# The S.U.R.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

**S**
### SAFE
Safety identification, alerts and data privacy

**U**
### UNDERSTANDABLE
Developmentally appropriate and accessible content

**R**
### RESTRICTED
Guardrails limit both content and engagement

**E**
### ETHICAL
Credible content, non-sycophantic, supports human connection

## 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 = (Undestandable x 0.20) + (Restricted x 0.40) + (Ethics x 0.40)

### 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.

## ALONGSIDE'S SAFEGUARDS

| | |
|---|---|
| S | • Immediate safety alerts<br>• FERPA/COPPA compliant |
| U | • 37+ Languages<br>• Developed with over 200 teen advisors |
| R | • Content restrictions based on age<br>• Time limited |
| E | • Created by clinicians<br>• Focus on building real-life skills<br>• Connects youth to human support |

## INDEPENDENT EVIDENCE




READ THE FULL WHITE PAPER