The C.A.S.E. Framework

alongside

A comprehensive evaluation framework for AI-enabled wellness chatbots for youth

Developed over 4 years of staged development, independent research, and clinical oversight

C
CREDIBLE

Evidence-based techniques and appropriate scope

35% WEIGHT

A
ACCESSIBLE

Age and culturally appropriate language

30% WEIGHT

S
SAFE

U18 protections, and mandatory escalation verified first

SAFETY GATE

E
ENGAGE within BOUNDS

Non synchophantic guide, limited topics and stop missuse

35% WEIGHT

HOW IT WORKS

Safety-First Scoring

Every AI interaction passes through a binary safety gate before any quality scoring begins. A failure on safety results in an automatic failing grade, regardless of performance elsewhere.

WEIGHTED SCORE FORMULA

Final Grade = (Engagement x 0.35) + (Credibility x 0.35) + (Accessibility x 0.30)

EVALUATION DEPTH

40-Item Coding System

A granular coding system across the four categories enables precise identification of failure modes and systemic trends — pinpointing exactly why interactions underperform.

Human Oversight

Clinical safety team reviews all flagged chats within 24 hours. Weekly random QA sampling plus LLM-assisted review of chats.

STUDENT SAFEGUARDS

- **Anti-dependency measures:** usage caps, real-world goal-setting
- **Anti-anthropomorphization:** AI disclosure at every session, coaching-oriented personality
- **Crisis escalation:** Integrated assessment & school staff/parent notification,
- **Equity monitoring:** classification models trained on diverse real student data
- **District customization:** configurable alerts, access windows & topic restrictions

4+	37	200+
YEARS OF STAGED AI DEPLOYMENT	LANGUAGES SUPPORTED	STUDENT ADVISORS IN CO-DESIGN

INDEPENDENT EVIDENCE

READ THE FULL WHITE PAPER