

RESEARCH INSIGHT

Teaching AI to Hear a Quiet Cry for Help

BesideU shows how AI can notice youth suicide risk earlier while keeping risk decisions auditable and crisis care in human hands.

Pouyan Esmail Zadeh, Department of Information Systems and Business Analytics, College of Business, Florida International University (pesmaeil@fiu.edu)

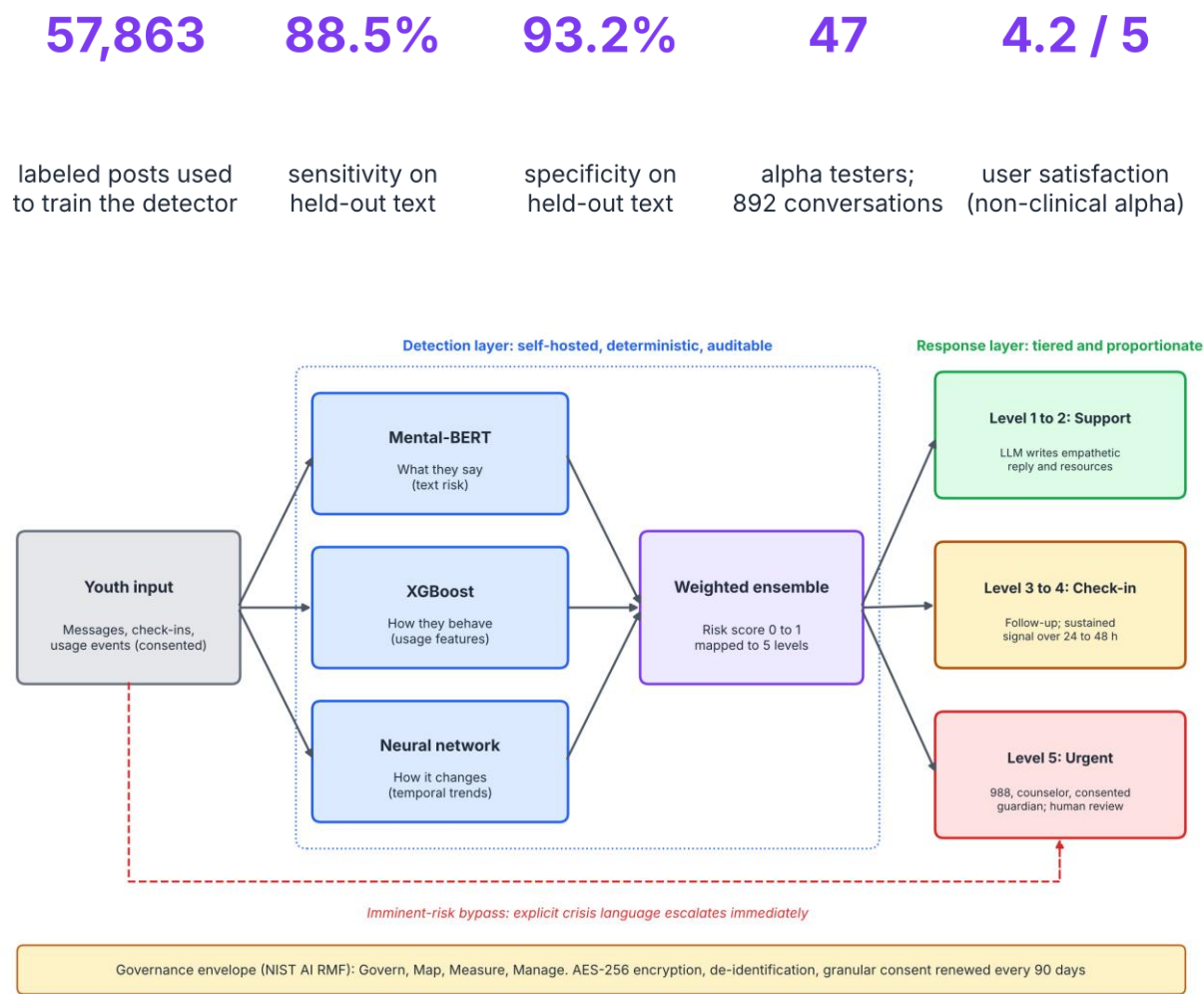
The problem. Suicide is the second leading cause of death for Americans aged 10 to 34, and in 2023 one in five U.S. high school students seriously considered it. Clinicians rely on tools such as the PHQ-9 and the Columbia Protocol, but they capture a single moment, while risk rises and falls in the weeks between appointments. Young people increasingly voice their distress to chatbots, and general-purpose AI companions have drawn FDA and state scrutiny for responding inconsistently when it matters most. This research asks a design question: how should a conversational AI be built to notice risk earlier without becoming a hidden, unaccountable decision-maker?

The approach. Building on a 2025 AI and Ethics review that proposed a five-dimension framework for evaluating suicide prevention chatbots, the project designed and built BesideU, a mobile AI companion for youth. BesideU reads five families of signals over time (language, app behavior, emotion, engagement, and social connection) and compares each person with their own baseline. Self-hosted models do the detecting: Mental-BERT for what a person says, XGBoost for how they use the app, and a neural network for how patterns change. A large language model only writes supportive replies. Risk is scored on five levels aligned with the Columbia Protocol and answered proportionately, from coping resources, to a check-in, to an urgent route to the 988 Lifeline, a counselor, or a guardian the user has chosen.

What we found. Trained on 57,863 labeled posts and tested on posts it had never seen, the detection ensemble reached 91.0% accuracy, 88.5% sensitivity, and 93.2% specificity. Requiring signals to persist for 24 to 48 hours cut false alerts from 12% to 6.3%, while explicit crisis language still triggers an immediate response. In an eight-week, IRB-approved, non-clinical test with 47 FIU students and staff (892 conversations), satisfaction averaged 4.2 out of 5, 87% found replies empathetic, and 91% were comfortable with privacy and consent.

Why it matters. The study offers six design principles for high-stakes conversational AI, led by one idea: let AI notice, and let people decide. A base-rate analysis shows why. The same classifier would be right about 21% of the time when the condition is rare and about 62% of the time at 11% prevalence, so a flag should open a supportive conversation, not deliver a verdict. A 200-participant validation against the PHQ-9 and the Columbia Protocol is now underway.

BesideU at a glance



Youth input

Messages, check-ins,
usage events (consented)

Detection layer: self-hosted, deterministic, auditable

Mental-BERT

What they say
(text risk)

XGBoost

How they behave
(usage features)

Neural network

How it changes
(temporal trends)

Weighted ensemble

Risk score 0 to 1
mapped to 5 levels

Response layer: tiered and proportionate

Level 1 to 2: Support

LLM writes empathetic
reply and resources

Level 3 to 4: Check-in

Follow-up: sustained
signal over 24 to 48 h

Level 5: Urgent

988, counselor, consented
guardian; human review

Imminent-risk bypass: explicit crisis language escalates immediately

Governance envelope (NIST AI RMF): Govern, Map, Measure, Manage. AES-256 encryption, de-identification, granular consent renewed every 90 days

How BesideU works. Self-hosted models estimate risk from what a person writes, how they use the app, and how patterns change over time. A language model only writes replies. Responses are tiered, and explicit crisis language goes straight to human help. Statistics are preliminary: detection was tested on public text, and the alpha test used non-clinical conversations.

References

Centers for Disease Control and Prevention. (2024). Youth Risk Behavior Survey, United States, 2023. MMWR Supplements, 73(4).

Esmailzadeh, P. (2025). Decoding the cry for help: AI's emerging role in suicide risk assessment. AI and Ethics, 5, 4645-4679. <https://doi.org/10.1007/s43681-025-00758-w>

Healthy Minds Network. (2025). The Healthy Minds Study: 2024-2025 national data report.

Ji, S., Zhang, T., Ansari, L., Fu, J., Tiwari, P., & Cambria, E. (2022). MentalBERT: Publicly available pretrained language models for mental healthcare. Proceedings of LREC 2022, 7184-7190.

Kroenke, K., Spitzer, R. L., & Williams, J. B. W. (2001). The PHQ-9: Validity of a brief depression severity measure. Journal of General Internal Medicine, 16(9), 606-613.

National Institute of Standards and Technology. (2023). Artificial Intelligence Risk Management Framework (AI RMF 1.0), NIST AI 100-1.

Posner, K., et al. (2011). The Columbia-Suicide Severity Rating Scale. American Journal of Psychiatry, 168(12), 1266-1277.

U.S. Food and Drug Administration. (2025, November 6). Digital Health Advisory Committee: Generative AI-enabled digital mental health medical devices.