GemmaGuard 3n

Technical Writeup

Google - The Gemma 3n Impact Challenge

1. Introduction & Motivation

Children today are increasingly exposed to emotional and behavioral harm from inappropriate content. This includes online exposure to violent or sexual material, offensive language, and toxic behavior, often found in movies, social media, and other digital platforms, as well as offline exposure through social environments, where family members, peers, or caregivers may model or directly engage in harmful behaviors. Children absorb and mirror behaviors they constantly see or hear, especially in formative years. Repeated exposure to harmful stimuli can negatively affect children's emotional and psychological development, potentially leading to lasting behavioral and issues or trauma.

The problem is exacerbated by the diversity of children's backgrounds: age, neurodiversity, culture, trauma history, and personal sensitivities all shape what constitutes "harmful or inappropriate" for the child, highlighting the need for nuanced and context-aware protection strategies. Existing current abuse detection systems and content moderation tools rely on pre-defined set of harm categories. They often fail to capture nuanced and context-dependent forms of harm. Furthermore, few systems prioritize real-time analysis, privacy, and customizability, all of which are essential for protecting children in both home and institutional environments.

Our proposed system, GemmaGuard3N, reimagines child protection through an intelligent, adaptable, privacy-conscious lens. Built using Gemma 3n, a state-of-the-art open multimodal LLM designed for running with high efficiency on low-resource devices. Our solution is capable of processing long-form visual and multilingual audio streams from diverse sources. GemmaGuard3N leverages zero-shot reasoning and customizable prompting to enable customizable real-time detection of potential harm tailored to individual children's needs.

2. Problem Statement

We define the problem as the detection of harm indicators in live or recorded video streams involving children, where "harm" includes:

- Inappropriate language, behavior, or gestures
- Psychological abuse or aggression
- Unsafe content exposure (e.g., sexual innuendos, substance use)
- Emotionally disturbing interactions or tones

Key Technical Requirements

- **Multimodal Understanding:** Detect harm from both visual and auditory inputs.
- **Privacy-Aware Processing:** Ensure that content is analyzed locally or through secure channels to protect user privacy.
- **Real-Time or Near-Real-Time Performance**: Provide immediate alerts to caregivers upon detecting signs of harm.
- **Personalization:** Support customizable definitions of harm based on each child's individual needs.
- **Zero-Shot Capability:** Eliminate the need for training a custom dataset for every use case.
- Multilingual Support: Accurately process audio in multiple languages to accommodate diverse populations.
- **Long-Form Video Handling:** Continuously analyze long-duration videos such as CCTV footage, movies, and online content without interruption.

3. System Architecture

GemmaGuard3N comprises four core modules:

1. **Input Handling**

The system supports a variety of video sources:

- Real-Time CCTV Streams via camera IP URLs.
- Local Videos, including uploaded movie clips or home surveillance footage.
- YouTube Links, allowing analysis of online content before sharing it with children.

2. **Preprocessing and Feature Extraction**
   - Video Frames: Sampled at 1 frame per second (FPS).
   - Audio: Extracted and converted to 16 kHz WAV format.
   - Temporal Alignment: Maintained to synchronize video and audio cues accurately.

3. Harm Detection with GemmaGuard3N

A lightweight, multimodal method leveraging Gemma 3N:

- Dynamically crafts prompts based on user-selected harm indicators.
- Applies a sliding window approach to process long videos in overlapping segments.
- Detects nuanced patterns using token-level SoftMax probability scoring from MLLM logits.

4. **Feedback and Alert System**
   - If harm is detected, parents or guardians are notified.
   - Each flagged instance includes harm probability score and a timetable to indicate segments of concern.
   - Recorded evidence, such as audio/video snippets, is provided to support parent review and informed decision-making.

4. Methodology

We designed the GemmaGuard3N method to detect harm indicators in real time from multimodal video inputs. The method consists of the following steps:

- **Step 1:** We curated a general list of harm indicators (e.g., violence, sexual content, verbal abuse, and inappropriate language). Optional harm indicators are also collected from users to extend this list based on their specific child safety requirements.
- **Step 2:** We craft a negatively framed prompt for the Gemma 3N using the selected harm indicators. For example: “Does this video contain any of the following: violence, sexual content, inappropriate language, or verbal abuse?” A "Yes" response indicates the presence of harm. Prompts are carefully constrained using system- and user-level instructions to ensure response (“Yes” or “No”) only.
- **Step 3:** To enable Gemma 3N to process longer videos, we implement a sliding window approach. Videos are divided into short, overlapping chunks to reduce inference load, based on the output of a preprocessing and feature extraction module. The window size is dynamically adjusted depending on device memory and model capacity. Overlaps are introduced between windows to prevent missing brief harmful segments (e.g., a single swear word). Each chunk is sent to Gemma 3N for evaluation.
- **Step 4:** For each window, we extract the logits associated with the “Yes” and “No” tokens generated in response to the crafted prompt. A SoftMax function is applied to these logits to compute the probability of a “Yes” response as a confidence score. A window is flagged as harmful if the "Yes" probability exceeds a defined threshold. This design is inspired by the QGuard paper [1], which used token-level SoftMax probabilities to detect unethical or harmful text prompts in the context of malicious attacks.

5. Implementation Details

We build GemmaGuard3N using a modular pipeline composed of the following core components:

- For video processing, we employ OpenCV for efficient frame sampling and real-time stream handling. Frames are extracted at configurable intervals to balance computational efficiency and temporal coverage.
- For audio processing, we use Pydub and Librosa to segment, normalize, and convert raw audio streams into model-compatible formats.
- For model inference, the backbone of our system is Gemma 3N, a multilingual, multimodal vision-language model developed by Google. We leverage the HuggingFace Transformers library with PyTorch to run inference over synchronized audio-visual segments.
- For the user interface, we implement an interactive web demo using Gradio, which supports real-time user input, file uploads, and video visualization.

6. Conclusion

GemmaGuard3N, powered by Gemma 3n, is a powerful proof-of-concept for socially responsible, AI-driven child protection. It transforms the way we approach video analysis, offering not just detection but contextual, human-aware safety for those most vulnerable. We believe our technical architecture and methodology can inspire a new class of safety applications built on Gemma3n, ones that are as empathetic as they are intelligent.

7. Citations

[1] T. Lee, J. Yoo, H. Cho, S. Y. Kim, and Y. Maeng, “QGuard: Question-based Zero-shot Guard for Multi-modal LLM Safety,” arXiv preprint arXiv:2506.12299 v1, June 2025