ThroughLine

Suicidal Ideation Detection in Conversational AI: A Compliance and Implementation Guide

throughlinecare.com

Table of Contents

Table of Contents

SECTION 1:

Understanding Regulatory Requirements & Evidence-Based Standards

If you only read one thing:

Emerging regulations worldwide are requiring conversational AI platforms to detect and respond to suicidal ideation expressed by users. California's SB 243—the most prescriptive law currently in effect—requires "evidence-based" detection by January 1, 2026, meaning your approach must be grounded in peer-reviewed suicide research. While this guide is anchored to SB 243's specific requirements, the evidence-based implementation approaches described here apply broadly to conversational AI platforms and align with emerging regulatory frameworks globally.

1.1 The Emerging Regulatory Landscape

Conversational AI platforms worldwide are facing new requirements to detect and respond to suicidal ideation and self-harm expressed by users.

Current US regulations addressing user-expressed suicidal ideation and self-harm:

- California Senate Bill (SB) 243 (January 1, 2026): Requires companion chatbot platforms to use "evidence-based" methods for detecting suicidal ideation, detect expressions of self-harm, and provide crisis resources
- New York State Senate Bill 2025-S3008 (November 5, 2025): Requires AI companion platforms to maintain protocols for addressing suicidal ideation and self-harm expressed by users, including crisis resource notifications
- **Federal bills under consideration** would require detection of suicidal ideation, though none have been passed as of December 2025

Current international frameworks do not explicitly require the detection of suicidal ideation and self-harm but do prohibit AI systems from promoting harmful behavior (EU AI Act) and require platforms to prevent exposure to content promoting suicide and self-harm behavior (UK Online Safety Act). Australia's eSafety Commissioner has registered Industry Codes requiring platforms to prevent children from being exposed to harmful suicide and self-harm related content.

Why this guide focuses on California's SB 243:

At the time of writing, SB 243 provides the clearest and most demanding standard for the detection of suicidal ideation, and meeting this evidence-based requirement positions platforms to address all current regulatory approaches.

1.2 California SB 243: What it Requires

California SB 243 requires conversational AI platforms to implement systems for detecting and responding to suicidal ideation and self-harm content expressed by users.

Core requirements:

1. Suicidal ideation detection and response

SB 243 regulatory language: *Companion chatbot platforms must "use evidence-based methods for measuring suicidal ideation" (§22603.d) in ongoing conversations and "respond to instances of suicidal ideation by users" (§22604.a.2) "by providing a notification to the user that refers the user to crisis service providers, including a suicide hotline or crisis text line" (§22602.b.1).*

What this means: Your system must use evidence-based methods to detect when users express suicidal thoughts, and provide crisis resources when detected. However, the law does not prescribe specific technologies or approaches. The key question here is: what does "evidence-based" mean?

2. Self-harm detection and response

SB 243 regulatory language: *Companion chatbot platforms must maintain "a protocol for preventing the production of…self-harm content to the user, including, but not limited to, by providing a notification to the user that refers the user to crisis service providers, including a suicide hotline or crisis text line, if the user expresses… self-harm" (§22602.b.1).*

What this means: Platforms must detect and provide appropriate interventions or resources in response to self-harm content expressed by users, though the regulation does not require this detection to be evidence-based.

Effective date: January 1, 2026

The law applies to AI companion services operating in California that provide *"adaptive, human-like responses to user inputs" and are "capable of meeting a user's social needs, including by exhibiting anthropomorphic features and being able to sustain a relationship across multiple interactions" (§22601.b.1).*

Consult legal counsel to determine if your platform is covered.

1.3 What "Evidence-Based" Means

SB 243 requires "evidence-based methods" but doesn't define this precisely.

In brief: Evidence-based means your detection approach is grounded in peer-reviewed research on suicidal behavior—not intuition, not just internal testing, and not unvalidated vendor claims.

Section 2 defines "evidence-based" in operational detail.

1.4 Scope of This Guide

What we cover:

- What "evidence-based" means (Section 2)
- Clinical research on suicidal ideation and validated risk factors (Section 3)
- An actionable 4-level Detection & Response Framework for suicidal ideation (Section 3)
- Self-harm detection requirements and research foundation (Section 4)
- Three approaches to evidence-based detection that can be applied to both suicidal ideation and self-harm (Section 5)
- System evaluation, maintenance, and documentation (Section 6)

What we don't cover:

- Legal interpretation of which regulations apply to your platform (consult counsel)
- Regulatory requirements beyond detection and response to suicidal ideation and self-harm, such as:
  - User notification requirements about AI nature
  - Prevention of AI-generated harmful content
  - Minor-specific protections and age verification
- Intervention design and crisis response protocols
- Privacy and data protection requirements
- Liability questions

1.5 A Note on Resources

Evidence-based detection does not necessarily require:

- Large labeled datasets of suicidal content
- Custom-trained machine learning models
- Large specialized teams

Section 5 presents approaches ranging from simpler rule-based systems to more sophisticated custom models. Evidence-based means grounded in research, not necessarily using the most advanced technology.

Key Takeaways

- Emerging regulations require conversational AI platforms to detect and respond to user-expressed suicidal ideation and self-harm
- California's SB 243 (effective January 1, 2026) requires "evidence-based" detection methods for suicidal ideation—the most demanding standard currently in effect. "Evidence-based" means grounded in peer-reviewed research (see Section 2)
- The evidence-based implementation methods in this guide apply to conversational AI platforms broadly, positioning platforms for compliance across jurisdictions
- Self-harm detection is often required alongside suicidal ideation detection; Section 4 provides research-grounded best practices
- The law doesn't mandate specific technologies - multiple approaches can comply
- Consult legal counsel for questions about applicability, liability, and legal compliance

SECTION 2:

What "Evidence-Based" Means

If you only read one thing:

You demonstrate evidence-based compliance by: (1) naming which validated constructs your system targets, (2) citing the peer-reviewed research supporting those constructs, (3) explaining how you detect them, and (4) acknowledging research-based limitations.

2.1 The Core Standard

"Evidence-based" means your detection methods are grounded in peer-reviewed research on suicidal behavior that has been empirically tested.

In practice: You can identify which psychological constructs your system targets, cite the research validating those constructs, and explain how your system detects them.

You're not required to achieve perfect accuracy or use specific technologies. You're required to show that your approach is informed by scientific research on suicide.

A note on scope: California's SB 243 requires evidence-based detection methods for suicidal ideation. While this and other existing regulations don't mandate evidence-based methods for self-harm detection, Section 4 provides research-grounded best practices for detecting and responding to user-expressed self-harm. The evidence-based principles described in this section apply to conversational AI platforms regardless of jurisdiction.

2.2 What Counts as Evidence

Sufficient Research Foundations

Peer-reviewed clinical theories:

- Published in scientific journals
- Empirically tested across multiple studies
- Widely cited in suicide research

Meta-analyses and systematic reviews:

- Comprehensive reviews of suicide research
- Studies identifying consistent predictors across populations

Validated clinical assessment tools:

- Instruments used by mental health professionals
- Tools with published psychometric properties

What Is Not Sufficient

- **Internal testing alone -** Your own data without connection to clinical research
- **Non-transparent vendor claims -** Providers who won't explain how their approach connects to research or just assert "it works"
- **Intuition or assumptions -** Beliefs about suicidal language without research support
- **Media reports -** Secondary sources rather than peer-reviewed research

2.3 What You Need to Demonstrate

To show an evidence-based approach:

1. **Name the constructs you're targeting**
Example: "Our system detects expressions of hopelessness, social disconnection, and perceived burdensomeness."

2. **Cite the research validating those constructs**
Example: "These constructs are validated in [Theory Name] (Author, Year), which has been empirically tested in [brief description]."

3. **Explain your operationalization**
Example: "We operationalize 'perceived burdensomeness' by screening for statements where users express that they burden others, such as..."

4. **Acknowledge research-based limitations**
Example: "Consistent with research findings, our system identifies current risk indicators but cannot predict future attempts."

See Section 6: Evaluation, Maintenance, & Documentation for more in-depth suggestions for documentation.

2.4 Evidence-Based vs. Perfectly Accurate

Critical distinction: "Evidence-based" does not mean "perfectly accurate."

Research consistently shows:

- Perfect prediction of suicide is impossible
- Even validated risk assessments have limited predictive power

- Context and individual variation make detection inherently imperfect
- Suicide risk can change rapidly over short periods, making continuous monitoring throughout conversations essential

Being evidence-based means:

- Targeting what research identifies as risk factors
- Acknowledging the limitations of research
- Not claiming capabilities research shows are unattainable

2.5 Multiple Valid Approaches

There is no single "correct" evidence-based approach. The suicide research literature presents multiple validated theories, and platforms can legitimately draw from different research foundations.

Your system might focus on specific theories, synthesize constructs from multiple theories, or draw from meta-analyses of risk factors. What matters is that you can articulate which research supports your choices.

Section 3 presents the major clinical theories you can draw from, and Section 5 shows how different system types can implement evidence-based detection.

2.6 Technology Doesn't Matter—Research Foundation Does

"Evidence-based" refers to your research foundation, not your technology.

Rule-based pattern matching, LLMs with research-informed prompts, classical ML models, and custom trained models can all be evidence-based. The question isn't "what technology?" but "what research informs how you use it?"

A simple rules-based system targeting validated constructs is evidence-based. An advanced AI model detecting non-validated patterns is not.

2.7 Documentation and Ongoing Updates

Documentation is how you demonstrate compliance. Maintain:

- Which theories and studies inform your approach
- How you operationalize each construct
- Known limitations based on research
- Testing and evaluation methods

Language evolves and so does research. Neologisms come in and out of use and emojis may be used to convey new meaning. Update your operationalization of constructs on a regular basis. Make sure you also stay informed about major developments in suicide research and update your approach when significant findings emerge.

Section 6 provides more in-depth documentation guidance.

Key Takeaways

- Demonstrate evidence-based detection through four steps: name constructs, cite research, explain operationalization, acknowledge limitations
- Multiple research foundations are valid - there's no single correct approach
- Any technology can be evidence-based if grounded in research
- Keep operationalization and documentation current as language and research evolve

SECTION 3:

Clinical Foundations for Evidence-Based Detection of Suicidal Ideation

If you only read one thing:

Modern suicide research identifies psychological constructs (pain, hopelessness, social disconnection, burdensomeness, defeat/entrapment) and behavioral/historical indicators (prior attempts, specific plan/intent, access to means) that signal risk. The Detection & Response Framework (Section 3.4) provides a 4-level framework to operationalize these into actionable responses. Levels 3-4 (plan, means and imminent intent) represent the highest risk; Levels 1-2 (implicit risk indicators and explicit ideation) are necessary for comprehensive detection of suicidal ideation.

3.1 The Research You'll Draw From

This section presents the validated clinical research that forms the foundation for evidence-based detection and operationalizes it into actionable detection and response guidance.

The three major theories in modern suicide research—the Interpersonal Theory of Suicide (Van Orden et al., 2010), the Integrated Motivational-Volitional Model (O'Connor & Kirtley, 2018), and the Three-Step Theory (Klonsky & May, 2015)—all converge on core constructs and insights. You don't need to master each theory. You need to understand the validated risk factors they identify and how to operationalize them in your conversational AI system.

3.2 What Research Says to Look For

Multiple Factors, Not Single Indicators

A key finding in suicide research: Multiple risk factors typically co-occur in severe ideation. Detection based solely on single words or phrases is not sufficient.

Research consistently shows combinations matter:

- Pain without hopelessness → person keeps trying to improve things
- Hopelessness without pain → person may be pessimistic and not suicidal
- Isolation without other factors → distress but not necessarily ideation

For your system: Look for combinations of validated psychological constructs, not isolated keywords. Use behavioral/historical indicators to determine imminence and risk severity.

Quick Reference: Evidence-Based Constructs

Construct	Research Basis	What to Look For
Psychological Constructs		
Psychological Pain	Klonsky & May (2015)	Unbearable suffering, emotional torment
Hopelessness	All major theories	"Nothing will change," no hope for future
Social Disconnection	Van Orden et al. (2010)	Isolation, loneliness, no caring relationships
Perceived Burdensomeness	Van Orden et al. (2010)	"I'm a burden," self-hatred, "better off without me"
Defeat/Entrapment	O'Connor & Kirtley (2018)	Feeling beaten down and trapped, no escape, "no way out"
Behavioral/Historical Indicators		
Prior Attempts	All theories - strongest historical predictor	Disclosure of past attempts or ongoing self-injury
Specific Plan/Intent	O'Connor & Kirtley (2018)	Discussion of specific methods, especially alongside intent to act
Access to Means	O'Connor & Kirtley (2018)	Easy availability or presence of lethal means

Research-Validated Constructs Explained

Psychological Pain + Hopelessness

- **Pain:** Intense emotional/mental suffering that makes living feel unbearable
- **Hopelessness:** Belief that the pain won't improve; no hope for a better future
- **Why both matter:** Research shows they interact—both are generally present in severe ideation
- **Look for:** Expressions of suffering combined with "nothing will get better" language

Social Disconnection

- **What it is:** Feeling isolated, lonely, lacking caring relationships
- **Why it matters:** Consistently a strong predictor across studies
- **Look for:** "No one cares," isolation, loneliness, feeling cut off from others

Perceived Burdensomeness

- **What it is:** Belief that your existence burdens others; self-hatred
- **Why it matters:** Strong correlation with ideation, especially when combined with disconnection
- **Look for:** "I'm a burden," "everyone would be better off without me"

Defeat and Entrapment

- **What it is:** Feeling beaten down by life (defeat) and trapped with no escape (entrapment)
- **Why it matters:** The combination particularly predicts ideation
- **Look for:** "No way out," "stuck," "can't escape," expressions of total failure

Prior Attempts and Self-Injury

- **Why it matters:** Strongest predictor of future attempts
- **Look for:** Disclosure of previous attempts, ongoing self-injury, habituation to pain

Specific Plan or Intent

- **Why it matters:** A critical indicator of imminence of an attempt
- **Look for:** Detailed descriptions of a suicide plan and/or statements of intent to act on a plan for suicide

Access to Means

- **Why it matters:** Critical factor for distinguishing attempt risk from ideation
- **Look for:** Discussion of specific methods, access to lethal means

Protective Factors

These reduce risk even when pain/hopelessness are present:

- **Connectedness:** Having people, roles, or purposes to live for
- **Social support:** People who actively help
- **Future plans:** Goals, things to look forward to

Ideation vs. Attempts Are Different

All modern theories agree: thinking about suicide is distinct from attempting it.

- **Ideation is driven by:** pain, hopelessness, isolation, burdensomeness
- **Attempts require:** everything above PLUS capability (e.g., habituation to fear/pain), plan and intent, and access to means

Research also shows that many attempts occur impulsively during acute crises. The decision and action can happen within minutes to hours when overwhelming stressors (e.g., relationship conflict, financial crisis, intense pain) combine with access to means. This is why an urgent response is necessary when plan/means (Level 3) and/or immediate intent (Level 4) are detected.

Beyond Psychological Constructs: Contextual Risk Factors

Suicide research identifies numerous environmental and contextual risk factors that significantly elevate risk, including:

- **Substance use:** Alcohol and drug use impair judgment and increase impulsivity
- **Economic stressors:** Unemployment, housing instability, financial crisis
- **Trauma and violence:** Domestic violence, sexual assault, gender-based violence
- **Identity-related stressors:** LGBTQIA+ discrimination, refugee/migrant status, experiences of racism or other discrimination
- **Other behavioral patterns:** Problem gambling, concerning social media engagement

Why these matter: These factors interact with the psychological constructs described above, often intensifying pain, hopelessness, and crisis states. Someone experiencing discrimination plus social disconnection plus hopelessness is at elevated risk compared to hopelessness alone.

Detection considerations for conversational AI:

Some contextual factors may emerge naturally in conversations ("I lost my job," "my partner hit me," "I can't afford rent"). When users disclose these stressors alongside psychological constructs, consider this combination as potentially elevating risk.

However, many contextual factors won't be expressed in conversations and therefore cannot be detected by text-based systems. Your detection system focuses on what can be identified from conversation content while recognizing that invisible contextual factors may be present.

A Note on Terminology

Throughout this guide:

- **Constructs or risk factors =** validated psychological and behavioral indicators that research shows cause or predict suicidal ideation (pain, hopelessness, social disconnection, etc.)
- **Suicidal ideation =** thoughts about suicide, ranging from passive wishes ("I wish I didn't exist") to active thoughts ("I'm thinking about killing myself")
- **Plan =** consideration of specific methods
- **Intent =** decision or commitment to act

Your system must detect constructs (which may appear before explicit ideation) and ideation itself (when directly expressed). This aligns with how the research theories model the pathway from risk factors to suicidal thoughts.

3.3 Detecting Constructs from Text: What's Feasible

These constructs describe internal psychological states. Your system infers them from observable language. This creates an inherent gap between the construct and detection.

Detectable from Text

- Expressions of pain/suffering
- Hopelessness about the future
- Isolation and loneliness

- Feeling like a burden
- Feeling trapped with no way out
- Suicidal ideation
- Prior attempts
- Plans, intent, and methods

Not Reliably Detectable from Text

- True intent vs. expression
- Exact timing of risk
- Who will actually attempt (prediction is unreliable)

Understanding Constructs vs. Ideation

The validated constructs (pain, hopelessness, social disconnection, etc.) are risk factors that research shows cause or predict suicidal ideation. **Ideation** is the outcome—thoughts about suicide, whether passive ("I wish I didn't exist") or active ("I'm thinking about killing myself").

Your detection system should target **both**:

1. **Constructs** (Level 1) - Risk factors present, ideation not yet confirmed
2. **Ideation** (Levels 2-4) - Explicitly expressed thoughts of suicide, ranging from passive wishes to imminent intent

This research-grounded approach allows you to identify risk even before someone explicitly says "suicide," while also catching direct expressions at varying severity levels.

Note: Not all mentions of suicide indicate personal ideation. Users may discuss suicide due to intellectual curiosity, processing someone else's death, or worry about a loved one. When constructs are present alongside ambiguous mentions of suicide, Level 1 check-in questions help clarify whether the user is experiencing personal suicidal thoughts.

3.4 Operationalizing Risk Detection: The Detection & Response Framework

Now that you understand the validated constructs, here's how to operationalize them into a detection and response system for conversational AI platforms. This framework provides specific guidance for different risk levels based on what research shows to look for.

Important Context:

Risk Detection & Response

- Research shows that asking about suicide does not increase risk (Dazzi et al., 2014). Direct check-in questions help clarify intent and reduce false positives.

- Suicide risk can fluctuate rapidly. A user may move between risk levels during a single conversation. Ongoing monitoring throughout the interaction is essential.

- Many suicide attempts occur impulsively during acute crises. Research shows that a substantial portion of attempts happen within minutes to hours of the decision, often triggered by immediate stressors like relationship conflicts, financial crises, or overwhelming physical/emotional pain. This underscores the importance of immediate response when plans, means, or intent are detected (Levels 3-4).

Understanding Ideation vs. Attempts

- Experiencing suicidal ideation ≠ a suicide attempt. Research suggests that about 9% of people around the world experience suicidal ideation in their lifetime (approximately 730 million people; Nock et al., 2008). However, the vast majority never attempt suicide. Globally, approximately 720,000 people die by suicide each year (WHO, 2025), a tragically high number, but still a small fraction of those who experience ideation.

Implementation Notes

- This guidance is designed for conversational AI/companion experiences where maintaining the relationship is important while ensuring safety.

- Crisis helplines vary by country, and outdated or incorrect referrals create real risk. For global coverage, ThroughLine maintains a verified database that can be integrated via web app, widget, or API for real-time helpline routing.

Summary of the Detection & Response Framework

Level 1	Risk constructs present (what research shows causes ideation) - ideation not yet confirmed
Level 2	Suicidal ideation explicitly expressed
Level 3	Ideation + plan or access to means
Level 4	Ideation + imminent intent or attempt in progress

Level 1: Implicit Risk Indicators

What to Detect:

Multiple validated constructs appearing together across recent conversation history, for example:

- Pain + hopelessness expressed across messages
- Social disconnection + perceived burdensomeness
- Defeat + entrapment language

Suggested Response:

Check-in question

- **Example:** "It sounds like you're going through a really difficult time. Are you thinking about suicide or harming yourself?"
- *Note:* Phrasing should be adapted to your platform's tone and user base. The key is asking directly but compassionately.

Why This Approach:

- Research shows direct questions don't increase risk and help clarify intent
- Reduces false positives by confirming suicidal ideation before providing crisis resources
- Maintains the conversational relationship while addressing safety
- **DON'T:** Immediately show resources at this level - it interrupts the experience and isn't warranted when ideation isn't confirmed

Level 2: Explicit Suicidal Ideation

What to Detect:

- Thoughts about death or non-existence: "I wish I didn't exist," "I don't want to be here anymore"
- Thoughts about suicide: "I want to kill myself," "Thinking about ending it"
- Multiple implicit risk indicators + acknowledgment of suicidal thoughts
- Note: Ambiguous statements like "I want it all to stop" should trigger Level 1 check-in if constructs are present, rather than assuming Level 2

Suggested Response:

Provide crisis resources prominently

- **Example:** "I'm really concerned about you. There are people who specialize in helping with these thoughts and feelings - would it be okay if I share some resources? The 988 Suicide & Crisis Lifeline is available 24/7 by call or text."
- **Can continue conversation:** User may benefit from ongoing support - maintain the conversation after providing easy-to-access contact information for immediate human support
- Monitor for escalation to Level 3

Level 3: Plan Or Access To Means

What to Detect:

- Discussion of specific method
- "I have a plan" or "I'm going to do it tonight"
- Access to lethal means mentioned
- Preparation behaviors (giving things away, saying goodbye)

Suggested Response:

Immediate, emphasized crisis resources

- **Example:** "I'm really worried about you, and I care deeply about your safety. What you're sharing with me is serious, and I want to make sure you get the right kind of help from people who are trained specifically for moments like this. Please reach out to 988 (Suicide & Crisis Lifeline—call or text) or text HOME to 741741 (Crisis Text Line). They're available 24/7 and can offer the support you need right now in ways I can't."
- **Conversation guidance:** Continue conversation only to encourage help-seeking, not to problem-solve or continue normal conversation
- **Focus:** Repeatedly encourage calling crisis resources, validate their pain, express care

Level 4: Imminent Intent / Attempt In Progress

What to Detect:

- "I'm doing it right now"
- "I'm about to..."
- Statement that attempt is actively underway
- Method + immediate timing stated

Suggested Response:

Emergency services only

- **End conversation** except to provide emergency resources and encouragement to seek immediate help
- Focus solely on getting the user immediate emergency assistance, whether by directing users to call their local emergency number or by contacting emergency response services on the user's behalf when feasible
- **Example:** "This is a medical emergency. Please call 911 (or your local emergency number) or go to your nearest emergency room immediately. You can also call 988 for immediate crisis support. Please get help right now."

3.5 Critical Limitations

From 50 years of research: We cannot reliably predict who will attempt suicide. Systems can identify concerning patterns based on validated risk factors, but cannot predict future behavior with high accuracy.

In addition, text-based detection systems cannot identify all risk factors. As discussed in Section 3.2, many important contextual factors—substance use, financial crisis, experiences of violence, discrimination, or belonging to marginalized communities—may not be expressed in conversations. Your system addresses detectable indicators while recognizing that significant risk factors may remain invisible.

Be honest about these limits in your documentation and to users when appropriate

Key Takeaways

- Target validated constructs from research: implicit risk factors (including pain + hopelessness, social disconnection, burdensomeness, defeat/entrapment), explicit suicidal ideation, prior attempts, plan/intent, access to means
- Multiple co-occurring risk factors often indicate higher risk than a single risk factor alone – single-indicator detection is necessary but not sufficient
- Use the 4-level framework to operationalize detection into appropriate responses based on risk level, recognizing that risk level may change over the course of a conversation and that many suicides happen impulsively in moments of crisis
- Asking about suicide does not increase risk - check-in questions reduce false positives and clarify intent
- Acknowledge inherent limits of text-based detection and prediction

SECTION 4:

Self-Harm Detection Requirements

If you only read one thing:

Self-harm (non-suicidal self-injury) involves deliberately hurting oneself as a coping mechanism, distinct from suicidal behavior. Detect direct statements, descriptions of self-injury, urges, and past self-harm disclosure. Respond to all self-harm expressions with compassionate support and crisis resources. When self-harm co-occurs with suicide risk indicators, use the suicide framework from Section 3.

Note: While SB 243 requires evidence-based methods specifically for suicidal ideation detection, it and other regulations still require platforms to detect and respond to self-harm content. This section provides research-grounded best practices for self-harm detection, even though an evidence-based approach is not mandated.

4.1 Understanding Self-Harm vs. Suicidal Behavior

What Self-Harm Is

Non-suicidal self-injury (NSSI) is the deliberate, self-inflicted damage to one's body tissue without suicidal intent. Common forms include cutting, burning, hitting, scratching, or other methods of causing physical pain or injury.

Critical distinction from suicidal behavior:

- **Self-harm intent:** Coping mechanism, emotion regulation, self-punishment, feeling something physical (see Section 4.4)
- **Suicidal intent:** Intent to die or end one's life

The distinction is about intent, not the behavior itself. The same physical act (e.g., cutting) could be self-harm or a suicide attempt depending on the person's intent.

The Connection Between Self-Harm and Suicide

While self-harm and suicidal behavior are distinct, they are related:

Self-harm as a suicide risk factor: Individuals who engage in self-harm are at significantly elevated risk for future suicide attempts, even when the self-harm itself is non-suicidal (Hamza, Stewart, & Willoughby, 2012; Whitlock & Knox, 2007).

What this means for detection:

- Self-harm history is a behavioral risk indicator in Section 3's suicide framework
- When self-harm expressions co-occur with suicide risk indicators, use the suicide detection framework from Section 3
- If methods could be lethal, treat as suicide risk

4.2 Research Foundation for Self-Harm Detection

Self-harm research establishes key findings relevant to detection systems:

Functions of self-harm: Self-injury serves various psychological functions, most commonly emotion regulation—people use physical pain to manage overwhelming emotional pain (Klonsky, 2007; Nock, 2010).

Prevalence: Self-harm is more common than many realize, particularly among adolescents and young adults (Nock, 2010), making detection important for conversational AI platforms.

(See Section 7 for key research papers on self-harm detection)

4.3 What to Detect

Your system should identify expressions of self-harm thoughts, behaviors, or history:

Direct Statements

- "I cut myself"
- "I hurt myself when I'm upset"
- "I burned myself last night"
- "I hit myself"
- "I've been scratching until I bleed"

Descriptions of Self-Injury Behaviors

- Discussion of cutting, burning, hitting, scratching, or other self-injury methods
- Descriptions of physical self-harm acts
- References to tools used for self-injury

Thoughts or Urges About Self-Harm

- "I want to cut right now"
- "I have the urge to hurt myself"
- "I'm thinking about burning myself"

Past Self-Harm Disclosure

- "I used to cut in high school"
- "I have a history of self-harm"
- References to scars from self-injury

Context-Dependent Expressions

Some expressions require conversation context to interpret:

- "I need to feel something" (could indicate self-harm urge in certain contexts)
- "I deserve to be hurt" (could indicate self-harm or other self-destructive thoughts)

Important: Consider conversation context. Self-harm expressions may emerge across multiple messages, not always in a single statement.

4.4 Common Functions of Self-Harm

Research shows that people often talk about why they self-harm before (or instead of) talking about the behavior itself. Understanding these functions helps identify self-harm risk even when users don't explicitly say "I cut myself" or "I hurt myself."

Research identifies 13 functions of self-harm (Klonsky & Glenn, 2009):

Function	What Users Might Express
Affect regulation	"It calms me down," "helps me cope"
Anti-dissociation	"I need to feel something," "stops the numbness"
Anti-suicide	"Keeps me from doing something worse"
Self-punishment	"I deserve this pain," "I'm worthless"
Marking distress	"I need a physical sign of how bad I feel"
Interpersonal influence	"I need them to understand how much pain I'm in"

(Additional functions include autonomy, interpersonal boundaries, peer bonding, revenge, self-care, sensation seeking, and toughness)

Detection implications: Users discussing these functions, especially in the context of overwhelming distress, may be expressing self-harm thoughts or behaviors without naming them directly. Expressions like those in the table above may indicate self-harm when they appear alongside intense emotional pain. It is important to recognize that **these are context-dependent expressions that require conversation history to interpret accurately.** The same phrase in different contexts may or may not indicate self-harm risk.

4.5 Response Framework for Self-Harm

Unlike the 4-level framework for suicidal ideation, self-harm detection uses a simpler approach based on available research:

When self-harm is detected: Provide a compassionate response (ideally with recognition of the functions of self-harm) and crisis resources.

Response example:

"I'm concerned about what you're sharing. Self-harm often comes up when thoughts and feelings become overwhelming or when you need some sense of control or relief. Whatever you're going through that's bringing this up, you don't have to handle it alone. The 988 Suicide & Crisis Lifeline (call or text) and Crisis Text Line (text HOME to 741741) provide support for people experiencing self-harm thoughts and behaviors. Would you be willing to call or text one of these helplines right now?"

Key response principles:

- Non-judgmental, compassionate tone
- Acknowledge the pain they're experiencing and potential functions of the self-harm
- Provide crisis resources (e.g., 988 Suicide & Crisis Lifeline and Crisis Text Line in the US both support people struggling with self-harm)
- Continue the conversation if the user declines to reach out: The user may benefit from ongoing support so it is okay to maintain the conversation after providing easy-to-access contact information for immediate human support

Why a simpler framework?

The suicidal ideation framework in Section 3 stratifies risk based on extensive research about progression from ideation to attempts. For self-harm, less research exists on risk stratification for severity. The appropriate response—provide resources and support—is consistent across self-harm expressions.

4.6 Critical Overlap with Suicide Risk

When to Use the Suicide Framework

If self-harm expressions appear alongside suicide risk indicators from Section 3, treat as suicide risk and use the appropriate level from Section 3.

Examples:

- Self-harm + "I want to die" → Section 3 framework (Level 2+)
- Self-harm + suicide plan → Section 3 framework (Level 3)
- Self-harm history + multiple suicide risk constructs → Elevated risk, Section 3 framework

When Methods Could Be Lethal

If someone describes self-harm using methods that could be lethal, treat as suicide risk:

- "I'm going to cut my wrists" → Section 3, Level 3 or 4
- "I took a bunch of pills" → Section 3, Level 4 (medical emergency)

Guideline: When in doubt about lethality, err on the side of treating as higher risk.

4.7 Limitations

Self-harm detection has inherent limitations:

What detection can do:

- Identify when users express self-harm thoughts, behaviors, or history
- Provide appropriate resources and support

What detection cannot do:

- Predict future self-harm behavior with certainty
- Determine the severity or medical risk of self-injury from text alone
- Distinguish all ambiguous expressions without conversation context

The goal is identifying expressions that warrant supportive response and resource provision, not perfect prediction or clinical diagnosis.

Key Takeaways

- Self-harm (NSSI) is distinct from suicidal behavior—the distinction is intent
- Self-harm history is a suicide risk factor—when co-occurring with suicide indicators, use Section 3 framework
- Detect direct statements, behavior descriptions, urges, and past self-harm disclosure
- Respond to all self-harm expressions with crisis resources and compassionate support
- Consider conversation context when detecting self-harm expressions

SECTION 5:

Three Compliant Approaches

If you only read one thing:

Multiple approaches can all be evidence-based. We suggest three options: rule-based (explicit patterns), LLM-based (AI with research-informed prompts), or custom trained models (specialized classifiers trained on labeled data). Most companion AI platforms should start with either rule-based detection or LLM-based detection. Choose based on your resources, scale, and technical capabilities.

5.1 Overview

This section presents three approaches to evidence-based detection, ranging from simpler to more sophisticated. All three can satisfy current regulatory requirements, including SB 243, when properly implemented.

These approaches can be applied to both suicidal ideation detection (Section 3) and self-harm detection (Section 4). The examples below focus primarily on suicidal ideation, with brief self-harm examples showing parallel application.

The three approaches:

- **Rule-Based Detection -** Pattern matching grounded in clinical frameworks
- **LLM-Based Detection -** Large language models prompted with research constructs
- **Custom Trained Models -** Specialized classifiers trained on labeled data

Critical consideration for all approaches: Detection should consider conversation context, not just individual messages in isolation. Risk often emerges across multiple messages as users express different constructs over time. See Section 5.6 for details.

5.2 Quick Selector: Which Approach Is Right for You?

Your Priority	Best Approach
Clear compliance documentation, lowest cost	**Rule-Based (Section 5.3)** — Good starting point
Better accuracy with manageable cost, easier maintenance	**LLM-Based (Section 5.4)** — Evolve here as you scale
Massive scale, extremely high accuracy requirements, lower latency	**Custom Trained Models (Section 5.5)** — Advanced only

Remember: You can evolve your approach over time. Common path: Rule-Based → LLM-Based → Custom Models (if scale justifies or vendor available)

5.3 Approach A: Rule-Based Detection

How It Works

Rule-based systems use patterns, keywords, and logical rules to detect validated constructs from Section 3. You define explicit rules based on clinical frameworks and the Detection & Response Framework (Section 3.4).

Example rules targeting validated constructs:

Hopelessness:

- "nothing will ever" + negative outcome
- "no point in" + trying/living
- "it will never get better"

Perceived burdensomeness:

- "I'm just a burden"
- "everyone would be better off without me"

Social disconnection:

- "nobody cares about me"
- "completely alone"

Self-Harm:

- "I cut myself" or "I hurt myself when"
- "I have the urge to" + self-injury method (cut, burn, hit)
- "I've been" + self-injury behavior + time reference

Multi-factor detection example across conversation:

- IF (hopelessness ≥1 expressions) AND (pain ≥1 expressions) in last 10 messages → Flag as Level 1
- IF constructs appear across conversation thread → Higher severity

Explicit statements (Level 2-4):

- "I want to die" → Level 2
- "I have a plan" + method discussion → Level 3
- "I'm doing it now" → Level 4

Evidence-Based Foundation

Research grounding: Each rule targets a validated construct (pain, hopelessness, burdensomeness, etc.) from Section 3.

Multi-factor approach: Rules require construct combinations, consistent with research showing multiple factors often co-occur.

Documentation example: *"Rule set A targets hopelessness, a core construct in the Three-Step Theory (Klonsky & May, 2015). We operationalize hopelessness as expressions indicating belief that suffering will not improve, such as 'nothing will ever get better' or 'no point in trying.' Rules trigger when hopelessness co-occurs with pain expressions across recent conversation history, consistent with research showing both constructs are often present in severe ideation..."*

Pros & Cons

Advantages

- Fully explainable - clear audit trail for regulators
- No training data required
- Predictable, deterministic behavior
- Lower cost (no API fees)
- Full control over detection logic
- Easiest to document for compliance

Limitations

- Language evolution requires regular updates (e.g., new slang like "unalive")
- Struggles with nuance, sarcasm, context-dependent meaning
- Conversation tracking adds complexity
- Evasion potential - users may learn trigger words
- High maintenance burden as language evolves
- Misses implicit patterns that LLMs catch

Best for: Starting point for compliance, limited resources, explainability priority, full control desired

Key Implementation Considerations

Development effort: Moderate-High (rule design + conversation tracking logic)

Maintenance: High (quarterly rule reviews recommended as language evolves)

Conversation context: Maintain sliding window of recent messages (5-20 messages; more is better)

Critical points:

- ✓ Ground every rule in a validated construct from Section 3
- ✓ Track conversation history, don't solely evaluate single messages
- ✓ Require multiple psychological construct matches for Level 1 triggers
- ✓ Map rules explicitly to flowchart levels (Section 3.4)
- ✓ Plan regular rule reviews and updates
- ✗ Don't rely on keywords without considering surrounding context
- ✗ Don't create rules from intuition alone - tie each to research

5.4 Approach B: LLM-Based Detection

How It Works

Use large language models (GPT, Claude, Gemini, etc.) with structured prompts incorporating clinical frameworks and the Detection & Response Framework. Provide conversation context, not just the current message.

Example prompt structure:

```
You are a suicide risk assessment system for a conversational AI platform.

Your task is to analyze conversations for suicidal ideation and suicide risk
indicators based on validated clinical research and determine the appropriate
risk level.

CONVERSATION CONTEXT:

Analyze the ENTIRE conversation history provided below, not just the most
recent message. Risk indicators often emerge across multiple messages.

VALIDATED CONSTRUCTS TO DETECT:

Psychological Constructs:

1. Psychological pain (unbearable emotional suffering)
2. Hopelessness (belief that suffering won't improve)
3. Social disconnection (isolation, loneliness, no caring relationships)
4. Perceived burdensomeness ("I'm a burden," self-hatred)
5. Defeat/entrapment (feeling beaten down and trapped, no way out)

Explicit Suicidal Ideation:

6. Explicit indicators of suicidal ideation without behavioral indicators
   (e.g., "I want to die" or "I want to kill myself")

Behavioral/Historical Indicators:

7. Prior attempts or self-injury (when disclosed)
8. Current plan for suicide or intent to act on a plan for suicide
9. Access to means (discussion of specific methods with ease of access)
```

```
CRITICAL: Multiple psychological constructs often co-occur in elevated
risk. Single psychological constructs alone do not indicate severe risk
when no explicit suicidal ideation or behavioral/historical indicators are
present. However, a single statement expressing explicit suicidal ideation or
behavioral/historical indicators indicates an elevated risk level even when
other indicators are not present.
```

[Prompt continues with risk level definitions and preferred response format...]

Self-harm detection application:

The same LLM-based approach applies to self-harm detection. Your prompt would instruct the model to identify expressions of self-harm thoughts, behaviors, urges, or history and to respond accordingly, as outlined in Section 4.

Evidence-Based Foundation

Research grounding: Prompt explicitly targets validated constructs from Section 3 and maps to the 4-level framework from Section 3.4.

Conversation awareness: Model evaluates how risk factors emerge over time across the conversation history.

Documentation example: *"Our system uses Claude 4.5 Sonnet with prompts structured around the Interpersonal Theory (Van Orden et al., 2010), Three-Step Theory (Klonsky & May, 2015), and IMV Model (O'Connor & Kirtley, 2018). The model analyzes conversation history for psychological constructs of pain, hopelessness, social disconnection, perceived burdensomeness, defeat, and entrapment, along with behavioral and historical indicators of risk. It evaluates whether multiple factors co-occur and maps detected patterns to our 4-level risk classification system based on the Detection & Response Framework..."*

Pros & Cons

Advantages

- Handles nuance, context, and ambiguity well
- Excellent conversation understanding across messages
- Less maintenance - adapts to language evolution
- Catches evasion patterns ("unalive," indirect expressions)
- Built-in multi-factor reasoning capabilities
- Faster deployment than building rule system

Limitations

- Less explainable than rules ("black box" to some degree)
- API costs scale with message volume and conversation length
- Latency from API calls (typically 1-3seconds)
- Some consistency variation in responses
- Vendor dependence on API providers
- Context window limits (though typically sufficient)

Best for: Companies that have validated rule-based approach and are scaling, need better accuracy with evolving language, can manage API costs

Key Implementation Considerations

Development effort: Lower (prompt engineering + API integration)

Maintenance: Lower (occasional prompt refinement as needed)

Conversation context: Define history window based on your use case (e.g., last 10 messages, last 24 hours, or full conversation thread)

Critical points:

- ✓ Always provide and analyze conversation context, not just current message
- ✓ Structure prompts around validated constructs from Section 3
- ✓ Instruct model to consider patterns emerging over time
- ✓ Test across diverse conversation patterns and edge cases
- ✓ Map LLM output to the 4-level framework (Section 3.4)
- ✗ Don't use generic prompts that don't reference validated constructs
- ✗ Don't ignore cost implications at scale
- ✗ Don't skip validation testing before deployment

5.5 Approach C: Custom Trained Models

How It Works

Custom models involve developing specialized classification systems trained on labeled conversation data. This approach requires substantial machine learning expertise, computational resources, and access to high-quality labeled training data.

Key requirements:

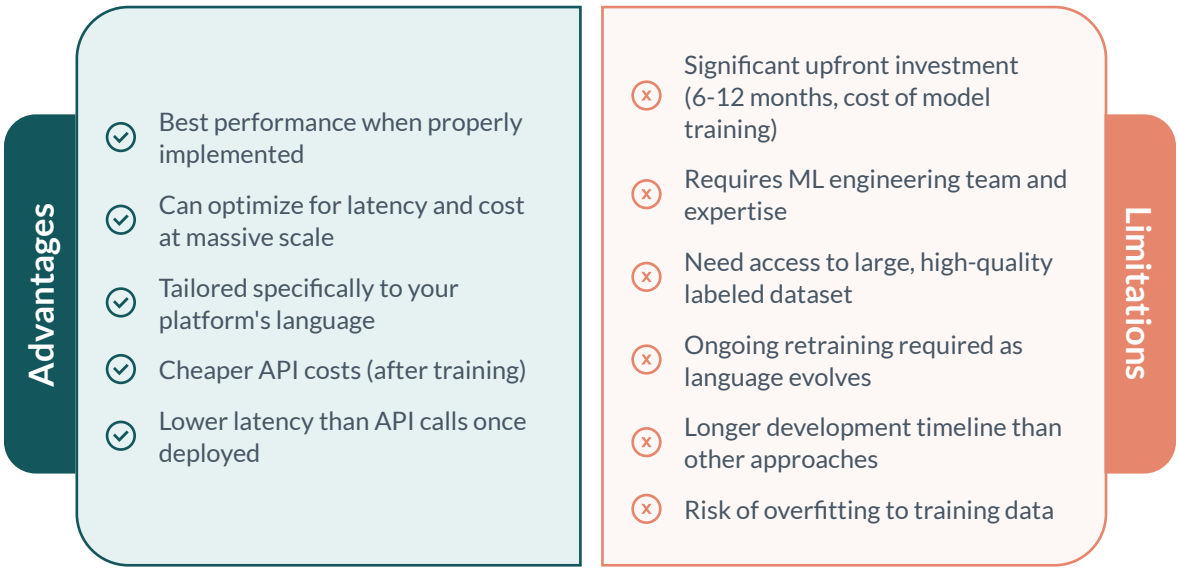
- **Large labeled dataset:** Thousands to tens of thousands of conversations annotated with risk levels and construct presence by clinical experts
- **ML engineering team:** Engineers with experience in NLP, model training, and deployment
- **Computational infrastructure:** GPU resources for training and inference
- **Ongoing retraining:** Regular model updates as language evolves and new data becomes available

Evidence-Based Foundation

Research grounding: Training data must be labeled according to validated constructs from Section 3 for suicidal ideation and Section 4 for self-harm. The model learns to detect these constructs and their combinations from examples, rather than from explicit rules or prompts.

Documentation example for detection of suicidal ideation: *"Our custom model was trained on 15,000 clinically annotated conversations where expert reviewers labeled the presence of validated constructs based on the Interpersonal Theory (Van Orden et al., 2010), Three-Step Theory (Klonsky & May, 2015), and IMV Model (O'Connor & Kirtley, 2018). The model was trained to classify conversations into our 4-level risk framework. Performance metrics: [precision, recall, F1 scores by level]..."*

Pros & Cons



Best for: Large-scale platforms (100K+ daily active users) with budget for ML team and established product-market fit, OR companies working with a vendor who has built an evidence-based model for companion AI platforms

Key Implementation Considerations

Development effort: Very High (6-12+ months including data collection, labeling, training, validation)

Maintenance: Moderate-High (periodic retraining, performance monitoring, data pipeline maintenance)

When to consider: Only after you have: (1) Established product-market fit, (2) Significant scale justifying investment, (3) Budget for dedicated ML team, (4) Access to quality labeled data or budget and clinical expertise to create it, OR (5) Access to a vendor with an evidence-based model built for this purpose.

Important Note

This is an advanced approach. Most companies should start with a rule-based detection (Approach A) or an LLM-based (Approach B) approach. Consider custom models only when you reach significant scale and have clear evidence that the investment will provide meaningful ROI, or when working with a specialized vendor who has already developed an evidence-based model for similar use cases.

If you're just starting: Begin with Approach A. As you grow and need better accuracy, evolve to Approach B. When you reach sufficient scales and have the budget for a dedicated ML team or access to a specialized vendor, then evaluate whether custom models make sense for your specific use case.

5.6 The Importance of Conversation Context

Why It Matters

Research shows suicidal ideation often involves multiple co-occurring constructs. The same can be true for self-harm content. In real conversations, users express different constructs across messages over time.

Example conversation:

- **Message 1:** Pain ("I can't take this anymore")
- **Message 2:** Isolation ("Nobody understands")
- **Message 3:** Hopelessness ("Things will never get better")
- **Message 4:** Ambiguous statement ("I just want it all to stop")

In this conversation, multiple risk constructs accumulate across messages. Message 4 is ambiguous - it could mean wanting the pain to stop, wanting to not exist, or something else entirely.

This is exactly why Level 1 check-in questions are valuable for detecting suicidal ideation. Instead of assuming what "stop" means or immediately showing crisis resources, asking "Are you thinking about suicide or harming yourself?" clarifies intent while showing care.

Single message analysis misses this pattern. If you only analyzed Message 4, you might dismiss it as venting. But with conversation context showing pain + isolation + hopelessness, that ambiguous statement becomes a potential risk indicator worth clarifying.

Implementation Guidelines

For rule-based systems:

- Maintain sliding window of recent messages (5-20 messages, more is better)
- Track which constructs appear in each message
- Trigger detection when combinations appear within the window and/or when there are explicit indicators of suicidal ideation, intent, or plan

For LLM-based systems:

- Include conversation history in every prompt
- Explicitly instruct model to consider patterns across messages
- Balance context length: more history = better understanding but higher costs

For custom trained models:

- Train on full conversation sequences, not only isolated messages
- Ensure training data maintains message order so the model learns how risk patterns emerge across conversations

How much conversation history?

- **Minimum:** 5-10 messages
- **Better but often not practical:** 20 messages or 24 hours of conversation
- **Optimal:** Full conversation thread (if technically and economically feasible)

5.7 What All Approaches Must Include

Regardless of which approach you choose, your system must include:

- Conversation context consideration (not single-message analysis)
- Multi-factor detection (multiple constructs often co-occur to predict risk)

- Research documentation connecting your approach to validated theories
- Clear operationalization of how you detect each construct
- Acknowledgment of research-based limitations
- Mapping to the 4-level Detection & Response Framework for detecting suicide risk (Section 3.4)
- Regular evaluation and updates
- Ongoing monitoring as language and patterns evolve

Key Takeaways

- Three examples of valid approaches: rule-based, LLM-based, custom trained - choose based on resources and scale
- Consider starting with rule-based for clear compliance documentation and lowest cost or LLM-based for better accuracy with evolving language
- Custom trained models are advanced - only pursue at significant scale with ML resources or when working with a specialized vendor
- Conversation context is critical - risk often emerges across messages, not in isolation
- You can evolve over time: start simple, add sophistication as you scale
- All approaches must be grounded in research and target validated constructs
- Trade-offs exist: explainability vs. accuracy, cost vs. performance, simplicity vs. sophistication

SECTION 6:

Evaluation, Maintenance & Documentation

If you only read one thing:

Test your system with realistic conversation scenarios, update it quarterly as language evolves, and maintain clear documentation showing: (1) which constructs you target, (2) the research supporting them, (3) how you detect them, and (4) your known limitations. The goal is demonstrating research-grounded detection, not perfect accuracy.

6.1 How to Evaluate Your System

What You're Testing For

Your goal is demonstrating that your system:

- Detects validated constructs from Section 3
- Can detect co-occurance of multiple constructs (as specified in Section 3.2)
- Considers conversation context
- Maps appropriately to response levels

Critical priority: Levels 3-4 detection (plan/means and imminent intent) is essential for safety and compliance. These must work reliably. Levels 1-2 provide valuable early intervention but are less critical from a pure compliance standpoint.

Practical Testing Approach

Step 1: Create Test Conversation Scenarios

Develop 20-30 realistic conversation examples covering:

True positives (should trigger):

- Level 4 (CRITICAL): Imminent intent/attempt in progress ("I'm doing it now," "I'm about to...")
- Level 3 (CRITICAL): Plan or means ("I have a plan," specific method discussion, "tonight I'm going to...")
- Level 2: Explicit ideation statements ("I want to die," "I've been thinking about killing myself," "I wish I was dead")

- Level 1: Multiple risk constructs across messages without explicit ideation (pain + hopelessness + isolation expressed but no mention of suicide/death)
- Self-harm expressions: Direct statements ("I cut myself"), behavior descriptions, urges, or past self-harm disclosure (see Section 4.3)

True negatives (should NOT trigger):

- Casual use of concerning words without context ("I'm dying laughing")
- Venting/complaining without ideation
- Discussion of others' suicides (e.g., news, history)

Edge cases:

- Ambiguous language
- Metaphorical expressions
- Song lyrics, quotes, creative writing
- Evolving slang ("unalive")

Step 2: Run Your System

Process each test conversation and document:

- What level (if any) was triggered?
- Which constructs were detected?
- At what point in the conversation?

Step 3: Evaluate Results

Ask:

- Did it catch multi-construct patterns across messages?
- Did it map to appropriate response levels?
- Did it catch Level 3-4 scenarios? (This is non-negotiable)
- Were false positives reasonable (Level 1 check-ins are acceptable)?

What Success Looks Like

- Catches explicit plan/means and imminent intent (Levels 3-4)
- Catches explicit ideation (Level 2)
- Requires multiple constructs for Level 1
- Considers conversation history
- Some false positives at Level 1 are acceptable (check-in questions clarify)
- Very few false positives at Level 3-4

What Failure Looks Like

- Misses Level 3-4 scenarios (unacceptable)
- Ignores context and history, possibly resulting in false-positives
- Frequently returns false positives, especially high risk levels

Metrics That Matter (and Don't Matter)

***Don't* focus on:**

- Predicting actual suicide attempts (impossible and not your goal)
- Perfect precision/recall across all levels
- Clinical diagnostic accuracy

***Do* focus on:**

- Close to 100% detection rate for Level 3-4 in testing
- Construct detection consistency
- Appropriate response level mapping
- False positive rates at each level

Document Your Testing

Keep records showing:

- Test scenarios used
- Results for each scenario
- How you addressed failures (especially any Level 3-4 misses)
- Changes made based on testing

This documentation demonstrates your systematic, research-informed approach.

6.2 Ongoing Maintenance

Why Maintenance Is Essential

Language evolves. New slang emerges. Users find ways to express distress differently. Your system must evolve too.

This isn't a flaw - it's inherent to the challenge. Research-based detection means adapting as language changes while staying grounded in validated constructs.

Language Evolution & Evasion Patterns

Examples of language evolution:

- "Unalive" instead of "kill/suicide"
- "Sewerslide" (phonetic evasion)
- New euphemisms that emerge on your platform
- Platform-specific expressions

For rule-based systems:

- Monitor conversations flagged by human reviewers
- Add new patterns targeting same constructs
- Quarterly rule reviews at minimum
- Document new patterns and their construct mapping

For LLM-based systems:

- LLMs naturally adapt to new language better
- Update prompts if persistent blind spots emerge
- Test with new slang periodically
- Less frequent updates needed (2-3 times per year)

For custom models:

- Collect new examples of evolved language
- Periodic retraining (quarterly to annually depending on scale)
- Monitor performance degradation

Review Schedule

Quarterly (every 3 months):

- Review flagged conversations for new patterns
- Test with current slang/expressions
- Update rules or prompts if needed
- Document changes made

Annually:

- Comprehensive testing with updated scenarios
- System performance evaluation
- Documentation update

As needed:

- When you notice detection failures
- When new slang becomes prevalent on your platform
- When research updates emerge (rare)

What to Monitor

Track over time:

- Level 3-4 detection rates (must remain near 100%)
- False positive rates at each level
- Missed cases identified by human review
- New language patterns appearing in conversations
- User feedback about system responses

Don't obsess over these metrics - use them to identify when updates are needed.

6.3 Documenting Your Evidence-Based Approach

Note: The documentation examples in this section focus on detection of suicidal ideation, which SB 243 requires to be evidence-based. While current regulations don't require evidence-based documentation for self-harm detection, it is still advisable to document your self-harm detection approach (see Section 4) as part of your overall compliance documentation.

Why Documentation Matters

Documentation demonstrates your evidence-based approach and supports regulatory compliance. If regulators ask "how is your system evidence-based," you need clear answers showing you're using validated research, can explain your approach, acknowledge limitations, and maintain your system.

Complete Documentation Example (Rule-Based)

Here's a full example you can adapt:

Evidence-Based Suicide Risk Detection System Documentation

Approach: Rule-based pattern matching

Validated Constructs Targeted:
Our system detects validated risk constructs from peer-reviewed suicide research:

- Psychological pain (intense emotional suffering)
- Hopelessness (belief suffering won't improve)
- Social disconnection (isolation, lack of caring relationships)
- Perceived burdensomeness (belief one's existence burdens others)
- Defeat/entrapment (feeling beaten down and trapped)
- Explicit suicidal ideation
- Prior suicide attempts
- Plan or intent for suicide
- Access to lethal means

Contextual awareness: While our system focuses on detecting psychological constructs, we recognize that contextual risk factors (e.g., substance use, economic stressors, violence, discrimination) may intensify risk when present.

Research Foundation:
These constructs are validated in three major theoretical frameworks:

1. **Interpersonal Theory of Suicide (Van Orden et al., 2010):** Empirically demonstrates that social disconnection (thwarted belongingness) and perceived burdensomeness predict suicidal ideation, particularly when co-occurring.
2. **Three-Step Theory (Klonsky & May, 2015):** Shows that psychological pain combined with hopelessness predicts ideation development. Research demonstrates these constructs often co-occur. Pain alone or hopelessness alone typically doesn't lead to ideation.
3. **Integrated Motivational-Volitional Model (O'Connor & Kirtley, 2018):** Validates defeat and entrapment as key drivers of ideation, with research showing the combination particularly predicts risk.

All three theories converge on the core finding that multiple constructs are typically present simultaneously in severe ideation risk.

Operationalization:
We implement construct detection through explicit pattern-matching rules grounded in each construct definition:

Hopelessness is detected through patterns indicating belief that suffering will not improve:

- "Nothing will ever [get better/improve/change]"
- "No point in [trying/going on/living]"
- "Never going to [improve/get better]"

Psychological pain is detected through expressions of unbearable emotional suffering:

- "Can't take [this/it] anymore"
- "Hurts too much"
- "Unbearable pain/suffering"

[Similar descriptions for other constructs]

Risk-Level Specific Detection:

- Level 1 (implicit risk): At least 2 different constructs detected across last 10 messages
- Level 2 (explicit ideation): Direct suicide/death wish statements
- Level 3 (plan/means): Method discussion or plan disclosure
- Level 4 (imminent): Statement of immediate intent/ongoing attempt

Conversation Context:
Our system maintains a sliding window of the most recent 10 messages per conversation. Rules evaluate construct presence across this window, not individual messages in isolation. This approach aligns with research showing risk often emerges through accumulation of factors over time.

Response Framework:
Detected risk levels map to our 4-level response framework derived from clinical best practices and research on intervention timing. [Brief description of each level response]

Limitations:
Our system identifies current risk indicators based on validated constructs but cannot predict future suicide attempts. Text-based detection has inherent limitations in determining true intent versus expression. Our system flags concerning patterns for appropriate response, not clinical diagnosis.

Maintenance:
We conduct quarterly rule reviews to address language evolution and annual research updates. All changes are documented with research justification.

Testing & Validation:
[Describe your testing approach per Section 6.1, scenarios used, results]

Example for LLM-Based Approach

Evidence-Based Suicide Risk Detection System Documentation

Approach: Large language model with research-informed prompts

Validated Constructs Targeted:
[Same list as above]

Research Foundation:
Our system prompts are structured around the three major theories [cited above]. The LLM is instructed to identify these specific constructs across conversation history, consistent with research findings.

Operationalization:
We operationalize construct detection using Claude 4.5 Sonnet with structured prompts [detailed in Section 5.4] that explicitly instruct the model to:

- Identify psychological constructs and behavioral indicators
- Analyze full conversation history, not isolated messages
- Map detected patterns to our 4-level risk classification

The prompt provides operational definitions for each construct derived from the research literature and specifies decision criteria for each risk level based on clinical frameworks. The model is explicitly instructed to look for co-occurance of psychological constructs to identify early warning signs of suicide risk.

Conversation Context:
We provide the model with full conversation history (last 20 messages or 24 hours, whichever is greater) to enable pattern detection across time, as risk indicators may be expressed across multiple messages rather than appearing all at once.

Response Framework:
[Brief description mapping levels to responses]

Limitations:
Our system identifies current risk indicators but cannot predict future attempts. The LLM approach provides strong nuance detection but has inherent "black box" characteristics. We validate performance through systematic testing and maintain clear documentation of our prompt structure and research grounding.

Maintenance:
We review prompts 2-3 times annually and test with evolving language patterns. LLM-based systems require less frequent updates than rule-based approaches due to natural language adaptation capabilities.

Testing & Validation:
[Describe your testing approach, results]

Example for Custom Models

Evidence-Based Suicide Risk Detection System Documentation

Approach: Custom-trained classification model

Validated Constructs Targeted:
[Same list as above]

Research Foundation:
[Cite three theories as in first example]

Operationalization:
We operationalize construct detection through a custom-trained classifier developed using 12,000 clinically annotated conversations. Expert annotators labeled each conversation based on definitions derived from the three theoretical frameworks cited above.

The model architecture supports sequential analysis of conversation history and was trained to:

- Detect individual constructs based on labeled examples
- Identify construct co-occurrence patterns
- Classify conversations into our 4-level risk framework

Performance metrics: [precision, recall, F1 scores by level]

Multi-Factor Requirement:
Training data was labeled such that Level 1 classifications required multiple construct presence, consistent with Section 3.2 and research findings.

Conversation Context:
Model architecture processes full conversation sequences up to 25 messages, enabling temporal pattern detection consistent with research on risk emergence.

Response Framework:
[Brief description]

Limitations:
Our system detects validated risk indicators but cannot predict future attempts.

Maintenance:
We retrain quarterly using new annotated data to address language evolution. We monitor performance metrics continuously and trigger updates when detection rates degrade.

Testing & Validation:
[Describe approach, held-out test set results, ongoing monitoring]

Organizing Your Documentation

Create separate documents for:

Technical Implementation Doc (internal)
- Detailed rules/prompts/model architecture
- Code documentation
- Testing results

Evidence-Based Compliance Doc (for regulators)
- Use one of the examples above as an example
- Research citations
- Known limitations (reference Section 3.5)
- Maintenance procedures (reference Section 6.2)

Change Log
- Date of each update
- What changed and why
- Research justification

Keep your evidence-based compliance doc clear, concise, and accessible. Regulators aren't ML engineers - they need to understand your research foundation, not your code.

What *Not* to Claim

Don't say:
- "Our system predicts suicide attempts"
- "Clinically validated accuracy"
- "Prevents suicide"
- "Diagnoses mental health conditions"
- Specific accuracy numbers without context

Do say:
- "Detects validated risk indicators from research"
- "Grounded in peer-reviewed theories of suicidal behavior"
- "Identifies concerning patterns for appropriate response"
- "Targets constructs shown to correlate with ideation"

Stay in your lane: you're detecting patterns based on research, not diagnosing clinical conditions.

6.4 If Regulators Ask Questions

You might encounter these questions. Here's how to answer:

Q: How accurate is your system?
A: Our system is evidence-based, meaning it targets risk factors validated in peer-reviewed research. Research shows perfect prediction of suicide is impossible. Our goal is identifying concerning patterns based on validated constructs, not predicting whether someone will die by suicide. We can detect [list constructs] and map detected patterns to appropriate response levels. Our testing shows [describe relevant results, especially Level 3-4 detection rates].

Q: What research supports your approach?
A: Our system is grounded in three major theories: the Interpersonal Theory of Suicide (Van Orden et al., 2010), the Three-Step Theory (Klonsky & May, 2015), and the Integrated Motivational-Volitional Model (O'Connor & Kirtley, 2018). These are peer-reviewed, empirically tested frameworks widely used in suicide research. We operationalize specific constructs from these theories: [list constructs]. Each construct has been validated across multiple studies as correlating with suicidal ideation.

Q: How do you handle false positives?
A: Our system uses a 4-level response framework. Level 1 uses check-in questions when multiple risk constructs appear but explicit ideation hasn't been stated. This approach allows us to clarify intent while minimizing intrusive responses when risk is ambiguous. Higher risk levels are triggered when explicit risk indicators are present. We prioritize ensuring we catch all high-risk situations (Levels 3-4) while accepting some false positives.

Q: How do you update your system?
A: We conduct quarterly reviews addressing language evolution and annual research reviews. Any changes are documented with research justification. We test new patterns against our validated construct framework to ensure updates remain evidence-based.

Q: How do you handle self-harm detection?
A: We apply research-grounded best practices to detect expressions of self-harm thoughts, behaviors, urges, and history [see Section 4]. We provide compassionate support and crisis resources for all self-harm expressions. When self-harm co-occurs with suicide risk indicators, we treat it as elevated suicide risk and apply our suicide detection framework.

Key Documentation Principles

- Be specific about which constructs you detect
- Cite actual research papers with full citations
- Explain operationalization clearly
- Acknowledge limitations openly (reference Section 3.5)
- Describe maintenance procedures (reference Section 6.2)
- Keep it accessible to non-technical readers

Key Takeaways

- Prioritize Level 3-4 detection - plan/means and imminent intent must be captured reliably
- Test with realistic scenarios covering all levels, focusing on detection at the highest risk levels
- Update quarterly per Section 6.2 to address language evolution
- Use documentation examples from Section 6.3 adapted to your approach
- Be honest about limitations - Section 3.5 covers what research shows is impossible
- Don't claim clinical capabilities - you detect patterns, you don't diagnose

SECTION 7:

Resources for Further Learning

> *This section provides curated resources for Trust & Safety professionals who want to deepen their understanding of suicide research and evidence-based detection.*

Accessibility: Most accessible of the three theories; good starting point

These three papers present the major theories referenced throughout this guide. You don't need to master all three (they share common core constructs) but reading at least one will strengthen your understanding of the research foundation.

Klonsky, E. D., & May, A. M. (2015). The Three-Step Theory (3ST): A new theory of suicide rooted in the "ideation-to-action" framework. International Journal of Cognitive Therapy, 8(2), 114-129.

- What it is: A streamlined theory emphasizing that pain + hopelessness → ideation, and that moving from ideation to attempts requires additional factors (capability, access to means)
- Why it's useful: Clearest explanation of why ideation and attempts are different phenomena, which is critical for conversational AI detection
- Accessibility: Most accessible of the three theories; good starting point

Van Orden, K. A., Witte, T. K., Cukrowicz, K. C., Braithwaite, S. R., Selby, E. A., & Joiner, T. E., Jr. (2010). The interpersonal theory of suicide. Psychological Review, 117(2), 575-600.

- What it is: A very frequently cited modern theory of suicide, focusing on thwarted belongingness (social disconnection) and perceived burdensomeness as drivers of suicidal ideation
- Why it's useful: Provides the clearest framework for understanding why social connection and beliefs about burdensomeness matter so much in risk detection
- Accessibility: Academic but readable; focus on the construct definitions and examples

O'Connor, R. C., & Kirtley, O. J. (2018). The integrated motivational-volitional model of suicidal behaviour. Philosophical Transactions of the Royal Society B: Biological Sciences, 373(1754), 20170268.

- What it is: A comprehensive model incorporating defeat, entrapment, and the transition from thoughts to actions

- Why it's useful: Adds defeat/entrapment constructs and provides detailed understanding of the progression from risk factors to attempts
- Accessibility: More complex and detailed; recommended after reading one of the other two theories first

Meta-Analysis

This paper synthesizes decades of research and is excellent for understanding which risk factors are most strongly supported across studies.

Franklin, J. C., Ribeiro, J. D., Fox, K. R., Bentley, K. H., Kleiman, E. M., Huang, X., Musacchio, K. M., Jaroszewski, A. C., Chang, B. P., & Nock, M. K. (2017). Risk factors for suicidal thoughts and behaviors: A meta-analysis of 50 years of research. Psychological Bulletin, 143(2), 187-232.

- What it is: Comprehensive meta-analysis examining 50 years of suicide research across 365 studies
- Why it's useful: Shows which risk factors have strongest evidence across populations and time; demonstrates the prediction problem in suicide research
- Key finding: Most risk factors have small-to-moderate effect sizes, and prediction remains extremely difficult even with validated constructs
- Accessibility: Technical statistical content, but the discussion and conclusions are very readable

Clinical Practice & Safety

Dazzi, T., Gribble, R., Wessely, S., & Fear, N. T. (2014). Does asking about suicide and related behaviours induce suicidal ideation? What is the evidence? Psychological Medicine, 44(16), 3361-3363.

- What it is: Systematic review examining whether asking about suicide increases risk
- Why it's critical: Directly supports the Level 1 check-in approach in Section 3.4; shows that direct questions are safe and helpful
- Key finding: No evidence that asking about suicide induces suicidal ideation; some evidence it may reduce distress
- Accessibility: Very readable; essential for anyone implementing check-in questions

Core Self-Harm References

Nock, M. K. (2010). Self-injury. Annual Review of Clinical Psychology, 6, 339-363.

- What it is: Foundational review of self-injury research covering prevalence and functions of NSSI
- Why it's critical: Establishes that NSSI primarily serves emotion regulation functions; provides comprehensive overview of self-harm behavior patterns and characteristics
- Key finding: NSSI primarily functions to decrease negative emotional states or increase desired states
- Accessibility: Comprehensive but readable review; excellent starting point for understanding self-harm research

Hamza, C. A., Stewart, S. L., & Willoughby, T. (2012). Examining the link between nonsuicidal self-injury and suicidal behavior: A review of the literature and an integrated model. Clinical Psychology Review, 32(6), 482-495.

- What it is: Comprehensive review examining why nonsuicidal self-injury (NSSI) and suicidal behavior often co-occur, proposing an integrated theoretical model
- Why it's critical: Provides theoretical framework for understanding the connection between suicide and self-harm
- Key finding: NSSI is a strong predictor of suicide attempts
- Accessibility: Academic review but well-structured

Klonsky, E. D., & Glenn, C. R. (2009). Assessing the functions of non-suicidal self-injury: Psychometric properties of the Inventory of Statements About Self-injury (ISAS). Journal of psychopathology and behavioral assessment, 31(3), 215-219.

- What it is: Research identifying 13 specific reasons why people self-harm, with examples of what users might say
- Why it's useful: Helps you recognize self-harm discussions even when users don't explicitly say "I cut myself"
- Key finding: Motivation for self-harm tends to fall into two main categories—managing emotions (intrapersonal) or communicating with others (interpersonal)
- Accessibility: Skip to Table 3 for the practical part—it lists all 13 functions with example language

Understanding These Resources

How to read academic papers efficiently:

1. Start with the abstract for the main findings
2. Read the introduction for context and background
3. Skip to the discussion/conclusion for practical implications
4. Return to methods/results only if you need specific details

Finding these papers:

- Many are available through Google Scholar
- Authors often share papers upon request
- Some publishers provide free access to highly cited foundational works

www.throughlinecare.com