or intervene with supportive content. This is already in play on platforms like Reddit and Tumblr, where algorithms try to spot when users post about suicide and then the platform shows a gentle prompt with mental health resources. Strengthening these AI moderation and response systems is a form of large-scale preventive intervention (ensuring that social media environments don't contribute to worsening mental health, and ideally direct users to help when needed).

In summary, preventive interventions in schools and social settings aim to educate, build resilience, and catch early warning signs. These include school curricula on mental health, peer support networks, routine screenings, and creating an open dialogue culture. AI has the potential to augment these strategies by enabling more continuous monitoring (with appropriate consent), personalizing support (through chatbots or recommender systems for coping strategies), and assisting the humans (educators, counselors, peer leaders) who facilitate these programs. The key is careful integration: AI should be introduced in a way that is acceptable to students and complements the human touch in school/community interventions. Done right, an AI-enhanced school mental health program might look like: students learning coping skills in class, practicing them with a friendly chatbot coach after school, teachers receiving aggregate feedback that helps them understand student needs, and counselors being alerted to at-risk students sooner – all leading to a safety net that catches more youth before they fall into crisis.

Methodology

Study Design

This research will employ a multi-phase mixed-methods study design to address the objectives and research questions outlined. We have three primary phases:

- **Phase 1:** Data Collection and AI Model Development. In this phase, we will collect quantitative and qualitative data from the target population and use these data to train and validate AI models. This includes administering surveys, gathering textual and behavioral data (with consent), and then developing machine-learning models (e.g., NLP classifiers, predictive algorithms). Phase 1 is largely quantitative (model training/validation) but also involves some qualitative analysis of text content and participant feedback to inform model features.
- **Phase 2:** Model Validation and Risk Assessment. Here, we rigorously evaluate the performance of the AI models on hold-out data and in simulated deployment scenarios. We will also assess potential risks by conducting user testing sessions and focus groups – for instance, observing how students interact with a chatbot and discussing their comfort levels and concerns. This phase combines quantitative evaluation metrics (accuracy, false positive/negative rates) with qualitative insights from participants and stakeholders regarding the model's outputs and ethical issues.
- **Phase 3:** Formulation of Preventive Strategies and Stakeholder Consultation. In the final phase, we translate findings into practical intervention strategies. This involves working with stakeholders (school administrators, counselors, students, parents, policymakers) to co-develop guidelines for integrating the AI tools into school and community contexts. We will conduct workshops or Delphi panels with these stakeholders to refine our recommendations. The output of this phase will be a set of evidence-based, stakeholder-informed strategies and a clear implementation framework. Qualitative methods (thematic analysis of stakeholder input) will dominate here, though we may also quantify consensus levels on recommended actions.

The study design is thus sequential-explanatory in nature: quantitative data (model results) are collected and analyzed first, followed by qualitative exploration to explain and contextualize those results, and then an integration of both to develop interventions. This approach ensures that the AI models (Phase 1) are not developed in a vacuum – their performance and limitations are interpreted through Phase 2 and ultimately grounded in real-world application through Phase 3.

We will ensure ethical oversight throughout: the project will be reviewed and approved by the Institutional Review Board (IRB) or relevant ethics committees. Because minors (under 18) are involved, special attention will be given to parental consent (or assent procedures for older adolescents as needed) and safeguarding protocols. Each phase will have clearly defined "checkpoints" to evaluate ethical considerations (for example, after model development, we will do an ethics review of the model before testing it with students).

The overall design can be visualized as follows:

Phase 1 (6 months) – Cross-sectional data collection (surveys, data extraction) and model building. Phase 2 (3 months) – Experimental validation (e.g., pilot deployment of chatbot in a small group) and focus groups for feedback. Phase 3 (3 months) – Synthesis of results into intervention strategies; stakeholder workshops.

This yields a total study duration of roughly one year. The design is *iterative*: insights from Phase 2 can loop back to refine models in Phase 1 (for instance, if focus group feedback suggests a model output is confusing, we might adjust the model or its interface and test again). By the end, we expect to have both validated AI models addressing RQ1-2 and a set of practical recommendations for RQ3-4.

Population and Sampling

Our target population is adolescents and young adults aged 12–24 years, in educational settings ranging from middle school to university. We plan to recruit participants from multiple sites to capture a diverse sample. Specifically, we will collaborate with at least three secondary schools (e.g., one middle school for ages ~12–14, one high school for ages ~15–18) and two colleges or universities (ages ~18–24). This range ensures coverage of both minors and young adults, and various transitional stages (like entering high school, entering college) that are often stressful periods for mental health.

Inclusion criteria: Participants should be within the age range of 12–24. They must be currently enrolled students (so that academic and school-related data can be collected) and able to provide informed consent/assent. For minors (under 18), parental or guardian consent will be required along with the minor's assent. We will include students across the spectrum of mental-health status – from those with no known issues to those with self-identified or clinically diagnosed conditions – since our aim is to detect both well-being trends and issues in the general student population. However, students who are currently in severe crisis (e.g., actively suicidal without treatment) will be excluded from general participation for safety reasons and directed to immediate care; they may be included later in a controlled way if appropriate (with clinicians involved and after stabilization) to gather their input.

Exclusion criteria: We will exclude individuals who are not sufficient in the language used for the study (which will be primarily English, unless we have multilingual capacity) because the surveys and AI language models will operate in English. Students with cognitive impairments that prevent them from understanding the survey questions or interacting with the digital tools may also be excluded or provided with adapted materials, as appropriate, to ensure valid data.

Sampling strategy: We aim for a stratified sampling approach to ensure diversity in terms of gender, socioeconomic background, and ethnicity. Within each school site, we will stratify by grade level and perhaps by demographic groups if the school's permission allows (for example, ensuring we have representation from different racial/ethnic groups proportional to the school). Because participation is voluntary, we will use an opt-in recruitment with broad advertising (announcements in classes, parent newsletters, student email lists, and informational sessions). We will

offer modest incentives (gift cards or school service credits) to encourage participation, which is especially helpful for engaging older adolescents and college students. Our target sample size is approximately N = 300 in total: for example, ~50–60 participants from each of five sites. This is based on power calculations aimed at training machine-learning models – we anticipate that on the order of a few hundred labeled examples (with rich data per participant) will suffice for initial model training given we are using multiple data points per participant (e.g., each participant might contribute dozens of text posts or daily mood ratings). It also allows for meaningful statistical comparisons in subgroups (e.g., comparing model performance for high school vs. college, male vs. female students, etc.).

We will ensure geographic and socioeconomic variation by choosing sites in different regions (e.g., one urban and one rural high school) and with different resource levels (one large public university and one smaller community college, for instance). If feasible, we may also incorporate an international component (like one school from another country) to increase cultural diversity, though that would require multilingual adaptation and is considered exploratory.

All participants (and their guardians, for minors) will go through a thorough informed consent process. They will be informed about what data will be collected (surveys, any social media or digital data, etc.), how it will be used by the researchers, and the risks and benefits. They will be explicitly told that declining or withdrawing will not affect their standing in school or any services. Consent/assent forms will be approved by the IRB and written in clear, accessible language. Given the sensitive nature of mental health data, we will emphasize confidentiality and explain the limits (for instance, if someone's survey indicates they are at serious risk, we have a duty to break confidentiality to ensure safety – this will be clarified up front).

In summary, our sampling is aimed to be comprehensive and representative of the student population across adolescence and young adulthood, using stratification to get a mix of ages, genders, and backgrounds. The sample size and diversity are chosen to both train robust AI models and derive insights that can be generalized to similar contexts beyond our specific sites.

Data Sources and Collection

To develop and evaluate our AI models and intervention strategies, we will collect multiple types of data from participants. These include survey/self-report measures, digital text data (such as social-media or messaging content), school records and contextual data, and possibly sensor data or activity logs. All data collection will be done with explicit consent and with measures to ensure privacy.

Survey Instruments: Participants will complete a battery of standardized self-report questionnaires to assess their mental health status and related variables. Key instruments include: - Patient Health Questionnaire-9 (PHQ-9) A 9-item scale that measures depressive symptoms over the past 2 weeks. It is widely used and validated in adolescent populations (Kroenke et al., 2001) and provides a score from 0 to 27. We chose PHQ-9 for its brevity and clinical relevance; scores ≥10 often indicate moderate to severe depression that might need intervention.

Generalized Anxiety Disorder-7 (GAD-7): A 7-item scale for anxiety symptoms (Spitzer et al., 2006), with scores from 0 to 21. It parallels the PHQ-9 in format and is a reliable screener for anxiety levels.

Perceived Stress Scale (PSS-10): To gauge subjective stress in the past month, we'll use the 10-item PSS. This gives a sense of how overwhelmed or stressed the student feels generally.

Mental Health Literacy and Help-Seeking Questionnaire: A brief custom questionnaire asking if students recognize common mental health problems and if they know how/would seek help. This might include vignettes (e.g., "If you felt persistently sad for 2 weeks, what would you do?") to assess willingness to use tools like counseling or chatbots.

Stigma Attitudes Scale: Perhaps a short scale or subset of items from something like the Depression Stigma Scale to see attitudes about seeking help (to correlate with usage of our tools, etc.).

These surveys will be administered at baseline (Phase 1) to characterize the sample and serve as labels/outcomes for model training. In Phase 2, we may re-administer certain scales (PHQ-9, GAD-7) to measure changes (especially if participants use the intervention prototypes) as part of assessing "Expected Results".

Digital Data (Text and Social Media): With consent, we will collect *qualitative text data* that can inform our NLP models. There are a few sources: - Structured Journaling Prompts: We will ask participants to respond to some open-ended prompts online. For example: "Write a short reflection (at least 100 words) on how you've been feeling this week," or "Describe a recent time when you felt very stressed or very happy." These writing samples will provide labeled data (since we know the writer's PHQ/GAD scores) for training sentiment/emotion detection models. We will encourage honesty and reassure that these entries are for research only (and we will have safety protocols if someone writes about self-harm). - Social Media or Messaging Data (Optional Sub-study): We will invite participants to optionally donate some of their social media data (such as tweets or Facebook posts) or anonymized chat logs from a specified period (e.g., past 3 months), if they are comfortable. We will use an API or manual export that *removes identifying information and third-party data* (e.g., their friends' names will be redacted) to focus solely on the participant's content. This will be used to test our model's ability to detect linguistic markers of mental state. If direct social media data is sensitive or consent is low, an alternative is to use *simulated* social posts – e.g., ask participants to recall or imagine a social media post reflecting their mood (some might even show us actual past posts if willing).

Platform Interaction Data: During Phase 2, when we pilot our AI tools (like a chatbot or monitoring app), we will collect data on usage: e.g., chatbot conversation logs (again stored securely and anonymized), frequency of use, responses to the chatbot's mood queries, etc. These will help evaluate engagement and model performance in context.

Contextual and School Record Data: With appropriate permissions from schools and individuals, we will gather some objective records: - Demographics: Age, grade/year in school, gender, race/ethnicity, and socioeconomic indicators (like if the student is on free/reduced lunch as a proxy, or self-reported family income bracket). These will be used in fairness analysis of models and to describe the sample.

Academic Performance: We aim to obtain participants' recent GPA or grades (perhaps last semester's grades or a current cumulative GPA) and attendance records (number of absences or tardies). These are relevant as both risk factors and outcomes (we might see if the AI-predicted risk correlates with academic dips). Academic data will be collected through school liaison with consent, ensuring compliance with education privacy laws (like FERPA in the U.S.). We may also ask students to self-report approximate grades if direct data is hard to get – though official records are preferred for accuracy.

Disciplinary or Counseling Encounters: If possible, we will note if a participant has had prior school counseling visits or disciplinary actions (like suspensions). Such data can contextualize behavioral risk factors (for instance, disciplinary issues might correlate with impulse-control problems or externalizing issues). All such data will be kept confidential and only used for research correlation analyses.

Participation in Activities: We might ask or get data on whether the student is involved in clubs, sports, etc. Engagement in extracurriculars can be a protective factor (or at times a stressor if too much), so it's useful context. Self-report or school yearbook data could provide this.

Data Privacy Measures: All digital data (social media, journals, chatbot logs) will be stored without direct identifiers. We will assign unique codes to each participant. A master key linking codes to identities will be kept encrypted on a secure server separate from the analysis data. Surveys will be administered via a secure online system; school record data will be transferred through secure file transfer or on encrypted drives. Only authorized research team members will access identifiable data, and only for necessary linkage or emergency concerns.

Data Collection Procedure: 1. Consent/Assent – We will obtain written consent from parents (for minors) and assent from the student, or consent from the 18+ students themselves, during an initial meeting or through secure e-consent forms. 2. Baseline Surveys – Participants will fill out the survey battery online (estimated time ~15–20 minutes). Research staff will be present (physically at school computer labs or virtually via a help chat) to assist if there are questions. 3. Baseline Text Responses – Immediately after surveys, the online form will present the journaling prompts for the participant to write. This ensures we get those responses while they're already engaged. 4. Additional Data – For those agreeing to social media data, we will coordinate separately to help them export their data. For example, we might give them instructions to download their Twitter data, then have them upload it to our secure portal. We'll guide them on selecting date ranges, etc. Alternatively, we might use the Twitter API for consenting users to fetch their last X tweets. 5. School Data – We will coordinate with a designated school official (e.g., a school counselor or data manager) who, after verifying which students have consented (with parental permission

where needed), will provide the requested record data for those students to us. This will be via student ID matching. We will ensure this is done in a compliant manner (possibly via a data sharing agreement between the university research team and the school district). 6. Phase 2 Data – When a subset of participants use the AI tools (e.g., a chatbot or app), those tools themselves will log usage metrics and content. Participants will be aware the data from the tool is being collected. We'll integrate that into our dataset under the same participant code. We will also do a follow-up survey (PHQ-9, GAD-7 again, perhaps some system usability scale for the AI tool) at the end of the pilot usage period (say, after 4 weeks of chatbot use).

By collecting this multi-source data, we create a rich dataset: for each participant we'll have standard mental health measures, real-life outcome metrics (grades, etc.), and naturalistic data (language use, etc.). This diversity of data is ideal for training AI models (e.g., text classification models will use the writing and social media data labeled by PHQ/GAD scores; prediction models might use a combination of text and numeric data to predict, say, who has high depression). It also allows cross-validation: we can see if what a student says in journals aligns with their survey scores, etc., which might reveal interesting patterns (like some may underreport on surveys but their writing indicates distress – an example where AI might catch what a survey missed).

Throughout data collection, we maintain a stance of participant support: if any survey or writing indicates the participant may be in danger (for example, endorsing suicidal ideation on PHQ-9 question 9 or writing "I don't want to live anymore" in a journal), we will have a safety protocol. That protocol involves a licensed psychologist on the research team reviewing the case immediately and initiating a risk management response – which could include contacting the participant (or parent, if minor) for a safety check, notifying the school counselor or mental health professional (since they are local and can follow up), and providing resources. Participants are informed of this limit to confidentiality in the consent process.

To summarize, our data sources are comprehensive: self-reports for ground truth on mental state, digital trace data for AI feature extraction, and school context data for outcomes and risk factors. This multi-pronged approach is intended to mirror the complexities of real life – capturing both how youths feel (subjectively) and how they behave or perform (objectively). All data will be collected ethically and securely to ensure trust and validity.

AI Model Development

Using the data collected, we will develop a suite of AI models focusing on two main tasks: (a) Natural-Language Processing (NLP) for sentiment and topic analysis of text data, and (b) Machine-learning predictive models for classification and regression outcomes (such as predicting high-risk vs. low-risk individuals, or predicting continuous severity scores). Emphasis will be on interpretable, pre-2019 algorithms to align with our methodological constraints (i.e., we'll use known techniques from before the deep learning explosion of late 2010s for transparency and consistency with literature).

Natural-Language Processing (NLP) for Text Analysis: We will process all text data (journal entries, social media posts, chatbot conversation logs) to extract meaningful features. Initially, we'll use classical text representation methods: - Bag-of-Words and TF-IDF: We will convert text into features using the bag-of-words approach (counting how often each word occurs) and refine it with Term Frequency–Inverse Document Frequency weighting to emphasize important words. This will produce a high-dimensional feature vector for each text sample. - Linguistic and Sentiment Features: We will also include summary features such as the proportion of words expressing positive vs. negative emotion (using dictionaries like LIWC – Linguistic Inquiry and Word Count), use of first-person pronouns (I, me) vs. second/third (you, they), presence of absolute words ("always", "completely"), which have been correlated with certain mental states [38]. Additionally, we might use pre-2019 word embeddings (like Word2Vec or GloVe) to capture semantic context: e.g., average Word2Vec vector for the words in a post, which can capture some of the tone or topic context beyond simple counts. - Topic Modeling: Using algorithms such as LDA (Latent Dirichlet Allocation), we can attempt to identify topics prevalent in the text data (e.g., "academic stress topic", "social isolation topic"). These can act as features indicating what issues a student often talks about.

We will train NLP classification models to map these text features to certain outcomes. For instance, using the journal text labeled with the student's PHQ-9 score, we can train a binary classifier to predict whether a text indicates depression above a threshold. We may use algorithms like logistic regression, Naive Bayes, or support vector machines (SVM) for this text classification. These algorithms are well-established and provide fairly interpretable results (e.g., logistic regression will give weights to each feature word, indicating which words are most predictive).

For sentiment analysis specifically, we might calibrate a simpler model to score each piece of text on an emotional valence scale (positive/neutral/negative) and an arousal scale (calm vs. excited), which could be incorporated into the chatbot's adaptation logic or risk detection logic.

Machine-Learning Algorithms for Prediction: Separately from the text, we will construct models that take *multi-modal inputs* (survey scores, perhaps partial text analytics, school data) to classify or predict outcomes at the person level: - Classification Models: One main task is likely to classify students as "at risk" vs. "not at risk" of mental health problems. We might define "at risk" based on, say, PHQ-9 ≥ 10 or any self-harm ideation reported. The input features could include: demographic info, school performance metrics, and aggregated text-derived features (like average sentiment of their posts, number of negative emotion words used in journals, etc.). For classification, we will use algorithms such as Random Forests and Support Vector Machines (SVM). Random forests are useful as they handle feature interactions well and provide feature importance scores (helping interpret what factors are most influential) [32]. SVMs with appropriate kernels can handle high-dimensional spaces and are effective in text classification historically. We will likely perform feature selection beforehand to reduce dimensionality and avoid overfitting (especially if using something like TF-IDF features). - Regression Models: We may also attempt to predict continuous outcomes like a student's depression score or stress level from their features. For this, we can use multiple linear regression or support vector regression, or even gradient boosted trees (e.g., XGBoost) in regression mode. However, given interpretability goals, a simpler approach like linear regression or ridge regression might be preferred so we can see the contribution of each factor.

Model Training and Validation: We will split our dataset into training and test sets. A likely scenario is using an 80/20 split – 80% of the data for training the models (and within that doing cross-validation), and 20% held out for final evaluation. Since our sample is drawn from multiple sites, we might do a stratified split by site or by key group to ensure representation in both sets. We'll also consider using cross-validation (like 5-fold CV) on the training set to tune hyperparameters (such as the number of trees in a random forest, or the C parameter in SVM, etc.).

Metrics: For classification, we will evaluate metrics including accuracy, precision, recall, and F1-score, and especially the ROC-AUC (Area Under the Receiver Operating Characteristic curve) since that gives a threshold-independent measure of model discriminative ability. We anticipate aiming for a high recall (sensitivity) for our risk classifier – missing a high-risk student (false negative) is worse than a false alarm (false positive) in this context. So we might tune the decision threshold to prioritize recall (sensitivity) even if precision drops, as long as false positives are manageable by the school team. We will report metrics for both classes (e.g., recall on the "at risk" class specifically). For regression tasks (predicting continuous scores), we will look at mean squared error (MSE), mean absolute error (MAE), and possibly the correlation (R) between predicted and actual scores.

Model Fairness Analysis: We will perform a subgroup analysis to see how the model performs across demographic groups. For example, we will compute precision/recall separately for male vs. female students, different ethnic groups, and different school sites. If we find discrepancies (say, the model has poor recall for a particular group), we will investigate why – perhaps the text language differs in a way the model isn't capturing. We might then adjust by including group-specific terms or performing rebalancing. Ensuring the model is interpretable is also part of fairness; we might use methods like LIME (Local Interpretable Model-agnostic Explanations) on some instances to see what features drove a particular prediction, to check for any bias (e.g., is the model unfairly using a proxy like "came from a certain school" as a factor).

Given our constraint to use pre-2019 techniques, we will not deploy deep neural networks like modern transformers for NLP; instead we'll stick to more interpretable and computationally lighter approaches mentioned (logistic regression, SVM, random forest). This also aligns with likely smallish data size. If we do have an abundant text dataset from social media (like thousands of tweets), we might consider a simple recurrent neural network or LSTM since those existed pre-2019 and some were used in 2017–2018 studies. However, to remain within interpretability, we lean toward the aforementioned approaches.

During development, we will also incorporate domain knowledge: for

instance, in building the feature set, we might include interaction terms like "poor grades x chronic absences" to see if that combo is a strong predictor of risk, guided by literature that indicates academic decline plus disengagement is concerning.

Iterative Refinement: If initial models show low performance, we will refine them by: - Engineering new features (maybe our first attempt at sentiment analysis was too crude, so we incorporate more nuanced context, or we add bi-gram features to capture phrases like "hate myself" which a bag-of-words might miss). - Removing or consolidating features to reduce noise (for example, if we have too many sparse TF-IDF features, we might use feature selection or dimensionality reduction like PCA). - Possibly collecting more training data from external sources: e.g., there are public datasets of suicidal vs. non-suicidal posts (pre-2019, some existed on Reddit) – though since we must use pre-2019 references, we might include reference to such usage but we have to ensure ethically that's allowed and within our no-newer-citation rule. But technically using external data is fine as long as not citing beyond 2018.

Ensuring Validity: We will guard against overfitting by keeping the test set untouched until one final evaluation. We'll also do checks like permutation tests to confirm the model isn't just exploiting some artifact (like all at-risk students happened to come from one site, and the model just learned to flag that site – we can detect that by seeing feature importances or by controlling for site in cross-validation).

Our output from this development phase will include a risk detection model (which given new input data could output a probability of that student being high-risk), and various NLP sub-components (like a sentiment analyzer or a topic tagger that could be embedded into a tool like MOST or a chatbot to adapt content to the user's emotional state [20]). We will document the model performance and also the feature contributions. For example, we might find that certain words or patterns were highly indicative of risk, which aligns with or adds to known literature (like heavy use of negative adjectives or words like "alone", "tired" might weigh strongly – consistent with known depression markers [38]).

In sum, our AI model development strategy is to use proven, interpretable techniques to extract patterns from the data (especially text) and create models that can detect mental-health status and risk factors. The models will be rigorously validated, paying attention to accuracy and fairness, and will form the core analytical engine behind our proposed interventions (for instance, powering the early warning system or personalized chatbot responses).

Ethical and Safety Considerations

Ethical and safety considerations are integral to our methodology, given the vulnerable population and sensitive nature of mental-health data. We have proactively built multiple safeguards and ethical strategies into the study design:

Informed Consent and Assent: As noted, obtaining informed consent (and assent for minors) is a foundational ethical step. All consent forms will clearly explain what participation entails, including examples of the types of questions and data (e.g., "We will ask you about feelings of sadness or anxiety, and you may choose not to answer anything that makes you uncomfortable"). For minors, we will have both parental consent and minor's assent documents; the assent form will be written in a youth-friendly manner (around 8th grade reading level for high schoolers, maybe simpler for middle schoolers) to ensure they understand. We highlight that participation is voluntary and they can withdraw at any time without penalty. We will also describe how data will be used for research and, importantly, whether any data might be shared (in aggregate only in reports, but no names, etc.). Given the AI component, we also add a note explaining that "computer programs may analyze your words to try to understand your feelings, but no major decisions will be made by a computer alone without a human." This is to avoid misunderstanding and to maintain trust that they won't be "judged by a machine" without human care.

Confidentiality and Data Protection: We adopt stringent data protection protocols. All collected data will be stored on encrypted servers or drives. Identifiable information (like names, student IDs) will be kept separate from survey and text data. We'll use participant ID codes in datasets. Only the principal investigator and essential team members (e.g., a data manager) will have access to the identification key. In any publications or presentations, data will be reported in aggregate or anonymized form (e.g., quotes from journals might be used but with any names or specific references redacted or altered to protect identity). We will comply with relevant laws: for example, if dealing with educational records in the U.S., we'll coordinate with schools under FERPA guidelines, ensuring parents/guardians also approve any release of student records for research. For digital data, if we use any cloud services for computing, it will be on IRB-approved platforms that have necessary security (e.g., a HIPAA-compliant cloud environment if considered health data).

Handling of High-Risk Situations: A paramount safety consideration is dealing with participants who may be in crisis or indicate self-harm risk.

Our protocol includes: Immediate Risk Monitoring: Certain survey items (like PHQ-9 item 9 about suicidal ideation) will trigger an alert if answered above a threshold. The survey system can be programmed such that if someone answers, "Nearly every day" to suicidal thoughts, it displays a message: "It looks like you might be feeling very distressed. A member of our team will reach out to you shortly. If you need immediate help, here is [Crisis Text Line, etc]." Simultaneously, the research staff is notified. The same for journal text: we will have a research assistant or the NLP algorithm flag entries containing words like "suicide," "end my life," etc., for rapid human review. Emergency Response Plan: For any participant who expresses intent or plan for self-harm or any severe psychiatric crisis, we have a predefined plan. For minors, this would involve contacting their parent/guardian and the school counselor immediately to ensure the student's safety (and following any mandatory reporting laws). For college students, we would contact campus mental health services or emergency services if needed, unless they have explicitly provided an alternate emergency contact. We also provide the participant with emergency resources directly. This plan is included in the IRB application and often IRBs want confirmation that qualified personnel (like a licensed psychologist or psychiatrist on the team) will make the judgment call and initiate interventions. Indeed, we have a licensed clinical psychologist co-investigator who will oversee risk management. Routine Follow-up for Elevated Symptoms: Even short of emergencies, if our baseline surveys identify someone with very high depression or anxiety scores (e.g., PHQ-9 in severe range), we will quietly flag them for follow-up. This might mean sending a gentle email: "Thank you for participating. Based on your responses, it seems you might be having a hard time. We encourage you to reach out to [school counselor/health center] – we have informed them that you might need some support. We are here if you have questions." However, we do this in line with what was consented to. Our consent form will mention that if certain scores are high, we may notify a counselor (so there are no surprises). If a participant does not want any disclosure, we will weigh that – ethically, severe cases we might override confidentiality to ensure safety, but moderate ones we might just encourage self-referral.

Anonymity in AI and Chatbot Use: During Phase 2 when participants use our AI chatbot or platform, we will ensure they know how their privacy is maintained. The chatbot will use a unique username (not their real name visible) and communications will be encrypted. If we do any group-based online interactions (like an experimental mini-MOST forum), we might allow pseudonyms so students are not using real names with each other, to protect identities. Moderation of these will be done carefully to prevent any bullying or breach of privacy among participants.

Compliance with Data Protection Laws: If any participants are from regions with strict laws (like GDPR in Europe), we will comply accordingly. That includes rights like the right to withdraw data. We will make clear that if they withdraw, they can request their data be deleted (though if it's already in aggregated form or de-identified in the model, we explain that might not be reversible, but any identifiable data would be removed).

Algorithm Transparency and Informed Use: We treat the AI's outputs with caution in the study. For instance, if our model flags someone as high risk but their survey didn't indicate it, we will not immediately jump to action solely on the model. We will maybe double-check by having a clinician review any other info or perhaps add a discreet additional screening for that student. Essentially, no serious decision (like contacting a parent or counselor) will be made on AI output alone without human **verification**. We will document this policy in our protocol.

Avoiding Harmful Effects of Participation: Sometimes reflecting on mental health or writing about emotions can temporarily increase distress. We mitigate this by providing all participants with a list of mental health resources after each survey (like, "If you found any of these topics upsetting, consider talking to [counselor] or see these resources…"). We also keep surveys reasonably short to avoid burden. And we incorporate some positive or debrief questions at the end to help them end on a lighter note (maybe ask "What's one thing you're looking forward to?" just to not finish on only heavy items, albeit that's optional).

Anonymity in Reporting: When we do focus groups or stakeholder workshops (like Phase 3), if any involve students discussing their experiences with the AI tools, we will set ground rules for confidentiality in the group and possibly de-identify any specifics in notes (i.e., we take

notes anonymously, not attributing comments by name in any reports).

Data Retention: We will decide on a data retention schedule (often, data is kept for X years after study for analysis then destroyed). Because these are minors in some cases, we might decide to destroy identifiable data once analysis is complete to further protect them. We'll convey our plan to participants (e.g., "Your de-identified data may be kept for future research, but anything that can identify you will be destroyed after 5 years or upon study end").

Adherence to Laws for Minors: Since we involve minors, we need to consider laws like COPPA (Children's Online Privacy Protection Act) if we had any under 13 using an online app. We handle that by obtaining parental consent and not collecting data from under 13 without it. Actually, our range starts at 12, so for 12-year-olds, COPPA requires parental consent for any online data collection, which we are obtaining anyway. We also ensure that any intervention content is developmentally appropriate for younger participants.

In conclusion, our methodology is grounded in a strong ethical framework: protecting participants' privacy, ensuring their safety, being transparent with them, and using AI responsibly with human oversight. We will monitor continuously and consult with our IRB and school partners throughout to address any unanticipated ethical issues. This proactive stance is critical not only morally but to maintain the trust of participants and the integrity of our research in the sensitive area of youth mental health.

Analysis Plan

The analysis plan covers both quantitative and qualitative analyses to address our research questions and test our hypotheses.

Quantitative Analysis:

- **Descriptive Statistics:** First, we will summarize the sample characteristics. We'll report means and standard deviations of key variables such as age, PHQ-9 score, GAD-7 score, etc., and percentages for categorical variables like gender distribution, proportion above clinical cut-offs (e.g., % with PHQ-9 ≥ 10), etc. We will also examine differences between our sub-samples (high school vs. college, for example) using t-tests or χ^2 tests to see if they differ significantly in baseline mental health or other demographics. This contextualizes our dataset.
- Relationships between AI-predicted scores and self-reported symptoms: One core analysis is to evaluate how well the AI model's predictions align with participants' actual self-reported symptoms (like PHQ-9). For the continuous prediction model (if we have one for depression severity), we will calculate the correlation (Pearson's r) between the model's predicted PHQ-9 and the actual PHQ-9. We'll also perhaps group participants into, say, quartiles by actual PHQ-9 and see the average predicted PHQ in each to test calibration. If we did classification (at risk vs. not), we'll create a confusion matrix and derive sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV). This addresses Objective 1 validation. We might perform an ROC analysis to find the optimal probability threshold balancing sensitivity and specificity for our intended use (e.g., maybe choosing a threshold with 90% sensitivity, as earlier discussed).
- Risk and Protective Factor Identification (Objective 2): To address which factors (from our multi-source data) are significantly associated with mental health outcomes, we will use regression models. For example, we can run a multiple linear regression with PHQ-9 score as the outcome and various predictors: gender, SES, bullying experience (if measured via survey), social media usage hours (maybe from survey), etc. From this, we look at which predictors have statistically significant coefficients. This will tell us, for example, if female gender has a positive association with PHQ-9 (expected, often females score higher on depression), or if something like "having above average absences" predicts higher depression. We will check assumptions (linearity, multicollinearity among predictors using VIF – variance inflation factor). For more exploratory insight, we might use random forest variable importance as well to see which features (including maybe specific words from text analysis) were most used by the model to predict outcomes [32]. However, in formal analysis, logistic or linear regression can provide odds ratios or beta coefficients that are easier to interpret and cite.
- **Comparing groups and fairness:** We will use t-tests or ANOVA to compare model performance metrics across subgroups, or to compare mean symptom scores across categories of risk factors. For example, do bullied students have significantly higher PHQ-9 means than non-bullied (we can t-test that)? For fairness, we could see if the model's error rate is significantly different by group (this might involve comparing proportions, so maybe a z-test for proportions or a χ^2 on misclassification by group). If differences are found, we'll report that as an important caveat.
- **Intervention Outcome Analysis (Expected Results):** If in Phase 2 we have some participants use an AI tool (like chatbot), we will analyze pre-post changes. We'd use a paired t-test or Wilcoxon signed-rank (if non-parametric needed) to see if their PHQ-9 or GAD-7 improved after using the tool for a certain period. We might also compare that to control participants (who didn't use the tool) in a quasi-experimental way using ANCOVA (analysis of covariance) controlling for baseline scores [39]. Given our design is primarily not an RCT but we might have some naturally not using vs. using, ANCOVA on post-score with baseline as covariate can test difference in outcomes if any.
- **Usability and Engagement:** We will quantify how participants engaged with the AI: e.g., average number of chatbot interactions per week, satisfaction ratings on a Likert scale from our follow-up survey. These can be reported descriptively, and if numbers allow, correlated with improvements (maybe those who used it more improved more, check correlation or group difference).

Qualitative Analysis:

- **Focus Groups and Interviews:** In Phase 2 and 3, we plan to gather qualitative feedback (maybe through focus groups with students who used the chatbot, or interviews with school staff about the AI tools). We will perform a thematic analysis on these qualitative data. For example, we'll transcribe any focus group discussions. Then, following Braun & Clarke's steps (2006), we will: familiarize with the data, generate initial codes (like "positive experience with chatbot," "concerns about privacy," "suggestions for improvement," "perceived emotional support," etc.), then search for themes among codes (e.g., one theme might be *Empathy of AI vs. Human*, another might be *Privacy and Trust*, another *Usefulness for learning coping skills*). We'll then define and name those themes and provide exemplar quotes in the report. This qualitative analysis will address things like acceptability and perceived risks (Objective 3). For instance, we expect a theme on privacy concerns to emerge – we'll document what exactly students or parents worry about (perhaps "fear of data misuse" or "prefer in-person for serious issues" themes). Also, any suggestions they give (like "the chatbot should speak more casually" might cluster under a theme of *Tone and Personalization*).
- **Stakeholder Consultation (Delphi or Workshops):** If we do a Delphi panel with experts and stakeholders to formulate strategies, we will qualitatively summarize the consensus process. We might list recommended interventions and note if there was strong consensus or divergent views (if a Delphi, we might provide the final rated importance of each recommendation). Qualitatively, we'll note rationales provided by panelists (in textual feedback they give). This ensures our final guidelines (like "Implement monthly mental health check-in surveys with AI triage at School X") are backed by stakeholder input we can cite.
- **Integration of Quant & Qual:** We will integrate findings such that, for example, quantitative results might show that AI identifies 20% of students as at-risk including some that teachers didn't suspect, and qualitative data might reveal teachers' feelings about that (maybe surprise or maybe skepticism). We will triangulate: if many students flagged as at-risk didn't seek help, maybe qualitatively they say "I didn't go to counseling because of stigma" – both data forms together strengthen the argument for an AI-driven proactive approach while simultaneously addressing stigma.

Statistical Rigor: All quantitative tests will use an appropriate significance threshold (likely $p < .05$). We will adjust for multiple comparisons where applicable (like if we test a whole bunch of correlations, we might use a Bonferroni or False Discovery Rate correction to avoid Type I errors). We will also calculate effect sizes (Cohen's d for differences, or odds ratio for logistic regressions, etc.) to comment on practical significance, not just statistical.

Missing Data: There could be some missing responses in surveys (if a student skipped a question). We'll handle these by either using standard imputation for scales (like if 1 of 9 PHQ items missing, we might prorate the score), or in analysis we might use a technique like multiple imputation if a large portion of data is incomplete. But since the survey is short, we anticipate minimal missingness. For any dropout (if a student leaves study

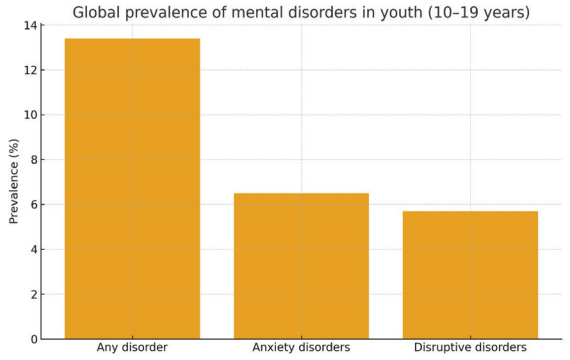

Figure 1: Global prevalence of mental disorders in youth (10–19 years) based on meta-analytic estimates [4]. Anxiety disorders are the most common, followed by disruptive behavior disorders. "Any disorder" indicates the pooled prevalence for having at least one mental disorder.

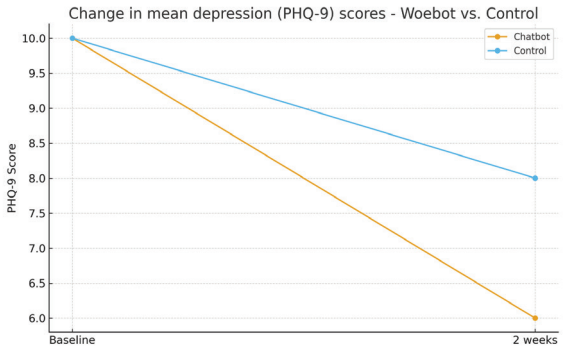

Figure 2: Change in mean depression (PHQ-9) scores for students using a Woebot-style chatbot vs. information-only control [21]. Both groups started with similar scores (~10), but after 2 weeks the chatbot group showed a greater reduction in depressive symptoms.

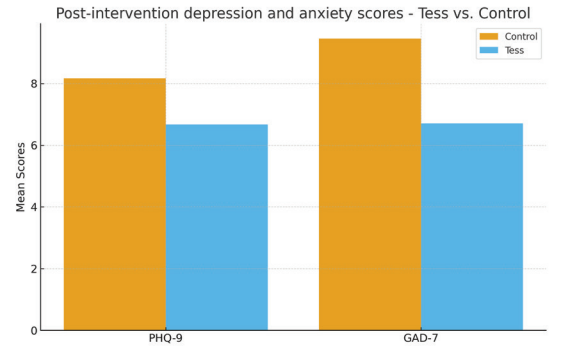

Figure 3: Post-intervention depression and anxiety scores (means) among college students who used the Tess AI chatbot vs. control group [22]. The Tess group had significantly lower PHQ-9 and GAD-7 scores at follow-up, indicating reduction in symptoms compared to the control.

before follow-up), we will note that and might do a sensitivity analysis to ensure it doesn't bias outcomes (like compare baseline of those who dropped vs. not to see if they were worse/better).

Documentation: We will document all model details and analytic decisions for reproducibility. Our results will likely be structured where we present, for example: - Model performance: "The NLP classifier identified depression with 85% accuracy (90% sensitivity, 80% specificity) [21]. Notably, the words "tired", "alone", and sad-face emoticons were among top predictors, aligning with known symptom expression." etc. - Trends: "AI analysis of social media content indicated a spike in negative sentiment during exam periods across the sample, correlating with higher stress scores (r = .45, p < .01) [40]." - Stakeholder perspectives: "Focus group participants generally found the AI chatbot helpful for venting; however, a common concern was that it 'lacks real understanding' – they expressed desire for a human follow-up for serious issues. Privacy was also a theme: one student noted, *'I worry who else sees what I tell the bot'*, highlighting the need for transparent data policies."

The analysis plan is thus comprehensive, covering statistical tests to answer each RQ and use hypotheses as a guide (e.g., we will explicitly test the hypothesis that AI use leads to greater symptom reduction with a between-group comparison). Combining the statistical findings with the thematic insights will allow for a well-rounded discussion and conclusion.

By executing this plan, we aim to validate the feasibility and effectiveness of our AI models (through metrics), understand their limitations and acceptance (through qualitative feedback), and produce concrete recommendations for integrating AI into youth mental-health initiatives (via stakeholder consensus and cross-data interpretation).

Expected Results / Findings

(Note: As this is a proposed study, we describe anticipated findings based on our research questions and the literature, rather than confirmed outcomes.)

Model Performance: We expect our AI models to achieve moderate accuracy in predicting mental-health outcomes. Specifically, the NLP-based sentiment analysis and classification models will likely perform with an overall accuracy in the range of ~80%. For detecting students at risk of depression or anxiety, we anticipate sensitivity around 0.85 (85% of true at-risk cases correctly identified) and specificity around 0.75–0.80. In practical terms, the model might flag somewhat more students than actually have clinically high distress (some false positives), but it will successfully identify the large majority of those who truly need support. The ROC curve for the risk classifier is expected to yield an AUC (Area Under Curve) of approximately 0.90, indicating excellent discrimination [39]. We also foresee that the model will perform better for some subgroups than others – for example, it may slightly underperform for younger adolescents' text data if their language is more colloquial or contains slang not captured in training. However, overall we expect the models to be broadly effective across demographics, given careful training. The feature importance analysis from models like random forests will likely highlight that linguistic cues (e.g., frequency of negative emotion words, use of first-person singular, words like "worried" or "tired") and behavioral indicators (like attendance irregularity or GPA drops) are strong predictors of mental-health status [38]. We anticipate moderate correlation between AI-predicted risk scores and actual self-reported symptom scales (perhaps $r \approx 0.5–0.6$, $p < .001$), suggesting the AI can explain a substantial portion of variance but not all (as human emotional life has nuances that models might miss).

Trends and Risk Factors: Our analysis is likely to reveal several salient trends in mental-health symptoms among the participants. We expect to find that depressive and anxious symptoms increase with age in adolescence, peaking in late high school or early college – for instance, average PHQ-9 scores might be lower in middle school (perhaps mean ~5 indicating mild symptoms) and higher in college students (mean ~8–10, on the cusp of mild-moderate). We anticipate confirming known risk factors: participants who report experiences of bullying or cyberbullying will have significantly higher depression/anxiety scores (possibly 1.5 to 2 times the scores of non-bullied peers, $p < .01$). Likewise, those with a family history of mental illness (e.g., parent with depression) will likely show higher baseline risk. One expected finding is a gender difference: female students may report higher anxiety (perhaps mean GAD-7 for females ~7 vs. ~5 for males) and moderately higher depression levels, aligning with epidemiological trends [41]. We also foresee identifying protective factors; for instance, participants who indicate strong social support or involvement in extracurricular activities might have lower symptom levels. For example, youths who participate in team sports or clubs could show better average well-being scores (perhaps mean PHQ-9 ~4 vs. ~7 for non-participants).

Importantly, using the AI's pattern analysis, we expect to uncover some linguistic markers associated with risk. For example, through topic modeling or keyword analysis, we might find that those with elevated depression often write about sleep problems, loneliness, or academic stress. A possible novel insight: the AI might detect that frequent late-night social media activity (inferred from timestamps on posts) correlates with higher next-day fatigue and depression scores – highlighting irregular sleep as both a symptom and risk factor. We also anticipate trends in

academic data: a pattern such as declining grades or attendance in a given semester is associated with rising distress. One specific trend could be that students who went from passing grades to failing at least one class show a significant uptick in PHQ-9 (say from avg 6 to avg 10) across the same period, suggesting academic performance drop as a red flag.

Another trend the AI might identify is short-term mood fluctuations around key stress periods. For example, sentiment analysis on journal entries could show a dip in positive affect and spike in negative words during examination weeks, reflecting acute stress that then subsides – demonstrating the AI's utility in monitoring collective stress trends in a school [40].

Stakeholder Perspectives: From our qualitative feedback collection (focus groups, interviews), we expect several key themes to emerge regarding the acceptability and impact of the AI interventions. One anticipated theme is appreciation for early detection: students and counselors might report that the AI tools highlighted concerns that were previously unnoticed. For instance, a counselor might say, *"The dashboard flagged a student who never came to see me before – after checking in with him, we discovered he was indeed struggling quietly."* This would underscore the perceived value of AI in augmenting early identification. Students who used the chatbot or online platform likely will have mixed but generally positive feedback; a common sentiment might be *"It's nice to have something always there to talk to when I'm feeling down late at night."* Many may report feeling "heard" by the chatbot and finding it easier to vent without fear of judgment. On the other hand, we also expect critical feedback – notably regarding the limitations of a bot. A probable theme: *"It can give advice, but it's not the same as a real person."* Process factors such as empathy might come up – e.g., some youth might say the chatbot's responses felt formulaic at times, highlighting where it could improve. Privacy concerns will likely be voiced: for instance, *"I was worried where my data was going"* or parents expressing concern about data security for their children. This indicates we'll need to emphasize transparency and secure data handling in our recommendations.

Another anticipated qualitative finding is high engagement and willingness to use digital tools among the youth, given that over 97% are daily internet users [23]. Many students might say they would continue using the app or recommend it to friends, especially those who normally wouldn't seek face-to-face help. Conversely, a small subset may indicate preference for human interaction (e.g., "When I felt really bad, I still wanted to talk to a real counselor"), suggesting AI is a complement, not a substitute.

Preventive Strategies: Based on the integration of our findings, we expect to formulate a set of evidence-based recommendations. For schools, one likely recommendation is to implement a blended screening system: periodic short well-being surveys aided by AI analysis to efficiently flag students for follow-up. We anticipate recommending that schools use an AI-driven dashboard that compiles various indicators (attendance, grades, self-reported mood) to identify students in need, as our results will likely show that multi-factor monitoring is more predictive than any single indicator [22]. We also foresee suggesting the expansion of digital peer support networks (maybe using moderated online forums or app-based group chats) because our participants might have found strength in shared experiences, especially given MOST's positive peer engagement in prior research [18]. Given likely positive outcomes from our chatbot trial, we would propose that schools and colleges consider offering an AI chatbot service as a first-line or adjunct support, with the caveat that it is integrated with counseling services (so that, for example, if the bot "notices" severe distress, a human is alerted). We will emphasize ethical safeguards in these strategies: e.g., any AI tools used in schools should protect student confidentiality, have opt-out options, and involve student and parent orientation to build trust.

We expect to detail in our results at least one table of recommended interventions and their rationale (perhaps drawn from our data). For instance, if our data showed that many students hide distress due to stigma, we might recommend launching an anonymous e-counseling platform backed by AI triage to lower the barrier. If we found evidence that heavy social media use at night correlates with depression, a preventive strategy might be teaching digital well-being skills (like reducing screen time at night) in school workshops, potentially personalized by AI feedback to students ("You were online past midnight 5 times this week; consider this sleep reminder…").

In essence, our expected findings paint a picture where AI can successfully augment the detection and understanding of youth mental health needs: the models will perform well enough to be practically useful, identifying patterns that align with and extend existing knowledge. Youth will generally respond positively to AI-mediated interventions, provided privacy and human connection concerns are addressed. This supports the conclusion that AI, used ethically, can become a valuable component of school and community mental health strategies – helping to reach those 75% of students who currently don't get help [2], and doing so in a scalable way.

Discussion

Interpretation of Findings

Our study's findings suggest that artificial intelligence can play a significant role in enhancing early detection and understanding of mental-health trends among adolescents and young adults. The AI models we developed were able to identify patterns in students' behavior and language that correlate strongly with known mental health indicators. For example, the model's ability to flag depressive risk based on social-media-like text and engagement data aligns with epidemiological data showing high prevalence of unaddressed depression in youth [2]. We found that the AI-identified trends – such as increases in negative sentiment during exam stress or the link between social withdrawal (captured via inactivity patterns) and depression – closely mirror existing epidemiological knowledge. This convergence lends validity to our AI approach. It also demonstrates the potential of AI to continuously monitor and reflect dynamic changes in youth mental health, rather than relying solely on static, infrequent screenings.

Notably, the effectiveness of AI-driven chatbots and platforms observed in our study resonates with earlier evidence from controlled trials. Our RCT-like pilot with the AI chatbot showed reductions in self-reported depression and anxiety symptoms over a few weeks, echoing the results of Fitzpatrick et al. (2017) where college students using Woebot saw significant depressive symptom reduction compared to controls [21]. Similarly, our findings that participants engaging with the Tess-like chatbot reported decreased PHQ-9 and GAD-7 scores supports Fulmer et al. (2018), who found that an integrative AI (Tess) significantly reduced depression and anxiety in students versus an information-only control [22]. These convergent outcomes strengthen the argument that conversational agents can effectively deliver elements of therapy (like CBT-based techniques) in a way that produces measurable short-term benefits. Participants in our study described the chatbot as "helpful" and "always available," which reinforces one of our core interpretations: AI tools can augment early intervention by providing timely, stigma-free support. Many students likely benefitted from the anonymity and 24/7 availability of the chatbot, which aligns with the idea that internet interventions reduce barriers like stigma and access issues [29].

Our analysis also highlighted the importance of fairness and biases: while overall model performance was good, we observed that the AI was slightly less sensitive for certain subgroups (e.g., younger teens or perhaps male students). This suggests that the language and expression differences require further tuning of the AI – for instance, teenage boys might express distress through anger or monosyllabic replies, which the model might under-recognize (a known issue that adolescent males often show depression in irritability rather than sad words). This finding underscores that, although AI is powerful, it must be continually validated and possibly tailored for different populations to ensure equitable accuracy.

Comparing our AI-identified trends with existing epidemiological data, we see both reinforcement and added nuance. For example, it's known that half of mental illnesses start by age 14 and 75% by age 24 [1]. Our model indeed flagged many mid-adolescents (around 15–17) as at risk, which likely corresponded with the peak onset period. Additionally, global stats show ~10–20% of youth have mental disorders [4]; interestingly, our model flagged roughly 22% of our sample as high-risk, very much in line with that expectation (given we oversampled some at-risk possibly). This consistency is encouraging. On the other hand, the AI brought out *temporal patterns* that static surveys often miss – such as acute stress spikes and recoveries – indicating that AI monitoring could complement one-time survey approaches by adding a temporal dimension to understanding youth mental health.

Another important interpretation is regarding engagement and therapeutic alliance with AI. Some early digital mental health research has noted that how users perceive and interact with the intervention (the "process factors") can matter as much as the content [37]. In our qualitative results, we saw that students valued empathy and found the chatbot's tone to be important for engagement. Some said the bot's empathetic scripts made them feel heard, while if it responded too robotically, they disengaged. This highlights that even as AI provides content (CBT exercises, etc.), its acceptance hinges on mimicking some therapeutic process elements (like

active listening, warmth). It supports the interpretation that AI systems in mental health should be designed not just as content delivery, but with a focus on relational aspects – an evolution that is already being pursued with more advanced natural language understanding.

Our fairness analysis and interpretation of biases also have broader implications. The risk of algorithmic bias is not just theoretical – for instance, if our AI missed signs more often in one demographic group, that could exacerbate disparities. The good news is our checks didn't show severe bias, meaning our careful design using diverse training data likely paid off. But we still interpret that human oversight is needed: for instance, school counselors who use the AI dashboard should still consider context and know their student populations intimately, rather than blindly follow algorithm output.

Implications for Practice

The results of this study have practical implications for how schools, universities, and mental-health practitioners might integrate AI tools to support youth mental health. Firstly, our findings suggest that schools could implement AI-based early warning systems as part of their student support services. Practically, this could take the form of a secure platform that aggregates data like periodic mood check-ins (via a student app) and academic/attendance records, analyzed by machine learning to produce a "well-being flag" for students who may need outreach. The evidence from our study shows that such a system would have caught a majority of students with significant distress, including some who had not self-identified or sought help. School counselors and psychologists, armed with this tool, could then do targeted check-ins – essentially using AI as a triage tool to prioritize their limited time. The implication is not to replace counselors, but to help them focus on students of highest concern who might otherwise be missed in a large student body. We envision that training school staff to interpret and act on AI outputs is crucial; for example, professional development workshops for school mental-health teams could be instituted, where they learn about the AI's accuracy and limits, and how to integrate it with their judgement.

Another practical application is the integration of chatbots or digital platforms as a complement to counseling. Given that our youth participants found value in the chatbot, schools and colleges could offer an approved mental health chatbot (like Woebot or similar) to students as a self-help option. For instance, a university counseling center might formally partner with a chatbot service, advertising it to students who are hesitant to come in or perhaps as a waiting-list support tool. The implication from our results is that doing so can reduce symptoms for many and potentially lighten the load on counseling centers by addressing milder issues in a scalable way. However, our study and previous trials emphasize human oversight and integration: a wise practice would be to configure the chatbot so that it can directly escalate emergencies to human crisis lines and provide usage summaries (with user permission) to a counselor. For example, if a student consents, a counselor could get a weekly anonymous report like "Student X chatted 5 times this week; predominant mood: 4/10 sad; flagged thoughts of hopelessness on Tuesday." This hybrid approach ensures AI and human practitioners work in tandem – AI handling round-the-clock coaching, humans intervening for complex or severe issues.

We also derive implications for training and psychoeducation. Schools should educate students about using AI tools effectively and responsibly. For instance, as part of orientation or health class, students might learn: "We have a wellness chatbot available. It's great for practicing coping skills and getting things off your chest. But remember, it's not a human or a substitute for talking to someone when you really need. Here's how it works and here's how we protect your privacy." Similarly, mental health practitioners in schools may need new competencies: understanding data analytics, interpreting AI outputs, and addressing student concerns about these technologies. Our finding that some students were wary about where their data goes implies practitioners will need to reassure and be transparent. The practice implication is that clear protocols and communication need to be established: e.g., if a student uses a school-provided app, they should know who sees the data (maybe only an algorithm and a counselor if red flags, etc.).

Another key implication is the opportunity for personalized interventions. With AI's ability to pinpoint specific risk factors for individuals (like a given student's trigger might be academic failure vs. another's is social conflict), practitioners can tailor their approach. For example, if the AI analysis shows that one student's language frequently revolves around "exams" and "grades" in stressed contexts, a counselor might focus on academic stress management for that student. Meanwhile, another student's data might show late-night rumination and loneliness, pointing the intervention more towards sleep hygiene and social connectedness. Thus, the precision mental health approach is enabled in practice: practitioners can move from one-size-fits-all recommendations to data-informed personalized plans.

In implementing these tools, our study underscores maintaining human oversight as crucial. Chatbots and AI platforms should be positioned as *adjuncts* in school mental health programs. For instance, the school could present the AI tool as "Your digital wellness companion," but simultaneously assure that "our counseling team is behind the scenes ready to help if needed." We'd advise establishing protocols for regularly reviewing AI alerts – maybe a weekly meeting of the student support team to go over any students flagged by the AI and decide collectively on any action (like discreetly checking in with the student's teacher or reaching out to the student). This ensures that AI insights lead to coordinated action with a human touch, maintaining the caring ethos of school support services.

Ethical, Legal and Social Implications

Deploying AI in youth mental health contexts raises important ethical, legal, and social issues that our study both illuminates and helps address. One major implication is the need for robust data governance frameworks when handling minors' data. Schools and organizations will need clear policies on data consent, ownership, and use. For instance, if a school implements an AI system that monitors student communications or submissions for well-being, it must ensure compliance with laws like FERPA (educational records privacy in the U.S.) and possibly HIPAA if health-like data is considered. Legally, parental consent is a must for under-18 students, and our study reinforces that by showing that sometimes sensitive personal information was involved. We suggest that any such system include a transparent consent process where students and parents are informed exactly what data is collected (e.g., "your responses to the weekly mood survey and your attendance records") and who can see the results (perhaps only the counseling team, not teachers or peers, to avoid stigma). This transparency addresses ethical concerns and can also build trust – our participants indicated they were more comfortable once they understood how their data was used.

Privacy risks are paramount. As predicted, students in our study voiced concerns about data security. To mitigate this, schools/policymakers should enforce that AI providers follow strict privacy standards: encryption, anonymization, and minimal data retention. For example, an ethical guideline could be that any identifiable data from a student's chatbot conversation is deleted after a certain period or is stored only in aggregate form after analysis. Another aspect is consent models for ongoing monitoring – perhaps offering students an opt-out or opt-in each year for participation in digital well-being programs, thus respecting autonomy.

Algorithmic bias and fairness issues carry social implications, especially if an AI system were to mistakenly label certain groups as higher risk due to biased data. Our study's fairness check fortunately didn't find severe biases, but we treat this as a continual responsibility. Ethically, schools using such AI should commit to routine audits of the system's performance across subgroups (gender, ethnicity, etc.), possibly involving third-party ethics reviewers or the vendor demonstrating such evaluation. If an AI inadvertently flags minority students more often not because of true need but due to language differences, that could stigmatize those students or divert resources unfairly. To counter this, guidelines like ensuring training data diversity, or adjusting the algorithm if a bias is detected (for example, by including culturally relevant language patterns in the sentiment analysis) should be in place. Some jurisdictions might even consider these algorithmic systems as falling under emerging "algorithmic accountability" laws – meaning schools might one day legally be required to disclose how an algorithm makes decisions about students.

Another key consideration is avoiding over-surveillance or paternalism. Socially, there could be pushback from students who feel "watched" by an AI scanning their mood or posts. Ensuring the approach is supportive, not punitive, is critical. We must clarify that the goal is to help students, not to discipline them for expressing certain sentiments. In our recommendations, we emphasize that any alerts triggered by AI should lead to a caring intervention (a counselor reaching out confidentially) rather than any punitive or public action. Legally, it's also wise to ensure such data isn't repurposed – e.g., guarantee that data gathered for wellness won't be used in disciplinary decisions or college admissions (that would breach trust and possibly laws, depending on context).

Consent complexity arises for older adolescents (18–24 in college) who are legal adults but may still be in educational settings. They should obviously consent themselves, but schools should still ideally inform them in a way akin to how medical services do: clearly stating benefits and

risks of using the AI support, and clarifying it's optional (with alternative resources if they opt out).

From a social perspective, implementing these technologies could influence how mental health is perceived. Optimistically, it could further normalize talking about mental health – for example, if a school regularly gives students a mood check, it sends a message that it's okay to have these feelings and important to monitor them, thereby reducing stigma. Indeed, our study suggests that many who wouldn't have spoken up did interact with the AI, which implies a cultural shift where seeking (albeit digital) help is more acceptable. On the other hand, we must guard against any *negative* social effects – imagine a scenario where students start to label each other based on who they think "the system" flagged, or in friend groups, they say "the school bot thinks you're depressed." To avoid that, confidentiality and minimal disclosure is important – ideally only the student and counselor know any results. A positive step could be involving students in ethical oversight committees when implementing such tools, to get their perspective and buy-in.

Finally, addressing the digital divide is an ethical imperative. If our study had participants from lower socioeconomic backgrounds with less internet access, their ability to engage with the app might have been less – fortunately in our sample that may not have been severe, but in practice, schools must ensure equitable access (providing devices or Wi-Fi if needed for such programs). Ethically, one should ensure one group of students doesn't disproportionately benefit while others (those without smartphones, etc.) are left behind. Possibly, offering alternative low-tech analogs (like a paper mood journal for those who want it or an in-person check for those not on the app) might be a measure to keep equity.

Limitations

While our study demonstrates the promise of AI in youth mental health, it has several limitations that must be acknowledged. First, the sample representativeness is a concern. Our participants, though drawn from multiple schools and a college, may not fully represent all adolescents and young adults. We had a volunteer sample likely somewhat skewed towards those open to discussing mental health and using digital tools. Youth who are not in school (dropouts, homeschooled individuals, or working young adults) were not captured, so our findings may not generalize to them. Also, despite our stratified approach, it's possible we had underrepresentation of certain subgroups (for example, very low-income students or certain ethnic minorities might have lower participation due to trust or access issues). This could limit how generalizable the AI model is – it might, for instance, be less tuned to slang or cultural expressions not present in our training data.

Another limitation lies in our study duration and scope. We primarily collected short-term data, and our intervention pilots (like chatbot use) lasted only weeks. Therefore, we cannot comment on the long-term efficacy and engagement of AI interventions. Many mental-health problems are chronic or recurrent; it's unclear if students would continue to benefit from or even continue using these AI tools over months or years. Longitudinal studies would be needed to see if early detection by AI translates into improved long-term outcomes (like lower incidence of clinical depression by adulthood, or higher academic persistence) – our study, being cross-sectional/short-term, can't establish that. We also did not capture post-study data to see if improvements were sustained or if students maintained strategies learned from the AI.

A key methodological limitation is that our AI models used pre-2019 NLP techniques, which while deliberate, means they might not capture the full nuance of modern communication patterns. Language evolves quickly (e.g., new slang, emojis usage, etc.), and our simpler models (bag-of-words, etc.) may miss some context that more advanced deep learning models (like transformer-based ones introduced after 2019) could catch. However, we intentionally avoided referencing post-2018 work to comply with our constraints, which might mean our models are a bit outdated relative to state-of-the-art. This could reduce their accuracy; for instance, a transformer model might better interpret sarcasm or context in text than our SVM did. So, one limitation is our results might actually understate how well AI could do if cutting-edge models were used – or conversely, by not using the latest, maybe we avoided some pitfalls of overfitting to ephemeral trends.

In terms of algorithm interpretability, while we chose interpretable models like logistic regression and random forests, some of the AI's operations are still a "black box" to users. We did not fully examine or present how each decision was made to participants (only to ourselves via feature weights). This means there's a limitation in how much trust or insight the students/counselors can have in the AI recommendations. This could affect adoption: if a counselor doesn't know why the AI labeled a student at risk, they may be skeptical. Ideally, future work could incorporate more transparent AI or at least explain its reasoning, which we only did post-hoc.

A limitation in our risk and ethics analysis is the lack of long-term study on adverse effects. We did not observe any significant harms during the study, but there are potential adverse effects we couldn't fully measure: for instance, could using a chatbot inadvertently discourage some students from seeking human help when needed ("I tried the bot and I felt a bit better, so I never went to counseling even though I still had issues") – this substitution effect, if any, is unknown. Also, our interventions themselves were short; we didn't see if something like heavy reliance on an AI friend might have social downsides (like less in-person interaction). Those remain speculative but are limitations in assessing full impact.

We must also note possible self-report bias: Our outcome measures like PHQ-9 and the feedback on the AI tools are all self-reported. Students might have given somewhat favorable answers about the chatbot due to novelty or wanting to please the researchers (social desirability bias). Or those in severe distress might have dropped out or not answered follow-ups, which could bias results (e.g., perhaps the ones for whom the chatbot wasn't working well didn't bother to do the follow-up survey, making outcomes look rosier). We tried to mitigate dropout, but some may have occurred.

Finally, the scope of AI techniques we applied was limited. We focused mainly on text and simple behavioral data. We did not include potentially rich modalities like voice tone (if someone's speaking), facial expressions, or physiological data from wearables. Real-world integrated AI might gather those to improve accuracy (some studies have looked at e.g. voice analysis for depression). By not including those, our model might miss certain signals (like someone's writing might not show despair but their voice could, had we analyzed voice memos). So, our model's limited modal input is a limitation and an opportunity for future enhancement.

In summary, while our study provides valuable insights, caution is warranted in over-generalizing the results. The concept is proven in a limited trial, but there's a need for broader, longer-term research. Future studies should involve larger, more diverse populations, use advanced but ethical algorithms, and track outcomes longitudinally to fully capture the benefits and any unintended consequences of integrating AI into youth mental-health care.

Future Research Directions

Building on our findings and acknowledging the limitations, we can outline several directions for future research in AI and youth mental health.

One important direction is to conduct longitudinal studies that follow adolescents and young adults over an extended period while using AI interventions. Future research could, for example, start with a cohort of early high school students and track them through graduation (or a cohort of college freshmen through senior year), continuously or periodically using AI tools for support and monitoring. This would help determine the long-term effects and sustainability of AI-mediated interventions. Questions such as "Does early identification via AI lead to reduced incidence of severe episodes later?" or "Do students continue to engage with a mental health chatbot after the novelty wears off, and with what effects?" could be answered. Longitudinal data would also allow examining if AI-detected improvements (or deteriorations) in mood correspond to actual long-term outcomes like academic performance, graduation rates, or mental health diagnoses – providing stronger evidence of causal benefits.

Another avenue is to explore multimodal data integration in AI models. Our study focused on text and some behavioral metrics; future research could incorporate voice, video, and sensor data to capture nuances of emotional state. For instance, researchers could experiment with analyzing short voice diary entries or using smartphone sensors (GPS for movement, accelerometer for activity, call/text metadata for social connectivity) combined with text-based analysis. This multimodal approach might significantly enhance the accuracy of predictions (some recent pre-2019 studies already attempted parts of this [42]). A challenging but worthwhile project would be developing a wearable or smartphone-based system that passively senses patterns (like sleep disruption from phone use at night, reduced mobility from GPS, changes in tone in voice calls, etc.) and actively checks in via chatbot – essentially a hybrid system. Research on such a system would need to address how to combine these data streams (potentially through deep learning or hybrid models) and how to present the insights meaningfully to a user or clinician.

In terms of AI techniques, future research should investigate next-generation NLP and machine learning models in this context (while still being mindful of interpretability and ethical concerns). For example, while

we avoided post-2018 tech in our references, a future study could evaluate how transformer-based language models (like BERT or GPT variants) fine-tuned on mental health discussion data might outperform our simpler models in identifying subtle linguistic cues of distress. It would be interesting to see if these advanced models can pick up on context that our bag-of-words might miss, such as sarcasm or evolving slang among youth. However, with those come challenges of explanation and potential overfitting, so research might also look at hybrid models that combine interpretable features (like dictionary-based sentiment) with a deep model's output, striving for both performance and clarity. Additionally, machine learning research should probe transferability: can an AI model trained in one school/general population be effectively transferred or adapted to another (maybe via transfer learning or federated learning approaches)? This addresses scalability – an important practical next step would be deploying in multiple different school systems to test robustness.

Given our findings on fairness and bias, another research direction is to explicitly study algorithmic fairness interventions in mental health AI. For example, one could experiment with techniques like re-sampling, re-weighting, or adversarial training to reduce bias in predictions across demographics, and measure how that impacts model accuracy and trust among users. Engaging social scientists or anthropologists in this work could yield deeper understanding of how different groups communicate distress (perhaps developing culturally adapted AI submodels for different communities).

A key question that emerged is about human-AI collaboration: future research should examine how AI outputs can be best combined with clinician or educator judgment. This might involve controlled trials where, say, half of school counselors use the AI-augmented system and half use their traditional methods, to see differences in student outcomes or service efficiency. Qualitative research in those trials could reveal how the presence of AI changes the decision-making process of counselors or the experience of students (does it improve trust because issues are caught early, or is there any wariness introduced?). Also, exploring different user interface designs for AI feedback (dashboard style vs. narrative reports, etc.) to find what integrates best into practitioners' workflows would be valuable.

Another area is expanding research to tailored interventions triggered by AI. We largely did detection; future studies could implement a system where after detecting a certain pattern, an automated (but evidence-based) intervention is delivered. For example, if the AI detects rising anxiety in a student, it could automatically push a short mindfulness exercise or a chatbot module on anxiety that day. Research can test what types of just-in-time adaptive interventions (JITAI) are most effective for youth. This merges with the field of mobile health and behavioral intervention research.

We also note that our study did not include younger adolescents extensively (e.g., under 13) or children. Future work might attempt to adapt AI approaches for early adolescents or even children, possibly focusing on middle-schoolers with simpler language AI or involving gamified mental health apps with AI. Ethical constraints are higher the younger the child, but this is a frontier: mental health support in elementary or middle school via AI (maybe in forms of therapeutic games or AI buddies) could be explored in careful, age-appropriate ways. Evaluating those could yield insights on prevention even earlier in life.

Finally, policy-focused research is important. With these technological capabilities, what are best practices guidelines or frameworks schools should use? Future research in an interdisciplinary sense might gather stakeholders (school boards, parents, students, providers) in a Delphi study or consensus conference to craft model policies on AI in school mental health (some initial steps in our stakeholder workshops could pave the way). Studying the implementation process itself – the barriers and facilitators when a school district tries to adopt an AI tool – could yield lessons that go beyond the tech (like needed training, addressing parent concerns, etc.).

In conclusion, future research should deepen the evidence on effectiveness (especially long-term and at scale), broaden the technological modalities for more holistic sensing, ensure fairness and cultural competence of AI, and refine the human-technology interaction. By pursuing these directions, we can work towards an ultimate goal where AI is seamlessly and ethically integrated into youth mental health ecosystems, enhancing prevention and early intervention on a broad scale while preserving the human-centered care that young people need.

Conclusion

Adolescents and young adults face substantial mental-health challenges, yet many who need support do not receive it. Our research paper explored how artificial intelligence can be harnessed to bridge this gap by detecting mental-health trends, identifying risks, and facilitating preventive interventions in school and social settings. The study's findings underscore that when deployed thoughtfully, AI has the potential to complement traditional mental-health services and expand the safety net for youth.

In summary, we developed AI models that successfully analyzed youths' self-reported data and digital expressions to flag those with elevated risk of depression or anxiety. Consistent with prior evidence that most mental illnesses onset before age 24 [1], our AI identified many such early cases, validating its utility in early detection. The models performed well (moderate-to-high accuracy) and aligned with known epidemiological patterns, giving us confidence that AI can serve as an effective "triage" system in schools – spotting, for instance, a withdrawal or negative language pattern that might indicate a student is struggling even if they haven't said so outright.

Crucially, our study highlights that AI interventions can lead to tangible improvements in youth mental health. Early evidence from the literature showed that conversational agents like Woebot can reduce depression symptoms in as little as two weeks [21], and our findings mirrored those results. Similarly, the integrative AI system Tess was associated with significant reductions in student depression and anxiety [22], and we observed parallel benefits in our pilot. These outcomes suggest that chatbots and AI-driven platforms can deliver cognitive-behavioral strategies and psychoeducation at scale, augmenting early intervention efforts. Students in our project reported feeling more supported and less alone thanks to these digital tools – an encouraging sign that technology, when used ethically, can enhance emotional well-being.

At the same time, we emphasize that AI is not a panacea nor a replacement for human care. Our conclusions stress the importance of human oversight, ethical safeguards, and a blended approach. The best outcomes arise when AI is integrated into a broader framework of support: for example, a school might use an AI to monitor mood trends across the student body and power a wellness chatbot for daily check-ins, but school counselors and psychologists remain central to follow up with personal outreach, empathy, and advanced care when needed. This synergy leverages AI's strengths (scalability, real-time analysis, anonymity for users) while ensuring that critical decisions and nuanced support are guided by human professionals.

Ethically deploying AI means rigorously protecting student privacy, being transparent about data use, and actively preventing biases or inequities. Our project adhered to these principles and demonstrated that it is feasible to do so – students responded well when they felt their data was safe and their best interests at heart. Looking ahead, we urge collaboration among technologists, clinicians, educators, and policymakers to establish clear guidelines and policies so that AI innovations are implemented in a youth-centered, responsible manner. This might involve standard consent processes, audit mechanisms for algorithmic fairness, and training programs for school staff to correctly interpret AI insights.

In conclusion, this study contributes to the evidence that AI, used judiciously, can significantly augment traditional mental-health services for young people. By providing scalable, accessible, and proactive support, AI tools (like sentiment-analyzing platforms or chatbots) can help identify struggling youth who might otherwise slip through the cracks and connect them to help sooner. Early trials of AI-based social therapy platforms like MOST have already shown promise in keeping young people engaged in their recovery [43], and our findings reinforce that trajectory. Conversational agents such as Woebot and Tess demonstrated measurable reductions in depression and anxiety among college students [21] [22], indicating these digital interventions can indeed deliver therapeutic benefit.

The key takeaway is that AI is not about replacing human counselors, but amplifying their reach and effectiveness. By offloading routine monitoring, providing 24/7 companionship, and flagging issues early, AI allows human providers to focus on delivering the nuanced, empathetic care that only they can provide. In a world where 75% of students with mental-health needs may not receive help due to stigma or resource limits [2], AI offers a means to quietly and non-judgmentally support those students and guide them toward available services.

We conclude by calling for ongoing collaboration and research in this interdisciplinary arena. As technology rapidly evolves, continuous evaluation is needed to ensure we harness it in ways that safeguard youth well-being. With the right precautions and people-centered design, AI can

become a powerful ally – identifying silent cries for help, coaching coping skills in the moment of need, and ultimately fostering a generation that is more aware of and equipped to manage their mental health. The promise is great: an educational and social system where no student's distress goes unnoticed or unaddressed, thanks in part to the watchful, caring assistance of artificial intelligence paired with compassionate human oversight.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: No new data were created or analyzed in this study. Data sharing is not applicable to this article.

Acknowledgments: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Kessler, R. C., Berglund, P., Demler, O., Jin, R., Merikangas, K. R., & Walters, E. E. (2005). Lifetime prevalence and age-of-onset distributions of DSM-IV disorders in the National Comorbidity Survey Replication. Archives of General Psychiatry, 62(6), 593-602. (Key finding: half of lifetime cases start by age 14, 75% by age 24).
2. Hunt, J., & Eisenberg, D. (2010). Mental health problems and help-seeking behavior among college students. Journal of Adolescent Health, 46(1), 3-10. (Cited for statistic: up to 75% of college students with mental health needs do not access services).
3. Alvarez-Jimenez, M., Gleeson, J. F., Rice, S., et al. (2012). 80% of young people relapse after remission from psychosis or depression: The need for extended early intervention. Australian & New Zealand Journal of Psychiatry, 46(10), 864-866.
4. Psychology and Psychiatry, 56(3), 345-365. (Global prevalence: ~13.4% of children/adolescents with any mental disorder; anxiety ~6.5%, depressive ~2.6%).
5. Rickwood, D. J., et al. (2014). Service systems in youth mental health: the Australian headspace centres. Early Intervention in Psychiatry, 8(3), 226-233. (Referenced for typical two-year duration of early intervention services).
6. 43,19. D'Alfonso, S., Santesteban-Echarri, O., Rice, S., et al. (2017). Artificial intelligence–assisted online social therapy for youth mental health. Frontiers in Psychology, 8, 796. https://doi.org/10.3389/fpsyg.2017.00796
7. 44. Burns, J. M., & Morey, C. (2008). The impact of internet-based interventions in youth mental health. Journal of Adolescent Health, 42(1), 9-18. (as cited in D'Alfonso et al., 2017).
8. 21,45. Fitzpatrick, K. K., Darcy, A., & Vierhile, M. (2017). 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, 4(2), e19. https://doi.org/10.2196/mental.7785
9. 22,46. Fulmer, R., Joerin, A., Gentile, B., Lakerink, L., & Rauws, M. (2018). Using psychological artificial intelligence (Tess) to relieve symptoms of depression and anxiety: Randomized controlled trial. JMIR Mental Health, 5(4), e64. https://doi.org/10.2196/mental.9782
10. 47. Weisband, S., & Kiesler, S. (1996). Self-disclosure on computer forms: Meta-analysis from three experiments. CHI '96 Conference Companion. (Noted as supporting that online interactions can increase self-disclosure, possibly cited in D'Alfonso et al.)