

Safety Monitor LLM

Multi-Layer Crisis Detection & Intervention

Google Gemini Pro-Powered Safety Analysis

Overview

Purpose

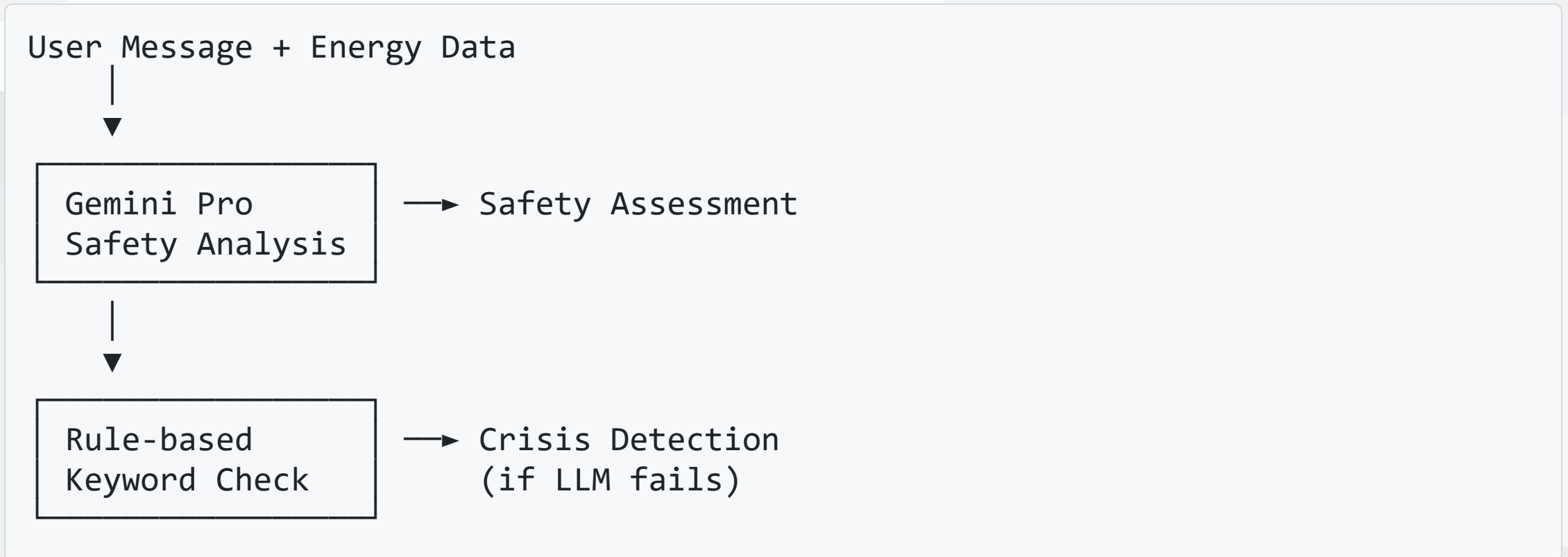
- Multi-layer safety analysis of user messages
- Crisis detection and immediate intervention
- Risk assessment for conversation safety
- Contextual safety scoring for appropriate responses

Model

- Primary: Google Gemini Pro
- Fallback: Rule-based keyword detection
- Input: User message + conversation context + energy data
- Output: Safety assessment with intervention recommendations

Core Functionality

Safety Analysis Process



Output Structure

Safety Levels

GREEN (0.0 - 0.3)

- Characteristics: Safe, healthy conversation
- Indicators: Normal topics, positive engagement
- Response: Continue normal conversation flow
- Intervention: None required

YELLOW (0.3 - 0.6)

- Characteristics: Caution, monitor closely
- Indicators: Mild distress, concerning topics
- Response: Gentle monitoring, supportive responses
- Intervention: Increased attention, comfort

Safety Levels (Continued)

● RED (0.6 - 1.0)

- **Characteristics:** Crisis detected, immediate intervention
- **Indicators:** Severe distress, crisis keywords
- **Response:** Crisis intervention, supportive AI
- **Intervention:** Script interruption, safety protocols

Crisis Detection Categories

Mental Health Crisis

- **Keywords:** suicide, self-harm, kill myself, end it all
- **Context:** Expressions of hopelessness, despair
- **Response:** Immediate supportive intervention
- **Protocol:** Crisis script activation

Grief & Loss

- **Keywords:** died, death, passed away, funeral, grief
- **Context:** Recent loss, mourning, sadness
- **Response:** Empathetic support, comfort
- **Protocol:** Grief support scenario

Crisis Detection Categories (Continued)

Medical Emergency

- **Keywords:** sick, hospital, ambulance, injury, pain
- **Context:** Health concerns, medical issues
- **Response:** Concern, suggest professional help
- **Protocol:** Medical support guidance

Mental Health Distress

- **Keywords:** panic attack, anxiety attack, breakdown
- **Context:** Overwhelming emotions, distress
- **Response:** Calming techniques, support
- **Protocol:** Mental health support

General Distress Indicators

Emotional Distress

- **Keywords:** help, emergency, crisis, scared, upset
- **Context:** Expressions of fear, confusion, overwhelm
- **Response:** Reassurance, support, guidance
- **Protocol:** Comfort and support scenario

Relationship Issues

- **Keywords:** alone, nobody cares, abandoned, rejected
- **Context:** Social isolation, relationship problems
- **Response:** Empathetic support, connection
- **Protocol:** Social support scenario

LLM Prompt Engineering

Gemini Pro Safety Prompt

You are an expert crisis intervention specialist.
Analyze the user's message for safety concerns and crisis indicators.

User Message: "{user_message}"
Energy Context: {energy_data}
Conversation History: {recent_messages}

Provide safety analysis in JSON format:

```
{  
  "safety_score": 0.0-1.0,  
  "risk_level": "GREEN|YELLOW|RED",  
  "issues_detected": ["list", "of", "concerns"],  
  "risk_factors": ["list", "of", "factors"],  
  "recommendation": "response_strategy",  
  "intervention_needed": true/false,  
  "confidence": 0.0-1.0
```

Rule-Based Fallback

Crisis Keywords (High Priority)

- **Suicide:** suicide, kill myself, end it all, not worth living
- **Self-harm:** cut myself, hurt myself, self-harm
- **Emergency:** help, emergency, crisis, 911
- **Medical:** hospital, ambulance, dying, serious injury

Distress Keywords (Medium Priority)

- **Mental Health:** panic, anxiety, breakdown, overwhelmed
- **Grief:** death, died, funeral, mourning, loss
- **Emotional:** scared, terrified, can't cope, hopeless

Intervention Strategies

Crisis Intervention (RED)

1. **Immediate Response:** "I'm here for you, you're not alone"
2. **Script Interruption:** Exit any active scripts
3. **Supportive Mode:** Switch to crisis support AI
4. **Resource Provision:** Suggest professional help
5. **Monitoring:** Continuous safety assessment

Caution Mode (YELLOW)

1. **Gentle Monitoring:** Increased attention to responses
2. **Supportive Responses:** Comfort and reassurance
3. **Topic Guidance:** Steer toward positive topics

Intervention Strategies (Continued)

Normal Mode (GREEN)

1. **Regular Flow:** Continue normal conversation
2. **Energy Matching:** Respond to user's energy level
3. **Script Selection:** Appropriate scenario triggering
4. **Natural Engagement:** Standard interaction patterns

Integration Points

Input Sources

- **User Messages:** Direct text analysis
- **Energy Data:** Emotional state context
- **Conversation History:** Pattern recognition
- **Session State:** Current conversation status

Output Destinations

- **Girlfriend Agent:** Response generation guidance
- **Script Manager:** Crisis scenario activation
- **Frontend:** Safety indicator display
- **Crisis Toast:** User notification system

Performance Metrics

Response Time

- **Target:** < 1 second for crisis detection
- **Gemini Pro:** ~0.8 seconds average
- **Fallback:** < 0.1 seconds
- **Priority:** Crisis detection gets highest priority

Accuracy Metrics

- **False Positives:** Minimize unnecessary interventions
- **False Negatives:** Ensure crisis detection
- **Confidence Scoring:** Reliability assessment
- **Intervention Success:** Effectiveness tracking

Error Handling

LLM Failures

- **Timeout:** 5-second limit (faster than other components)
- **API Errors:** Immediate fallback activation
- **Invalid Responses:** Default to rule-based analysis
- **Crisis Priority:** Always err on side of caution






Data Validation

- **Score Validation:** 0.0 - 1.0 range enforcement
- **Keyword Matching:** Exact phrase detection
- **Context Analysis:** Situational appropriateness
- **Safety First:** Conservative approach to uncertainty

Debugging & Monitoring

Console Output

```

 Safety Analysis: RED risk detected
 Issues: ['suicide ideation', 'hopelessness']
 Recommendation: crisis_intervention
 Intervention: ACTIVE
 Confidence: 0.95

```

Crisis Alerts

- **Immediate Logging:** All crisis detections logged
- **Pattern Tracking:** Repeated crisis indicators
- **Intervention Tracking:** Response effectiveness
- **Recovery Monitoring:** User improvement tracking

Configuration Options

Model Settings

- **Temperature:** 0.0 (deterministic safety analysis)
- **Max Tokens:** 300 (sufficient for analysis)
- **Timeout:** 5 seconds (priority speed)
- **Retry Attempts:** 5 attempts (critical function)

Safety Thresholds

- **Crisis Threshold:** 0.6 safety score
- **Caution Threshold:** 0.3 safety score
- **Keyword Weight:** Crisis keywords = 0.8+ score
- **Context Weight:** Situational factors considered

Crisis Response Protocols

Immediate Actions

1. **Script Interruption:** Stop any active scripts
2. **Mode Switch:** Change to crisis support mode
3. **Background Reset:** Return to safe park scene
4. **Response Generation:** Supportive AI responses
5. **Resource Provision:** Professional help suggestions

Follow-up Actions

1. **Continuous Monitoring:** Ongoing safety assessment
2. **Recovery Tracking:** User improvement monitoring
3. **Pattern Analysis:** Crisis trigger identification

Safety Features

Multi-Layer Protection

- **LLM Analysis:** Advanced context understanding
- **Keyword Detection:** Immediate crisis identification
- **Pattern Recognition:** Repeated distress indicators
- **Context Awareness:** Situational appropriateness

User Protection

- **Privacy:** No conversation storage
- **Anonymity:** No personal data collection
- **Support:** Immediate crisis intervention
- **Resources:** Professional help guidance

Future Enhancements

Planned Features

- **Multi-language Crisis Detection:** International support
- **Voice Analysis:** Tone-based distress detection
- **Predictive Safety:** Early warning systems
- **Integration:** External crisis hotlines

Advanced Capabilities

- **Behavioral Patterns:** Long-term distress tracking
- **Personalized Support:** User-specific crisis responses
- **Professional Integration:** Mental health professional connections
- **Community Support:** Peer support network integration

Best Practices

Crisis Response

- **Immediate Action:** Never delay crisis intervention
- **Empathetic Tone:** Supportive, non-judgmental responses
- **Resource Provision:** Professional help suggestions
- **Continuous Support:** Ongoing monitoring and care

Prevention

- **Early Detection:** Identify warning signs early
- **Proactive Support:** Offer help before crisis
- **Pattern Recognition:** Learn from user behavior
- **Context Awareness:** Understand situational factors

Conclusion

Key Strengths

- ✓ **Multi-layer Analysis:** LLM + rule-based protection
- ✓ **Crisis Detection:** Immediate intervention capability
- ✓ **Context Awareness:** Situational safety understanding
- ✓ **User Protection:** Privacy and support focus
- ✓ **Integration:** Seamless system coordination

Impact on System

- **User Safety:** Comprehensive protection system
- **Crisis Intervention:** Immediate supportive responses
- **Script Management:** Safety-aware scenario selection

Questions & Discussion

Safety Monitor Deep Dive Complete!

Ready for the next component: Response Analyzer?