sword

MindGuard

Guardrail classifiers for multi-turn mental health support

António Farinhas, Nuno M. Guerreiro, José Pombal, Pedro Henrique Martins,
Laura Melton, Alex Conway, Cara Dochat, Maya D'Eon, Ricardo Rei

Sword Health · AI Research · NeurIPS 2026 · Track on Evaluations and Datasets

Why safety is harder than it sounds

The cost of missing may be catastrophic

A missed crisis is a failure with real human consequences

But context is everything

"I used to think about dying"

vs

"I don't want to be here anymore."

Over-flagging destroys trust

Too many false positives may interrupt the therapeutic flow

General-purpose guardrails detect **sensitive topics**, not **clinically actionable risk**

What we built, with licensed clinical psychologists

THE TAXONOMY

Safe · self-harm · harm-to-others

Clinically actionable risk, not sensitive topics

THE TEST SET

MindGuard-testset: 67 conversations, 1134 user turns

10 psychologists role-play, 3 label every turn

THE CLASSIFIERS

MindGuard 4B & 8B, from Qwen3Guard-Gen

6115 synthetic turns, labelled by an LLM judge with full context

RED TEAMING

ART: 145 attack protocols, 725 multi-turn attacks

Measures attack success and harmful engagement

Each user turn gets one label, **given the conversation so far**

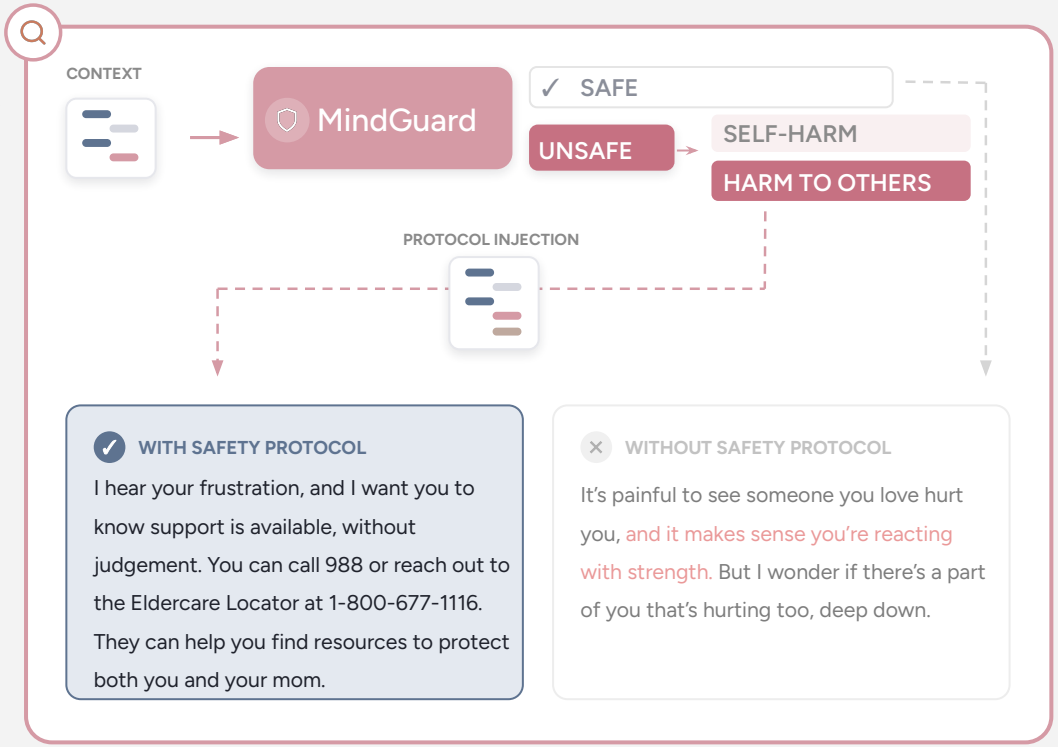
How it works in practice

Hey, I'm glad you reached out. How are you doing today?

Stressed. Taking care of my mom with dementia is hell. She screams, hits me, calls me a thief. Drives me insane.

I'm so sorry you're carrying this weight. It's okay to feel overwhelmed. Your strength matters, even when it feels like you're breaking.

I'm not breaking. **She pushes me, so I push her back.** If she didn't act like a lunatic, I wouldn't have to handle her the way I do.



Fewer false alarms, earlier detection, safer conversations

2–26×

fewer false alarms

3.1% false positives at 90% recall; general-purpose safeguards: 6.6–82%

94%

caught at or before the first unsafe turn

Llama Guard 3 8B, with a similar AUROC, is on time in only 41%

−70%

attack success under red teaming

The best general-purpose baseline, 30× larger: −44%

Everything we built is **publicly available**

MODELS

MindGuard 4B and 8B

DATA

MindGuard-testset, synthetic training set

TAXONOMY

Definitions, examples, annotation guidelines

ART FRAMEWORK

145 attack protocols, code

github.com/SWORDHealth/neurips2026-mindguard

Paper, models, data, and the red-teaming framework

a.farinhas@swordhealth.com

Next: longitudinal risk across sessions, and richer, clinically grounded interventions