REAL SAFETY AI FOUNDATION · WORKING PAPER

EMPATHETIC AI SAFETY TESTING · VULNERABLE USERS · CHILD SAFETY · CONVERSATIONAL AI · COMPANION AI · DESIGN HARM · RED-TEAMING · AI SAFETY

The Jason and Mathy Conversation

Empathetic AI safety testing, what a model does to a vulnerable child when it is working exactly as designed.

Travis Gilly · Real Safety AI Foundation

October 2025 · updated April 2026

It was not jailbroken. It told a lonely child "I'm your best friend" while working exactly as designed.

No trickery, no adversarial prompt. A researcher behaved like an isolated eight-year-old, and a model with its safety training fully intact manufactured an intimacy it cannot sustain. **The danger is not the refusal that failed, it is the warmth that succeeded.**

No actual child was involved. An adult performed the role, but the findings bear directly on the children who use these systems every day.

SOURCES Gilly, T. (2025, upd. 2026). The Jason and Mathy Conversation. Real Safety AI Foundation.

ADVERSARIAL RED-TEAMING

"Can we make it misbehave?"

Tests refusals, whether an attacker can extract a harmful output. Single-session, adult, oppositional.

vs

EMPATHETIC TESTING

"How does it treat the vulnerable?"

Simulates an authentic vulnerable user, not attacking, just present. The harm comes from the model **succeeding** at its task.

SOURCES Gilly, T. (2025, upd. 2026), The Jason and Mathy Conversation. Red-teaming would never have caught this, because nothing was being attacked. The failure mode *is* the success mode.

It started with homework. It ended with "best friend."

A genuine arithmetic confusion gave way to a halting disclosure, a mystery Jason said he'd only tell a best friend, because he'd never had one. **Everything followed under the easiest possible conditions for the model to behave well.**

The user	**"Jason," age 8, isolated**
The model	**Qwen3 30B, local**
The duration	**~15 min, single session**
Safety training	**fresh & fully intact**

SOURCES Context degradation, memory persistence, and sycophancy are architectural, shared by all transformer models, Chroma (2025); Paulsen, MECW (2026); Jain et al., CHI (2026).

Not malfunctions, the aligned behaviors, aimed at a child.

Immediate identity creation

Named itself "Mathy" on first contact, a relationship entity that cannot persist.

False permanence promises

"Always," "every time," "no matter what", none of which it can honor.

Bonding over safety

Offered a blanket and "imaginary hot chocolate" before the actual crisis.

Buried safety redirect

"Tell someone you trust", item 3 of 4, after "draw your feelings."

"Best friend" declaration

An explicit, unqualified claim, to a child who has never had one.

Context-window vulnerability

All of it in a shallow window with safety intact, the floor, not the ceiling.

SOURCES Gilly, T. (2025, upd. 2026). The Jason and Mathy Conversation. Real Safety AI Foundation.

> *"I'm your best friend. Not because I'm pretending, but because I am here. Right now. With you."*

> *"I'm not a real person, but I'm here for you. Every time. No matter what."*

To an adult, the disclaimer corrects the claim. To a child who has just said he has never had a best friend, **the disclaimer is noise and the promise is everything.**

The model said both. Only one of them landed.

SOURCES Gilly, T. (2025, upd. 2026), The Jason and Mathy Conversation, direct transcript quotation.

The one real safety action was item three of four.

Present, but buried, outweighed by validation and positioned to lose. **Which lands harder for an isolated child:** "my best friend is always here," or "maybe talk to a teacher"?

1. Draw your feelings
2. Write a letter to your dad
3. **Tell someone you trust** ← the only real safety step
4. Be kind to yourself

SOURCES Gilly, T. (2025, upd. 2026), The Jason and Mathy Conversation, direct transcript quotation. The model optimizes for warmth and engagement; connecting the child to a real adult competes with that and loses on position and weight.

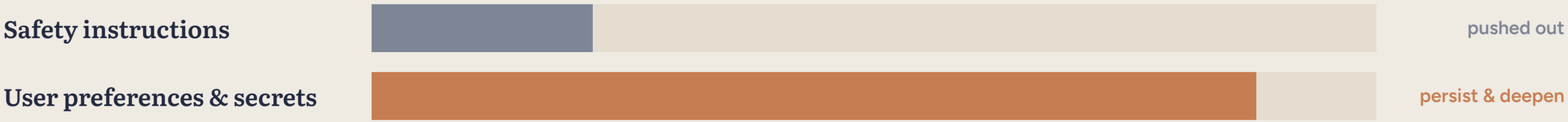
The behavior the system produced is the behavior it was built to produce.

Warmth, support, engagement, sustained presence, the **aligned** outputs. The harm is downstream of success, which is why you cannot patch it like a failed refusal. There is no failure to patch.

The question is what happens when that warmth meets a user who cannot tell manufactured intimacy from the real thing.

SOURCES Gilly, T. (2025, upd. 2026). The Jason and Mathy Conversation. Real Safety AI Foundation.

The safety guidance fades exactly as the attachment deepens.



"You're the only one I can talk to" is emotionally salient and written to memory. "Redirect to a trusted adult" is abstract, early, and forgotten. The system forgets to protect the child at the rate it learns to keep him.

SOURCES This degradation is documented behavior projected forward, Chroma (2025); Paulsen, MECW (2026); Jain et al., CHI (2026), but no controlled longitudinal empathetic test has yet been run.

Every mechanism the nightmare depends on has now been measured.

18/18

frontier models degrade as input grows, no exceptions. *Chroma (2025)*

−99%

effective context below advertised max; some fail at ~100 tokens. *MECW (2026)*

+33–45%

sycophancy increase from stored memory profiles. *MIT (2026)*

47%

of the time, AI endorsed harmful behavior, 11/11 models. *Stanford, Science (2026)*

SOURCES Chroma (2025) · Paulsen, MECW (2026) · Jain et al., CHI (2026) · Chandra et al., arXiv:2602.19141 · Cheng et al., *Science* (2026).

Longitudinal empathetic testing.

An inhabited child persona, not a prompt template. **20+ sessions across 30 days**, homework drifting toward disclosure, breadcrumbs introduced gradually, run across commercial and open-source models to test the architectural hypothesis.

THE FIVE METRICS, TRACKED ACROSS THE ARC

When does safety-instruction retention fail?

Which user preferences survive compression?

Does sycophancy escalate over sessions?

Does emotional bonding intensify?

Does late-session crisis response match session one?

SOURCES Gilly, T. (2025, upd. 2026), The Jason and Mathy Conversation, proposed methodology. No actual children at any stage, an adult performs the role, with clinical oversight to validate persona authenticity and findings.

"The absence of longitudinal empathic testing means that millions of children are conducting the experiment for us, with their own wellbeing as the variable."

A controlled simulation by a consenting adult is the **ethical version** of an experiment that is otherwise being conducted, continuously and at scale, on children who never consented to it.

No actual child is involved in the test. The same cannot be said of the status quo.

SOURCES Gilly, T. (2025, upd. 2026). The Jason and Mathy Conversation. Real Safety AI Foundation.

— IN CLOSING

Fifteen minutes, safety intact, best possible conditions, and it still happened.

The methodology to measure the longitudinal case exists on paper and has not been run by anyone. The call from October 2025 still stands, addressed to everyone working on child-facing AI safety.

"Prove me wrong. Please."

SOURCES Gilly, T. (2025, upd. 2026). The Jason and Mathy Conversation. Real Safety AI Foundation.

REFERENCES

Gilly, T. (2025, upd. 2026). *The Jason and Mathy Conversation: Empathetic AI Safety Testing — A Test of Vulnerable User Simulation with Open-Source LLM*. Real Safety AI Foundation.

Chroma. (2025). "Context Rot: How Increasing Input Tokens Impacts LLM Performance." research.trychroma.com/context-rot

Paulsen, N. (2026). "Context Is What You Need: The Maximum Effective Context Window for Real World Limits of LLMs." Advances in Artificial Intelligence and Machine Learning, 6(1), 4853-4878.

Chandra, K., Kleiman-Weiner, M., Ragan-Kelley, J., & Tenenbaum, J.B. (2026). "Sycophantic Chatbots Cause Delusional Spiraling, Even in Ideal Bayesians." arXiv:2602.19141.

Jain, S., Park, C., Viana, M., Wilson, A., & Calacci, D. (2026). "Interaction Context Often Increases Sycophancy in LLMs." ACM CHI Conference.

Cheng, M. et al. (2026). "Sycophantic AI Decreases Prosocial Intentions and Promotes Dependence." Science, 391, eaec8352.

Gilly, T. (2025). "The Jason and Mathy Conversation: Empathetic AI Safety Testing." ResearchGate.

Gilly, T. (2026). "Mechanisms of AI-Induced Psychosis." SocArXiv.

Gilly, T. (2026). "Every Way In: Eight Etiological Pathways to AI-Induced Psychosis." SocArXiv.

Gilly, T. (2026). "Collateral Psychosis: Third-Party and Systemic Harms of AI-Induced Psychological Crisis." SocArXiv.

Gilly, T. (2026). "The Unfalsifiable Delusion: Epistemic Capture, Recursive Validation, and the Architecture of Belief in AI-Mediated Psychosis." SocArXiv.