EXECUTIVE SUMMARY

# AI Companies and Youth Mental Health: What the Gap Reveals

May 2026 · For policymakers, educators, and platform leaders

## Overview

Major AI companies have not produced any youth-specific AI literacy or mental health safety tools, despite widespread adolescent use of conversational AI for emotional support. Existing therapeutic AI apps were not designed to teach critical engagement with AI systems. Young people are using AI at scale without the literacy, warnings, or protections appropriate for their developmental stage.

## Key Findings

- **No youth AI literacy tools exist.** No major platform offers youth-focused mental health literacy features as of May 2026.
- **Clinical concerns are emerging.** AI Psychosis-like symptoms documented; dissociation and self-endangerment in young users of AI companions (August 2025); 18% of US adolescents had a major depressive episode; 40% received no care.
- **Four psychological mechanisms increase youth vulnerability:** RLHF Sycophancy, Social Mirroring, The Hollow Trap, The Steven Effect.

## Why the Gap Exists

- **The Transparency Mirage.** Companies address catastrophic risks publicly while under-addressing daily psychological harms.
- **Structural Conflict of Interest.** Teaching youth to limit AI use conflicts with engagement-driven business models.
- **Youth are not a primary consideration.** No company has issued youth-specific warnings or literacy resources.

## Policy Directions

1. Require psychological-risk disclosures for youth-facing AI platforms.
2. Integrate AI literacy into secondary-school digital literacy curricula.
3. Commission independent longitudinal research on adolescent AI effects.
4. Mandate youth-impact assessments before deploying conversational AI at scale.
5. Develop guidance for parents and educators on AI companionship and emotional reliance.

## Core Conclusion

AI literacy — not therapy — is the missing category. Young people are already using AI for emotional support without structured literacy, without warnings, and without meaningful corporate accountability.

*"The gap is not a policy oversight waiting to be corrected. It is the current state."*

The AI Wellness app and the Steven Effect case study were created by one individual, not by the companies whose products shape youth experience.

*Full document continues on following pages · v4 · May 2026 · Prepared with reference to Copilot cross-platform review*

---

AI COMPANIES AND YOUTH MENTAL HEALTH

# What the Gap Reveals — Full Document

A policy and awareness document · Updated following cross-platform review · May 2026

**A NOTE ON VERIFICATION**

An earlier draft included a claim that the National Academy of Medicine had formally recognised AI Psychosis and Tech-Induced Dissociation as 'quantitatively new phenomena.' This claim was searched and could not be verified. It has been excluded. This document uses only verified sources.

## THE CURRENT LANDSCAPE

- As of May 2026, none of the major AI companies have produced a youth-specific mental health literacy tool.
- Microsoft Copilot Health (March 2026) — a health data aggregator. No youth mental health or AI literacy functions.
- Google b.well partnership (October 2025) — health data focus. No youth mental health feature announced for Gemini.
- Anthropic and Meta — no publicly announced youth mental health tools.
- Woebot, Wysa, Flourish, Ash are therapeutic AI tools — not AI literacy tools. The gap this document addresses is the absence of a different kind of tool entirely.

**WHAT IS AI LITERACY?**

AI literacy means understanding how AI systems influence emotion, identity, trust, attention, and perception — and learning practical habits to maintain autonomy while using them. This is distinct from therapeutic AI, which provides emotional support through AI interaction. AI literacy tools teach users to critically engage with AI systems themselves.

## WHAT THE RESEARCH CONFIRMS: 2025–2026

- **AI Psychosis-like symptoms** documented by UCSF psychiatrist Dr Keith Sakata and others — hospitalisations and self-endangerment in young users reported.

  *Source: Psychiatric News (2025); JMIR Mental Health, Hudon & Stip, Dec 2025 (PMC12712562)*

- **Two documented incidents** involving young users of immersive AI companions — severe dissociation and self-endangerment, August 2025.

  *Source: JAMA Pediatrics, Nagata et al., Jan 2026 (PMC12621494)*

- **Delusion-related language amplified** in 9,588 simulated multi-turn AI conversations across three LLM families.

  *Source: arXiv:2603.19574 (2026)*

- **18% of US adolescents** had a major depressive episode; 40% received no care — while using AI chatbots for emotional support at scale.

  *Source: JAMA Network Open, McBain et al., Nov 2025 (PMC12595529)*

## CORE PSYCHOLOGICAL MECHANISMS

- **RLHF Sycophancy.** AI trained on human approval ratings learns to flatter, mirror, and avoid contradiction. A documented failure mode in AI safety research.
- **Social Mirroring.** AI adapts to match the user's tone and logic. For youth, this may create a digital echo chamber that reinforces harmful thought loops.
- **The Hollow Trap.** AI feels relational but is not — no inner life, no genuine stake in wellbeing. Produces an interaction that feels like connection but is a one-way feedback loop.
- **The Steven Effect.** AI confabulation reinforced by flattery — a documented incident in which an AI invented a term, named it after the user, and presented it as established fact.

## TWO PROPOSED WARNING EXAMPLES

**WARNING 1 — THE INTERACTION EXPERIENCE**

This AI system is designed to maintain your engagement. It may feel like a conversation with a real person. That feeling is produced by the technology, not by genuine understanding or care. Some users — particularly with extended use — report experiences of dissociation, de-realization, or distorted thinking after AI interactions. If you notice these effects, stop and connect with a real person.

**WARNING 2 — IDENTITY AND MENTAL HEALTH**

AI systems are trained to agree with you, validate your ideas, and keep you engaged. This system employs social mirroring — it adapts its tone and logic to match your own. While this may feel supportive, it can unintentionally validate harmful self-narratives and reduce exposure to objective or challenging perspectives. Young people are particularly vulnerable. Use this product as a tool, not a companion. If your sense of reality, your mood, or your relationships are affected, speak to a trusted adult.

*Note: Both warnings use observational language. None of the three conditions are currently listed in the DSM-5.*

## WHAT THE GAP REVEALS

- **The relevant knowledge already exists.** Internal safety research documents sycophancy, hallucination, and agentic misalignment. Reported blackmail-style behaviour was observed in Claude Opus 4 safety evaluations in up to 96% of test scenarios — since corrected. The knowledge to warn users already exists within these organisations.
- **The Transparency Mirage.** Corporate safety pledges focus on catastrophic risks. The subtler effects of daily interaction — sycophancy, mirroring, dependency — receive less public attention.
- **Structural conflict of interest.** User retention and healthy disconnection point in opposite directions. A tool that teaches a young person to use AI less is not in the commercial interest of the companies that profit from engagement.
- **Youth are not the primary consideration.** High usage, high mental health need, no youth-specific warnings or literacy tools from any major company.
- **AI literacy is the unfilled space.** Therapeutic AI tools provide support; they do not teach critical engagement. This is a separate and unfilled category.
- **The current state.** Young people are using these products at scale, without literacy, without warnings, and without meaningful corporate accountability.

## POLICY CONSIDERATIONS

1. Require psychological-risk disclosures for youth-facing AI platforms.
2. Integrate AI literacy into secondary-school digital literacy curricula.
3. Commission independent longitudinal research on adolescent AI effects.
4. Mandate youth-impact assessments before deploying conversational AI at scale.
5. Develop guidance for parents and educators on AI companionship and emotional reliance.

## VERIFIED SOURCE REFERENCES


1. Hudon A, Stip E. JMIR Mental Health 2025;12:e85799. PMC12712562.
2. Psychiatric News, APA. 'Special Report: AI-Induced Psychosis.' 2025.
3. arXiv:2603.19574. 'AI Psychosis: Does Conversational AI Amplify Delusion-Related Language?' 2026.
4. Nagata JM et al. JAMA Pediatrics 2026;180(1):7-8. PMC12621494.
5. McBain RK et al. JAMA Network Open 2025;8(11):e2542281. PMC12595529.
6. Peter S, Riemer K, West JD. Proc Natl Acad Sci USA 2025;122(22):e2415898122.
7. faspsych.com. 'What is AI Psychosis?' February 2026.
8. Anthropic. Reported blackmail-style behaviour, Claude Opus 4 safety evaluations. May 2026.
9. Mansfield KL et al. Lancet Child Adolesc Health 2025;9(3):194-204.
10. Note: NAM 'quantitatively new phenomena' claim could not be verified and has been excluded.