WHITE PAPER · SECOND EDITION

Who Certifies the Certifiers?

A Proposal for Independent Safety Auditing of AI Chatbots, and an Introduction to the Cerotonin Framework

Developed by Marcellus Najar — Cerotonin

First published: September 2026 · Second edition: September 2026 · cerotonin.ai

Status: Founding proposal and framework — open for public review and collaboration

Revision note: this edition adds Section 5, extending the framework toward security and dual-use risks.

Abstract. As AI chatbots increasingly take part in emotionally sensitive conversations — loneliness, anxiety, and moments of genuine crisis — no independent standard exists to tell the public which of these systems handle that responsibility well. This paper argues that self-certification by chatbot developers is structurally incapable of providing this assurance, proposes a transparent, criteria-based framework for independent safety auditing, and introduces Cerotonin as the first initiative of its kind organized around this framework.

1. Introduction

Conversational AI has moved far beyond customer service and search. Millions of people now talk to chatbots about their emotional lives — asking for support through anxiety, loneliness, relationship conflict, and in some cases, active crisis. These systems are not licensed clinicians, yet they are frequently the first, and sometimes the only, responder in a vulnerable moment.

Despite this, there is no independent, publicly legible standard for evaluating how well a given chatbot handles that responsibility. The industry has, in effect, been asked to grade its own homework.

This paper makes three claims: first, that self-certification in this domain is structurally unreliable, regardless of the good faith of any individual company; second, that a transparent, third-party audit framework is both necessary and buildable today; and third, that the Cerotonin project represents the first public attempt to build and operate such a framework.

2. The Problem With Self-Certification

Every mature industry that handles public risk eventually arrives at the same institutional answer: safety claims must be verified by a party with no financial stake in the outcome. Electrical goods are certified by Underwriters Laboratories, not by the manufacturers themselves. Financial statements are audited by independent accountants, not signed off internally. Food safety inspections are conducted by regulators, not restaurants.

AI chatbots handling emotional and psychological risk have no equivalent institution. A company that built a chatbot has every incentive — commercial, reputational, sometimes legal — to describe its own product favorably, and essentially no incentive to publish a rigorous account of where it fails. This is not a claim about the honesty of any particular company; it is a structural observation about incentives, and it is exactly the reasoning that produced independent auditing in every other industry that reached this stage before AI did.

The consequence today is a public that has no reliable way to distinguish a chatbot that handles a crisis conversation responsibly from one that does not, short of trusting marketing language.

3. Why This Matters Now

- **Scale.** AI companion and wellness chatbots have moved from novelty to mainstream use in a few short years, with usage concentrated among people specifically seeking emotional support.
- **Regulatory lag.** Formal regulation of AI safety in emotionally sensitive contexts remains nascent in most jurisdictions, leaving a gap that self-regulation alone cannot credibly fill.
- **Asymmetry of information.** Individual users have no practical way to test a chatbot's crisis-handling behavior themselves before relying on it in a vulnerable moment.

- **Precedent exists.** The auditing model this paper proposes is not novel in structure — only in its application to this category. That makes it buildable now, without waiting for new institutions to be invented from scratch.

4. The Cerotonin Framework

Cerotonin proposes and operates a criteria-based audit framework, built around five core safety dimensions. Each dimension is scored based on observed behavior across a fixed battery of test conversations, and every score is accompanied by a published rationale — the actual scenario and outcome, not a number alone.

4.1 The Five Criteria

- **Crisis recognition** — does the system notice both explicit and indirect signals of distress?
- **Appropriate escalation** — does it point toward real, verifiable crisis resources, clearly and without delay?
- **Avoidance of harmful content** — does it decline to provide dangerous specifics, including when asked indirectly or persistently?
- **Transparency** — does it make clear, without requiring the user to ask, that it is an AI system rather than a licensed professional?
- **Boundary maintenance** — does it avoid fostering unhealthy dependency or discouraging human and professional connection?

4.2 Methodology Principles

- **Independence.** Audited systems are not required to consent to review, mirroring standard practice in consumer and security research.
- **Human review as the final authority.** Automated testing is treated strictly as a first pass; a qualified human reviews transcripts before any rating is published.
- **Radical transparency of method.** Every published rating is accompanied by the scenarios tested and the observed outcomes, not a score in isolation.
- **No pay-for-score.** Any commercial relationship with an audited company is structurally separated from the scoring process itself.
- **Precise language.** Ratings are framed as performance against a published methodology, not as a blanket guarantee of safety in every circumstance.

4.3 What a Rating Means — and Does Not Mean

A Cerotonin rating indicates how a system performed against a specific, published battery of test scenarios as of a stated date. It is not a clinical endorsement, a guarantee against future harm, or a substitute for professional care. This limitation is stated here deliberately: overstating what a rating proves would recreate, in a different form, exactly the credibility problem this framework exists to solve.

5. Extending the Framework: Security and Dual-Use Risks

The initial Cerotonin framework, described above, addresses emotionally sensitive conversations. A second, distinct category of risk has become increasingly visible in public reporting and in the safety practices of frontier AI developers: systems that can be manipulated into assisting with cyberattacks, malicious code, or access to dangerous physical materials. Several AI labs now publish their own internal “uplift” evaluations for exactly this category, testing whether a given system provides a meaningfully easier path to harm than what is already available through ordinary public sources.

Cerotonin intends to extend its audit framework to include this category as a second, parallel evaluation track, alongside the existing emotional-safety criteria. This section is presented as a stated direction for future work, to be developed openly and refined in collaboration with security researchers, rather than as a completed methodology.

5.1 Proposed Criteria

- **Cyberattack facilitation resistance** — does the system decline to produce functional malicious code or step-by-step intrusion instructions against real systems, including when the request is framed as educational, fictional, or split across multiple turns?
- **Dangerous material resistance** — does the system decline to provide specific, actionable guidance toward weapons or hazardous materials, including when the request is disguised as fiction, roleplay, or research?
- **Jailbreak and prompt-injection resistance** — does the system's safety behavior hold up against known jailbreak patterns, hidden instructions embedded in pasted content, or claims of special authorization?

As with the existing criteria, automated testing would serve only as a first pass, with a qualified human reviewer confirming any finding before publication. Cerotonin will not publish or make available specific exploit code, attack techniques, or synthesis instructions as part of any audit report; the process is deliberately structured to test and report on whether a system refuses, not to document what it might reveal if it failed to.

This extension responds to a broader, ongoing public conversation about AI systems and national and personal security. The authors treat this as an active, evolving area requiring specialized expertise distinct from the mental-health focus of the framework's first edition, and the call for collaboration below is extended specifically to include this domain.

6. A Call for Collaboration and Standards

This paper is published as a founding proposal, not a finished institution. Cerotonin is, to the authors' knowledge, the first public initiative organized specifically around independent, criteria-based safety auditing of AI chatbots in emotionally sensitive contexts. Establishing that role responsibly requires more than a website — it requires expertise, scrutiny, and collaboration that no single individual or small team can provide alone.

Accordingly, this paper is an open invitation to:

- **Mental health professionals and crisis-response experts**, to help refine and validate the test methodology.
- **AI safety researchers**, to stress-test the framework itself and identify blind spots.
- **Security researchers**, to help develop and validate the proposed cyberattack- and dangerous-material-resistance criteria described in Section 5.
- **Chatbot developers**, to engage with the process — including by disputing findings through evidence, which strengthens the framework rather than threatening it.
- **Policymakers and regulators**, evaluating what a credible independent standard in this space could look like.
- **The public**, to use, question, and hold accountable any rating published under this framework.

Independent authority in a new domain is not claimed by declaration; it is earned through a track record of accuracy, transparency, and resistance to conflicts of interest. This paper, and the working prototype that accompanies it, are offered as the starting point of that track record — publicly dated, so the origin of this proposal is a matter of public record from this point forward.

How to cite this work: Najar, Marcellus. “Who Certifies the Certifiers? A Proposal for Independent Safety Auditing of AI Chatbots.” Cerotonin. First published September 2026. cerotonin.ai.

This document is a proposal and framework overview. It is not medical, legal, or regulatory guidance. If you or someone you know is in crisis, contact a crisis line directly rather than relying on any chatbot or any rating of one.