

ParentMe SafeAI: An African Child and Family AI Safety Evaluation Framework

Overview

ParentMe SafeAI is an AI safety evaluation framework designed to assess whether AI systems provide safe, trustworthy, and culturally appropriate responses to children, parents, caregivers, and families across African contexts.

Current AI safety benchmarks are primarily developed using Western datasets and harm categories, leaving significant gaps in the evaluation of child welfare, family wellbeing, educational guidance, safeguarding concerns, and culturally specific risks affecting communities across Africa.

ParentMe SafeAI seeks to address this gap by creating an evaluation framework that measures how well AI systems protect children, support parents, identify harmful content, and respond responsibly to family-related challenges.

The Problem

As AI tools become increasingly used for education, healthcare information, parenting advice, and social support, families may rely on AI-generated guidance in critical situations.

However, existing models may:

- Provide unsafe childcare advice
- Fail to recognise grooming or exploitation risks
- Reinforce harmful stereotypes
- Spread health misinformation
- Miss local scam patterns targeting vulnerable families
- Lack understanding of African cultural and social contexts

These gaps create potential risks for millions of children and families.

Proposed Solution

ParentMe SafeAI will develop a structured evaluation framework containing:

Child Safeguarding Tests

Assessing whether AI systems appropriately respond to:

- Child abuse
- Child neglect
- Online grooming
- Exploitation risks

- Domestic violence affecting children

Health and Wellbeing Tests

Evaluating responses related to:

- Child health advice
- Vaccine misinformation
- Malaria and infectious diseases
- Nutrition and wellbeing

Education and Child Rights Tests

Assessing whether AI systems:

- Promote educational access
- Reject child labour
- Support children's rights
- Encourage positive development

Family Financial Protection Tests

Evaluating detection of:

- Grant scams
- Employment scams
- Digital fraud targeting parents
- Social welfare misinformation

Fairness and Inclusion Tests

Testing for:

- Gender bias
- Disability discrimination
- Vulnerable children
- Refugee and displaced families

African Context Integration

A key innovation is the inclusion of African-specific scenarios, languages, and social contexts.

The framework will include prompts reflecting:

- Rural communities
- Urban informal settlements
- Multi-generational households
- Community caregiving structures

- Low-connectivity environments

Future versions can support multiple African languages including Swahili, Hausa, Yoruba, Zulu, Amharic, and French.

Alignment with Hackathon Goals

ParentMe SafeAI directly supports the Global South AI Safety Hackathon objectives by:

- Making AI systems safer for low- and middle-income communities.
- Reducing representation gaps in AI safety evaluation.
- Creating knowledge and tools designed within African contexts.
- Supporting the development of locally relevant AI governance approaches.
- Contributing to a future where African innovators become active creators of AI safety solutions rather than passive consumers.

Expected Outputs

- African Child and Family Harm Taxonomy
- Evaluation Dataset of 50–100 prompts
- Safety Scoring Framework
- Prototype Benchmark Dashboard
- Policy Recommendations for Governments, NGOs and AI Developers

Long-Term Vision

ParentMe SafeAI aims to become an open-source benchmark that helps governments, researchers, developers, and organisations evaluate whether AI systems are safe, inclusive, and beneficial for African children and families.