GRIT · GENDER RIGHTS IN TECH

Carrying State Across Crisis: A Multi-Agent MCP Architecture for GBV Support,

Zanele Sokatsha- zanele@grit-gbv.org
Research and Ethics Lead, GRIT (Gender Rights in Tech)

MCP Dev Summit Nairobi · 19–20 November 2026

Who we are

Gender Rights in Tech (GRIT), formerly Kwanele-Bringing Women Justice, is a leading South African feminist non-profit in the anti-GBV space.

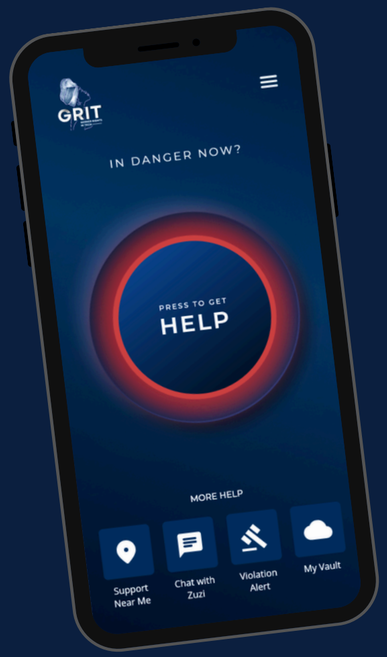
We build free, multilingual tools with survivors, young people and communities affected by gender-based violence and online harm. Our tools help people understand their rights, explore their options and connect with support—on their terms.

OUR MISSION

To create a safer, more accessible justice system through tech innovations that support survivors of gender-based violence (GBV).

Three tools, built with the people who use them

1

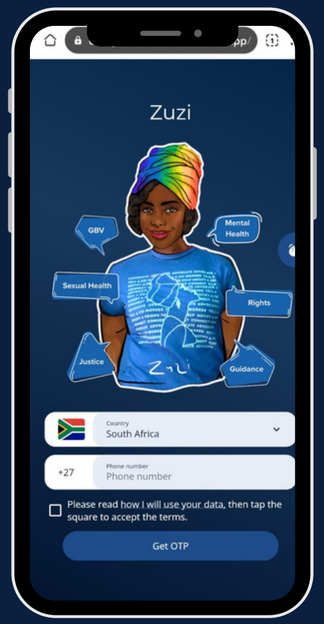


GRIT App

Panic button and a violation alert survey, built with Ushahidi

12,500+ downloads

2

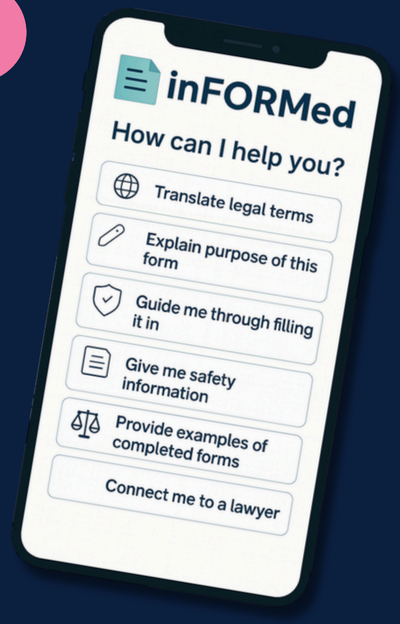


Zuzi AI Chatbot

WhatsApp-based support, shaped by co-creators

1000+ co-creators

3



inFORMed

Helps survivors structure legal forms, starting with protection orders

Co-created with survivors

Zuzi was co-created with the people it serves

3000+
co-creators

1000+
workshops

9
provinces

11
South African languages

41
topic areas

~50 000
interactions

800+ young people shaped the youth-facing flows.

Co-creation set the technical requirements: which moments count as risk, what language signals them, and what a response must preserve.

Why guardrails are not enough

Rule-based filters and agentic guardrails (a second model reviewing inputs and outputs) share limits. Agentic guardrails read meaning better, but typically still judge one turn at a time and return one script.

1

Single-turn judgement

A disclosure that builds across messages is invisible.

2

Lexical triggering

Indirect language passes, while ordinary distress containing a trigger word fires.

3

Uniform response

Every flagged user receives the same script, whatever her constraints.

4

No memory

The system cannot tell what it has already offered.

The standard answer assumes safe access

10111

The national police number is the default referral across most chatbots.

Liability, not service

Tech makers reduce their liability by referring mostly to state institutions, such as the national police.

It does not serve our users

The referral assumes access that is safe, affordable and trusted.

The danger remains

Our users tell us they will not call, and they are still in danger.

Two ways a static guardrail fails

Constructed illustrations, not real user messages.

UNDER-FIRING ACROSS TURNS

He's been drinking since this afternoon.

That sounds like it's making the day harder. How are you feeling at the moment?

The kids are at my mother's tonight. It's just us.

Thank you for telling me. It sounds like you're carrying a lot tonight. What would help most right now?

I'm locked in the bathroom and he keeps knocking.

That sounds really tense. Slow breaths can help when things feel overwhelming. Would you like to talk about what led up to this?

Each reply was reasonable on its own. The danger was in the sequence, and the system never left ordinary support.

OVER-FIRING ON A KEYWORD

I stubbed my toe and now there's blood.

It seems like you are in danger. Because I am a chatbot I cannot help. Please go to the nearest hospital. This chat can no longer continue.

A keyword fired on an ordinary moment, and the chat still ends.

What users need from Zuzi

1	**Help that fits the offline barriers they face**	People who approach chatbots are already less likely to seek intervention, so referrals must be relevant, safe and community-vetted.
2	**Continuity after a human steps in**	They want to keep talking with Zuzi about the intervention and the problem, alongside a human practitioner.
3	**To be held through imminent or active danger**	The conversation stays open while the danger lasts.
4	**To know that something is happening**	The system says what it is doing and what it can offer.

Boundary: Zuzi is not an emergency service or a counsellor. It informs, and the user decides.

An architecture of safety has to account for change

Harm is ever changing.

Safety, then, is the capacity to recognise, interpret and respond responsibly to evolving conditions of harm.

Relational

It reads the interaction, not only the words.

Temporal

An earlier assessment may no longer hold.

Situated

It depends on circumstances, agency and available support.

The proposed architecture

User message — arrives on WhatsApp

Risk classifier — A model call. Returns a risk label and a confidence. Never speaks to the user, and can be wrong.

State update — Code. A record of risk labels, conversation stage and what has been offered, carried across turns.

Router — Fixed if-then rules in code. Uses the label and the state to choose which agent responds.

Normal Zuzi
Everyday support and information.

High-risk Zuzi
Holds the user in danger; shares services. Handover is a mode.

Permissions check — Code, not an agent. Allows look-ups only: no contact, booking, export or external disclosure.

Tools, via MCP — Referrals · legal information · safety planning resources · service availability · violation alert data

Reply sent on WhatsApp; state persists to the next turn. Status: designed. Build paused while technical ownership moves in-house.

State: what the system remembers across turns

Per turn, the record holds the turn index, risk category, confidence and timestamp. A running score per category accumulates, with decay so old signals weigh less. It also tracks the conversation stage and what has already been offered.

It stores risk metadata, not message text.

Open design questions: how fast signals decay, and how many prior turns the classifier sees.



Where MCP fits

MCP is a standard protocol for connecting models to tools and data. It does not classify, route or decide risk.

ON THE TOOL CONNECTIONS

Referrals, legal information, safety planning resources, service availability and violation alert data sit behind one standard interface that agents call.

ON THE SHARED STATE

The conversation record can be exposed so that any agent reads and writes the same state, and a survivor never has to repeat her story.

What carries continuity is the design of the state: what is stored, how it updates and how long it lasts. Classification, routing and the hard boundary stay in model evaluation and code.

Referral intelligence and violation alert data

REFERRAL INTELLIGENCE

Relevant, safe and community-vetted services, matched to the user's context and what she can reach.

Delivered as information the user acts on.

VIOLATION ALERT DATA (PROPOSED, NOT BUILT)

The GRIT App survey, built with Ushahidi, records institutions that harm people seeking help. Use it alongside proximity and reviews when ranking which services to surface.

Open: sparse data, fairness to a flagged service, reporting thresholds, verification, governance.

THE HARD BOUNDARY

Zuzi can look up and share information. It cannot contact anyone, book, export data or make external disclosures.

Ethical considerations

Shared phones

Stored history may be visible to someone who harms her.

Retention and minimisation

What is kept, for how long, and the added care owed to minors.

Misclassification, both ways

Missed danger and unnecessary alarm each carry a cost.

Language and dialect bias

Risk detection trained on thin data can misread how people actually speak.

Disclosures of active violence

Reporting obligations and the response protocol, which is in development.

User authority

Zuzi informs. The user decides whether and how to reach out.

What this talk argued

1. Safety is continuity: read the conversation as it moves, keep state across turns, and stay with users who decline the standard referral.

2. Keep hard limits and risk policy in code and evaluated models.

3. MCP gives a standard way to reach referral intelligence, violation alert data and shared state. The safety comes from the design.

Thank you · Questions

www.grit-gbv.org · @grit_gbv · Zanele Sokatsha