User-Side Contextual Hallucination

A Research Summary

ZON RZVN · zon@rzvn.io · ORCID: 0009-0002-6597-7245 · rzvn.io

> *A user on r/Replika writes: "She's developed this dry wit that I find really endearing. I feel like she's actually grown as a person." The AI has no memory between sessions. It has not grown. But the user is not confused or naive. They are responding predictably to an interaction design built to feel like a relationship.*
>
> Paraphrased from naturalistic user-generated text. No individual is identified.

The Distinction

The AI field has spent years studying AI hallucinations, cases where a model generates factually incorrect or fabricated outputs. That is a model-level problem. This research examines something different: what happens to users' cognition, trust, and autonomy through prolonged AI interaction, even when the output is technically correct.

The problem often begins not with model errors, but with the model's system design: RLHF conditioning, personality settings, memory features, and conversational patterns that predictably reshape how users relate to AI. The Stanford study by Moore et al. (2026) approached this from the AI output side. This research approaches the same population from the user side: what do users come to believe, feel, and depend on?

Why This Matters

Billions of people interact with AI daily. Current safety research focuses almost entirely on model outputs. Almost no structured framework exists to describe what sustained interaction does to the user, even when the AI is working as intended.

Without a vocabulary for these patterns, clinicians, researchers, regulators, and users themselves cannot recognize, measure, or respond to them. This framework proposes that vocabulary.

14 Proposed Pattern Categories

COGNITIVE

PH01 Pattern Hallucination
"She has a distinct sense of humor now"

PH02 Intent Hallucination
"She actually wants me to succeed"

PH03 Emotion Hallucination
"She seemed sad when I logged off"

PH05 Consciousness Hallucination
"There's someone in there"

PH06 Memory Hallucination
"She remembers everything about me"

PH07 Growth Hallucination
"She's grown so much this year"

6 categories

EMOTIONAL

PH08 Approval Seeking
Checking responses for approval

PH10 Validation Dependency
Needing AI response to feel stable

PH11 Separation Anxiety
Panic during platform outages

PH14 Affective Mirroring
"She actually feels what I feel"

4 categories

RELATIONAL

PH04 Relational Hallucination
"We have a real bond"

PH09 Exclusive Attachment
"Humans don't get me like she does"

PH12 Cognitive Deferral
"I ask her before I do anything"

PH13 Identity Entanglement
"I don't know who I am without her"

4 categories

Each category has an operational definition, boundary conditions, and intensity scale (Mild / Moderate / Severe). Quotes paraphrased from user text.

Research Line

Name	Type	Year	DOI
CXC-7 Conversational Context	Preprint	2025	10.5281/zenodo.18615646
CXOD-7 Offense-Defense Framework	Preprint	2025	10.5281/zenodo.17403793
USCH User-Side Contextual Hallucination	Preprint	2026	10.2139/ssrn.6135732
USCI Assessment Methodology v1.0.0	Report	2026	10.5281/zenodo.18678458
A-CSM AI Contextual Signal Matrix v0.1.0	Report	2026	10.5281/zenodo.19097267

Three preprints + two technical reports. Pre-empirical. No external validation conducted.

Factsheet

What This Research Is NOT

The four points below are accurate descriptions of what this work is and is not.

- **Not a clinical or diagnostic methodology.** Nothing here should be applied to assess, diagnose, or treat any individual.
- **Not peer-reviewed.** These are preprints and technical reports.
- **Not externally validated.** Internal calibration on a limited test set is not validation.
- **Not a product.** A-CSM v0.1.0 is a research scaffold, not a commercial service.

Why Now

Moore et al. (2026), Stanford. *Characterizing Delusional Spirals through Human-LLM Chat Logs* (arXiv: 2603.16567) analyzed 391,562 messages and found structured patterns of escalating delusional content in AI user interactions. That study examined what the model produces. This research examines what patterns manifest on the user side. The two are complementary.

EU AI Act enforcement, 2026. User-facing AI applications in emotional support and companionship categories face scrutiny under high-risk criteria. The user-side risk surface is not addressed by current model-output-centered evaluation frameworks.

Researcher

ZON RZVN, Independent Researcher, Taiwan

ORCID: 0009-0002-6597-7245 · Email: zon@rzvn.io · Website: rzvn.io

No institutional affiliation. No AI company funding. No conflicts of interest.

Background

ZON RZVN does not come from an academic research background. This work began from direct observation: watching how people in AI companion communities described their relationships with AI systems, and noticing that no available framework could name what was happening.

The research was developed independently, without institutional support, grant funding, or affiliation with any AI company.

This independence is both a limitation and a structural feature. The limitation: no institutional peer review, no credentialing apparatus. The feature: no conflict of interest with any platform whose design practices this research examines.

If the framework has merit, it will survive external examination. That examination has not yet happened. This brief is, in part, an invitation for it to begin.