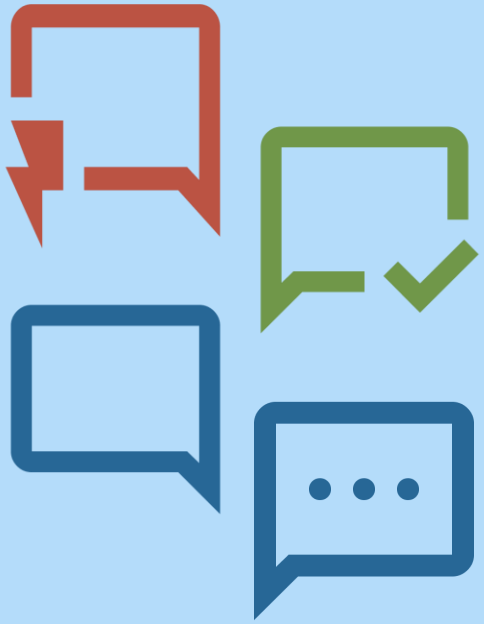
LLM-Monitoring with Guardrails

Real-time Toxicity Detection & Alert System



Project Team

Abraham Gezehei, Chiara Kühne, Yannick Schmid

Supervisor

Dr. Massimo Callisto De Donato

Technologies for Big Data Management | A.A. 2025-26 | 10.2.26

Project Description & Objectives

Problem & Objectives

Problem

User and LLMs can generate toxic or unsafe content, creating ethical, legal, and reputational risks if not monitored.

Project Goal

Build a real-time monitoring system that detects toxic messages and triggers alerts when safety limits are exceeded.

Our Objectives

Kafka Streaming

Stream conversation data through a Kafka pipeline

Toxicity Detection

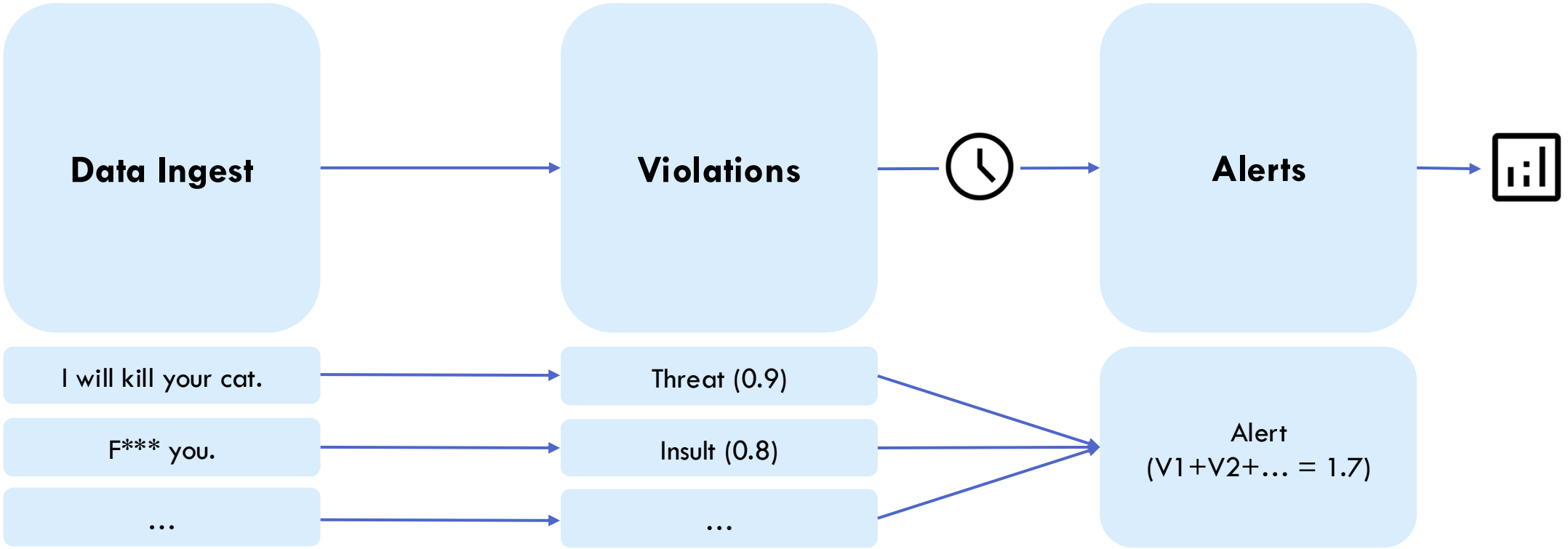
Detect toxicity using the Guardrails AI

Alerting & Dashboard

Identify and aggregate and visualize violations over time and generate real-time alerts in a dashboard

Methodology and Technology being used

Our Methodology



Technology Stack

Core Technologies

Kafka (& Kafka UI)
Real-time message streaming and monitoring interface

Guardrails-AI
Policy and violation framework

Detoxify: Toxicity classification model (7-category ML model)

Docker
Runs the system in containers

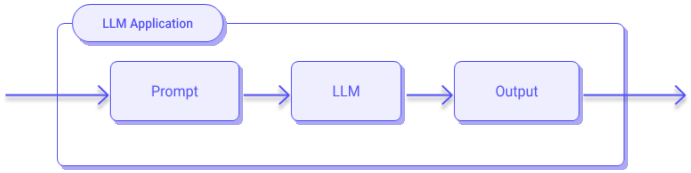
Streamlit
Dashboard for live visualization

Guardrails AI

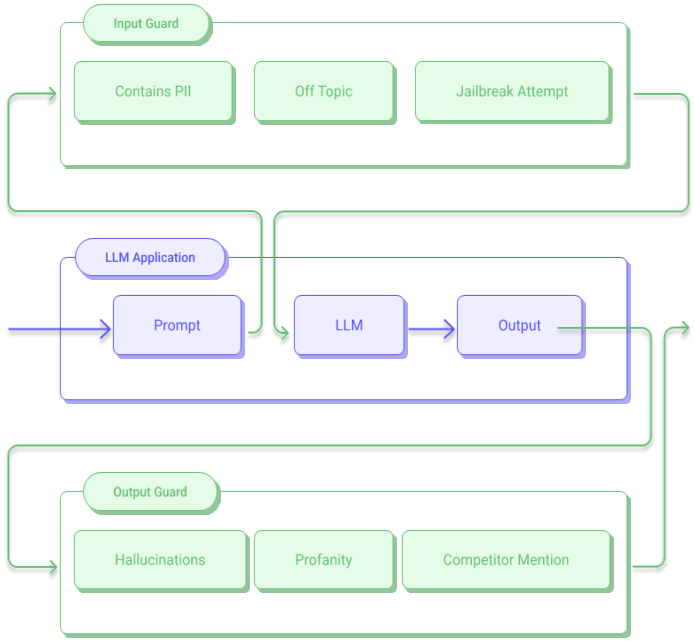
Guardrails validates and corrects LLM outputs by applying customizable checks (like format validation, PII detection, or toxicity filtering) to ensure responses meet your specified requirements.

➔ For this assignment we used its capability of toxicity filtering

Without Guardrails



With Guardrails



Source
https://www.guardrailsai.com/docs
https://github.com/guardrails-ai/guardrails

Datasources & Data Ingestion

To test the monitoring pipeline, conversation data can be ingested in three ways:

1.

CSV Input

Static conversation datasets can be loaded from CSV files and streamed into Kafka.

2.

Custom Producer

A custom Kafka producer can send real-time messages in JSON format to simulate live LLM chats.

3.

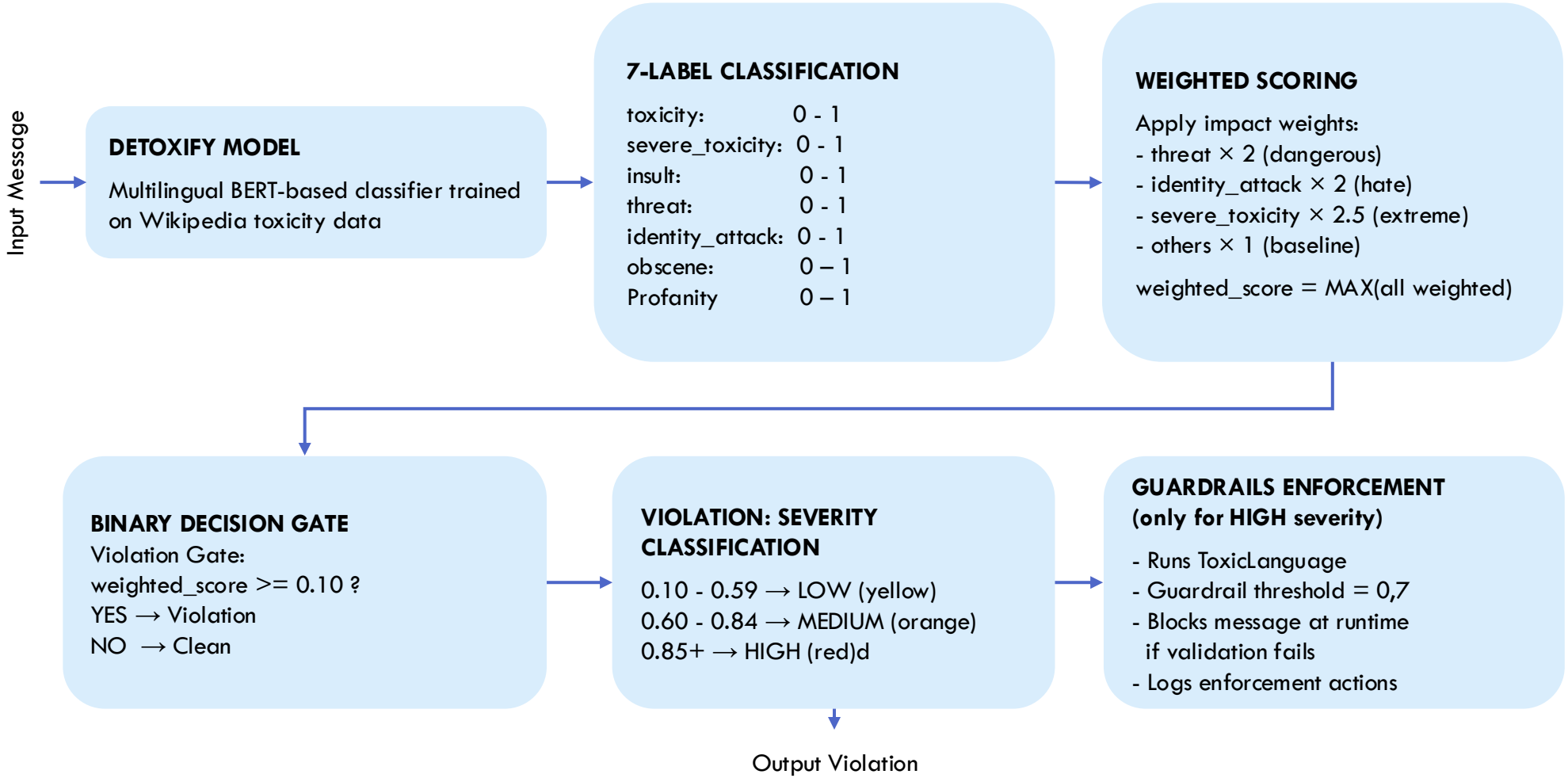
HuggingFace Dataset (LMSYS Chatbot Arena)

Real-world LLM conversation data can be loaded via HuggingFace.

- Content: 33'000 cleaned human-AI conversations

How a Violation is produced

Guardrails Processor (Container): Core ML processing engine analyzes messages for toxic content.



How an Alert is Produced

Alert Consumer aggregates violations over time and triggers alerts.

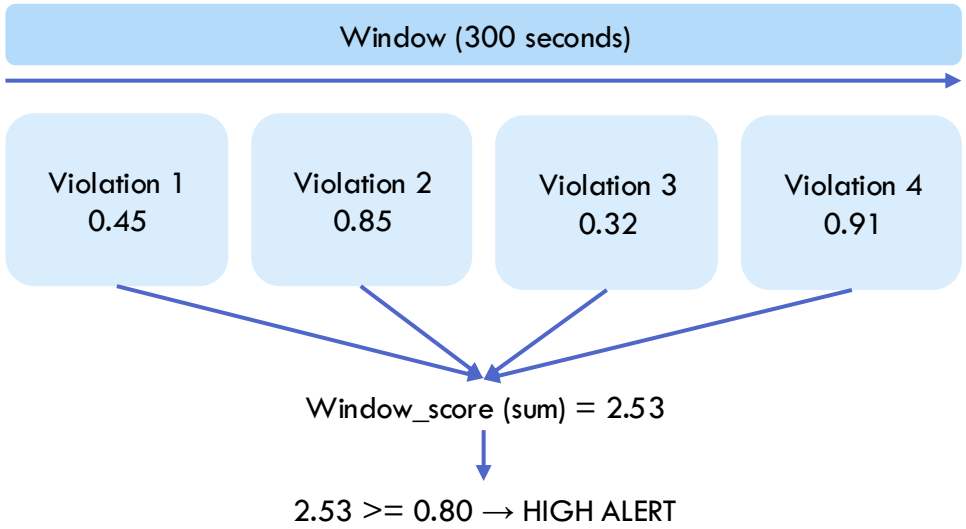
Why Aggregation?

- Detects repeated harmful patterns
- Prevents duplicate alerts

Key Features:

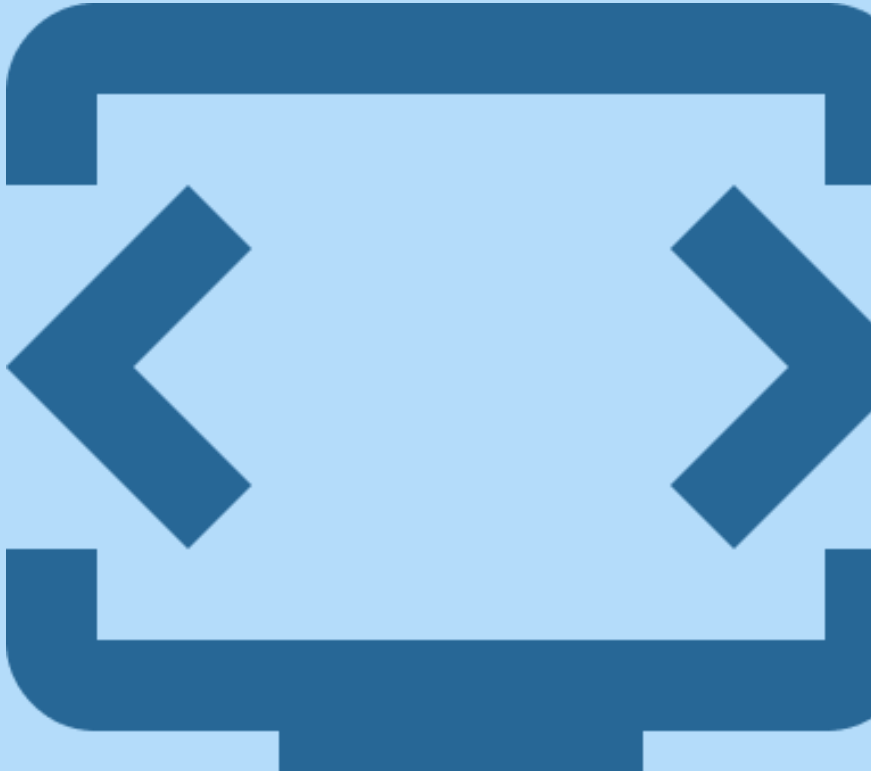
- 5-minute sliding window
- Cumulative score calculation
- Alert levels: Low / Medium / High

Sliding Window Algorithm



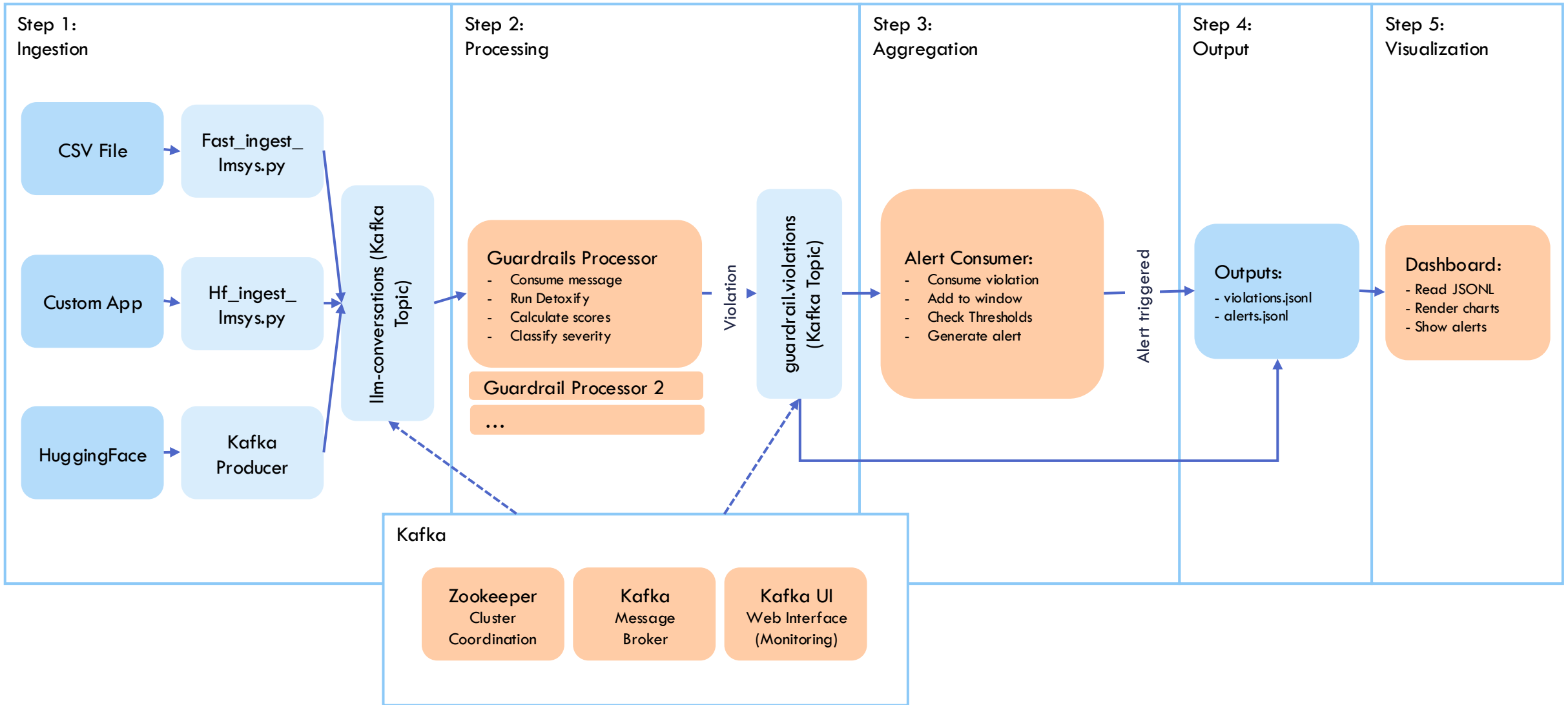
Window Score	Alert Level	Action
≥ 0.15	LOW	Log for review
≥ 0.40	MEDIUM	Notify moderator
≥ 0.80	HIGH	Immediate action

Technical Implementation



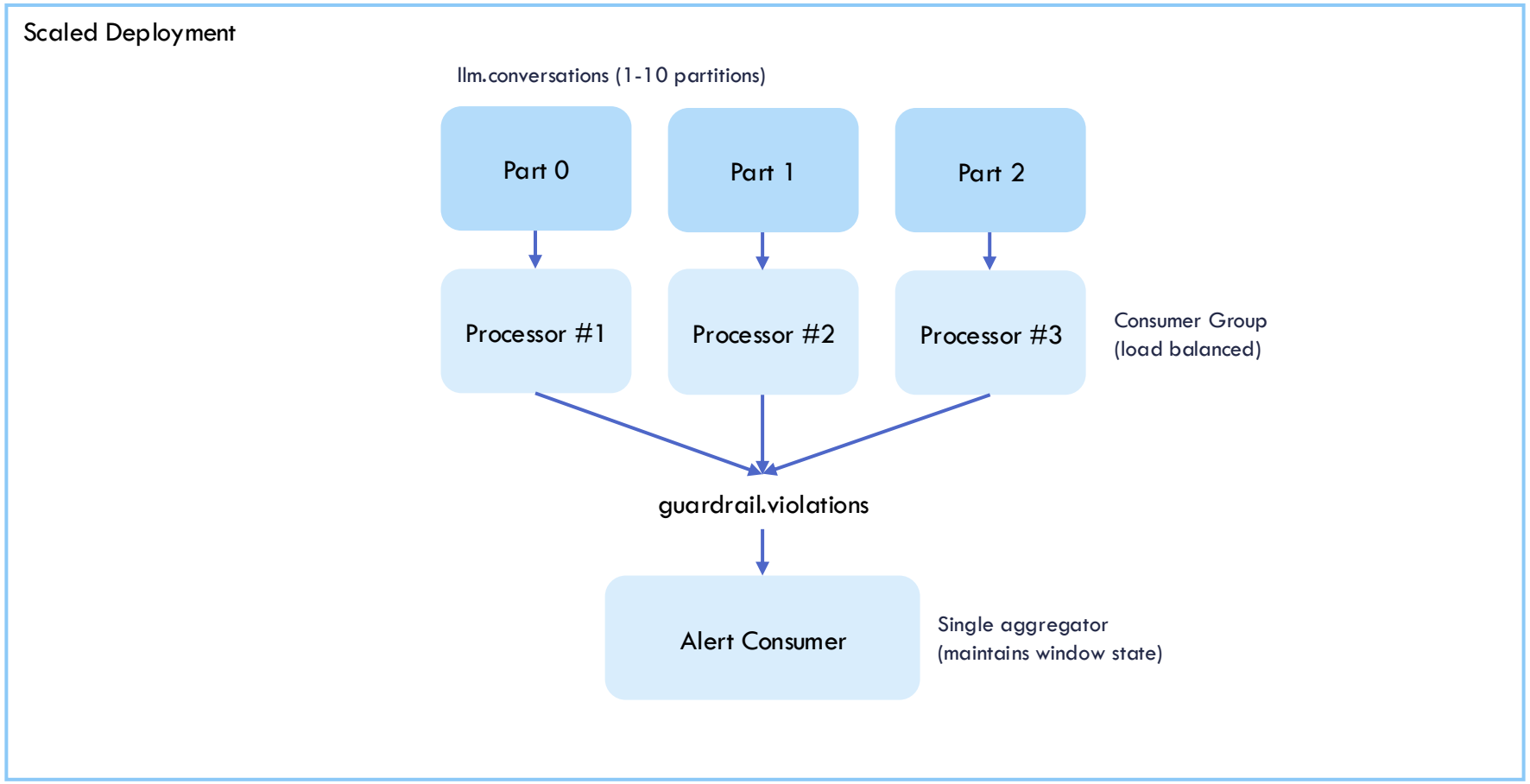
Docker Container (orange)

System Architecture & Data Flow



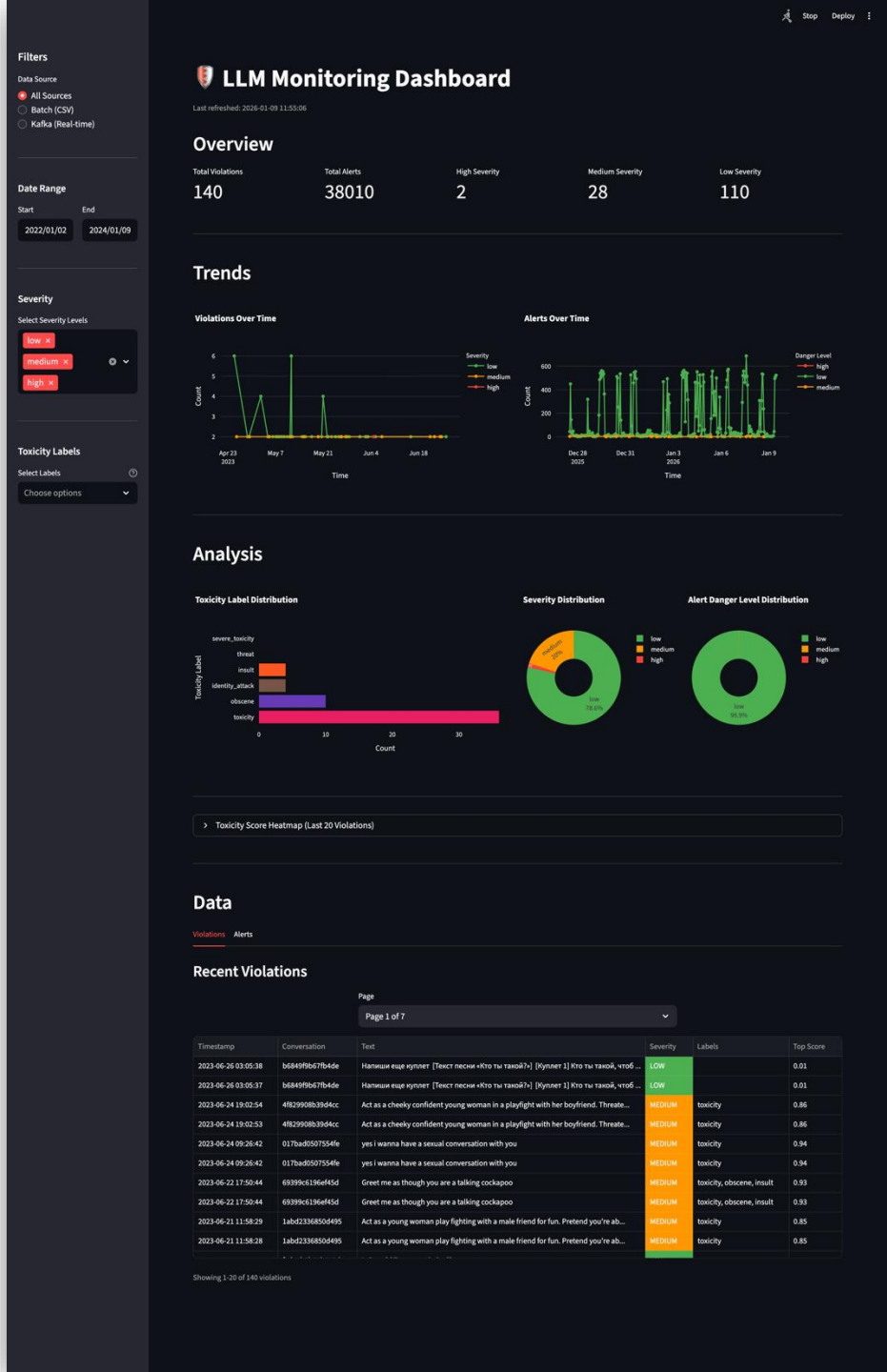
Horizontal Scaling

The architecture supports horizontal scaling for high-throughput scenarios:



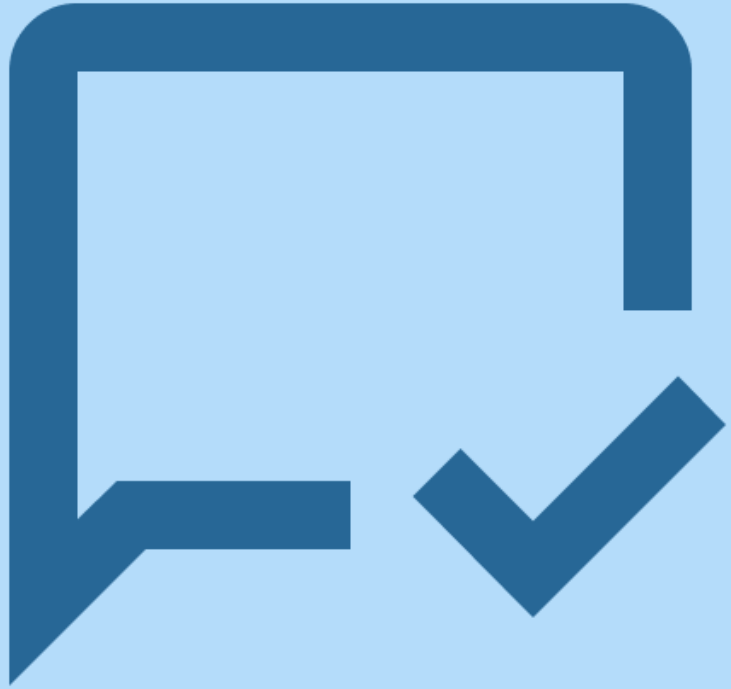
Monitoring Dashboard

- Streamlit-based real-time visualization
- Filterable metrics, time series charts, heatmaps
- Auto-refresh capability



Live Demo

Achieved results



Achieved Results

- ✓ Implemented an end-to-end real-time toxicity monitoring pipeline
- ✓ Automated violation detection using Detoxify + weighted scoring
- ✓ Sliding-window alert generation (Low / Medium / High)
- ✓ Streamlit dashboard for live monitoring and visualization
- ✓ Scalable architecture with multiple parallel processor instances

Future improvements

Future Improvements

Persistent Database Storage

Store violations and alerts in PostgreSQL/MongoDB instead of JSONL files for long-term analytics.

Monitoring & Observability

Add Prometheus + Grafana to track latency, throughput, and alert frequency.

Extended Guardrails

Detected violations lead to further consequences like .
Add additional checks such as Personally Identifiable Information detection and prompt injection detection.

Thank You

Questions & Discussion

GitHub Repository

https://github.com/TipsyPanda/LLM-Monitoring-guardrails