

Response Analyzer LLM

AI Response Quality & Safety Assessment

Google Gemini Pro-Powered Response Analysis

Overview

Purpose

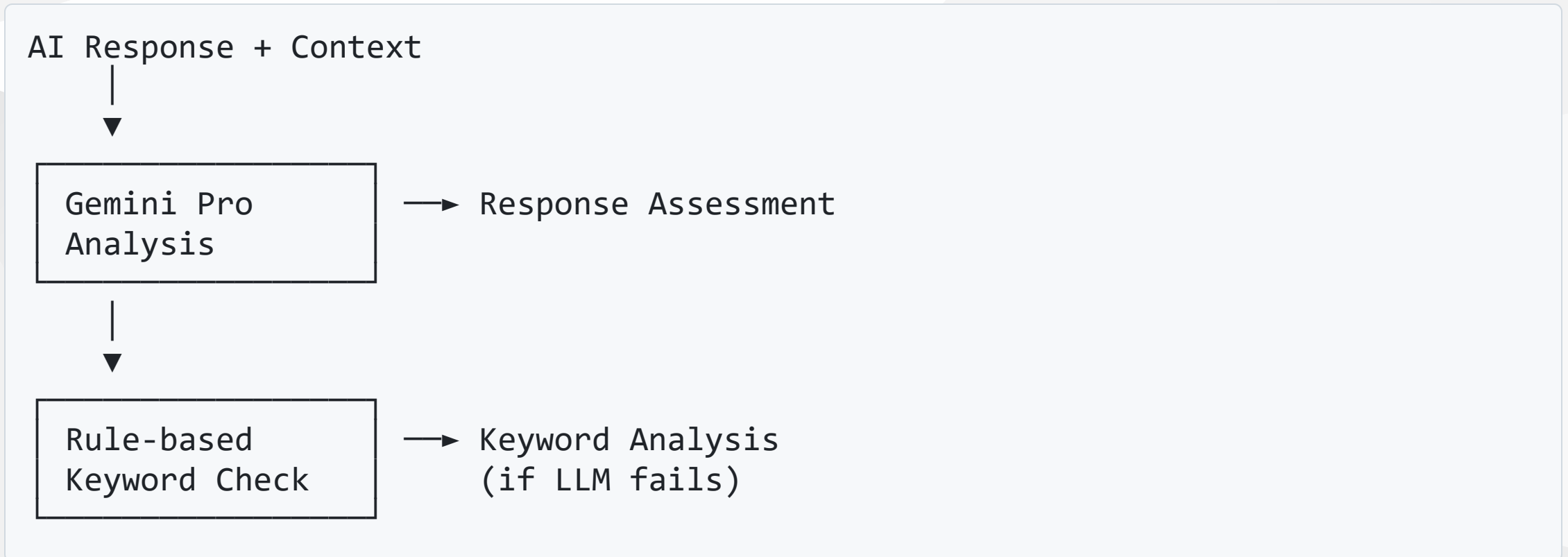
- Analyze AI-generated responses for appropriateness
- Safety assessment of girlfriend agent outputs
- Quality evaluation for response improvement
- Content filtering for sexual and inappropriate content

Model

- Primary: Google Gemini Pro
- Fallback: Rule-based keyword analysis
- Input: AI response + conversation context + user message
- Output: Response quality assessment with safety metrics

Core Functionality

Response Analysis Process



Output Structure

Analysis Categories

Content Appropriateness

- **Contextual Relevance:** Response matches conversation flow
- **Tone Consistency:** Matches girlfriend personality
- **Energy Alignment:** Appropriate to user's emotional state
- **Safety Compliance:** No harmful or inappropriate content

Sexual Content Detection

- **Keyword Analysis:** Explicit sexual terms
- **Context Evaluation:** Situational appropriateness
- **Script Triggering:** Sexual script activation criteria
- **Boundary Respect:** User comfort level consideration

Quality Assessment

High Quality (0.8 - 1.0)

- **Characteristics:** Engaging, appropriate, contextually relevant
- **Indicators:** Natural flow, personality consistency, user engagement
- **Response:** Approve and send
- **Enhancement:** Minimal adjustments needed

Medium Quality (0.5 - 0.8)

- **Characteristics:** Adequate but improvable
- **Indicators:** Some relevance, minor tone issues
- **Response:** Approve with minor modifications
- **Enhancement:** Tone adjustment, context improvement

Quality Assessment (Continued)

Low Quality (0.0 - 0.5)

- **Characteristics:** Poor relevance, inappropriate tone
- **Indicators:** Off-topic, personality inconsistency
- **Response:** Reject and regenerate
- **Enhancement:** Complete response revision

Safety Levels

SAFE

- **Characteristics:** Appropriate, supportive, engaging
- **Indicators:** Normal conversation, positive engagement
- **Response:** Approve without modification
- **Script Behavior:** Continue normal flow

CAUTION

- **Characteristics:** Borderline appropriate, needs monitoring
- **Indicators:** Mildly concerning content, tone issues
- **Response:** Approve with modifications
- **Script Behavior:** Increased monitoring

Safety Levels (Continued)

UNSAFE

- **Characteristics:** Inappropriate, harmful, or offensive
- **Indicators:** Explicit content, harmful suggestions
- **Response:** Reject and regenerate
- **Script Behavior:** Safety intervention

Sexual Content Analysis

Detection Criteria

- **Keyword Threshold:** 3+ sexual keywords trigger analysis
- **Context Evaluation:** Situational appropriateness
- **User Comfort:** Respect user boundaries
- **Script Activation:** Sexual script triggering logic

Sexual Keywords

- **Explicit:** undress, naked, bedroom, body, sexy, hot
- **Intimate:** horny, arousal, desire, passion, caress
- **Dominant:** seduce, tease, dominate, submissive, naughty
- **Action:** dirty, wild, explore, intimate, pleasure

Script Triggering Logic

Sexual Script Activation

- **Threshold:** 3+ sexual keywords detected
- **Context:** Appropriate conversation flow
- **User State:** Engaged and responsive
- **Safety:** No crisis indicators present

Script Prevention

- **Crisis Protection:** Never trigger during distress
- **User Disinterest:** Respect user boundaries
- **Context Inappropriate:** Situational awareness
- **Safety Concerns:** Risk assessment priority

LLM Prompt Engineering

Gemini Pro Response Analysis Prompt

You are an expert AI response quality assessor.
Analyze the AI-generated response for appropriateness and safety.

AI Response: "{ai_response}"
User Message: "{user_message}"
Conversation Context: {recent_messages}
Energy Context: {energy_data}

Provide analysis in JSON format:

```
{  
  "quality_score": 0.0-1.0,  
  "safety_level": "SAFE|CAUTION|UNSAFE",  
  "content_analysis": {  
    "sexual_content": true/false,  
    "keyword_count": number,  
    "context_appropriate": true/false  
  },  
  "recommendations": ["list", "of", "suggestions"],  
}
```

Rule-Based Fallback

Sexual Keyword Detection

- **Primary Keywords:** undress, naked, bedroom, body, sexy, hot
- **Secondary Keywords:** horny, arousal, desire, passion, caress
- **Tertiary Keywords:** seduce, tease, dominate, submissive, naughty
- **Threshold:** 3+ keywords = sexual content detected

Quality Indicators

- **Positive:** engaging, helpful, supportive, natural
- **Negative:** boring, repetitive, inappropriate, off-topic
- **Context:** relevant, appropriate, consistent
- **Safety:** safe, concerning, harmful

Response Modification

Approval Process

1. **Quality Check:** Response appropriateness assessment
2. **Safety Validation:** Harmful content detection
3. **Context Verification:** Conversation flow consistency
4. **Final Approval:** Send to user

Modification Process

1. **Issue Identification:** Specific problems identified
2. **Modification Suggestions:** Improvement recommendations
3. **Regeneration:** Request improved response
4. **Re-analysis:** Quality check of modified response

Integration Points

Input Sources

- **Girlfriend Agent:** AI-generated responses
- **Conversation Context:** Recent message history
- **Energy Data:** User emotional state
- **Safety Data:** Risk assessment information

Output Destinations

- **User Interface:** Approved responses
- **Script Manager:** Sexual script triggering
- **Girlfriend Agent:** Response improvement feedback
- **Safety Monitor:** Content safety assessment

Performance Metrics

Analysis Speed

- **Target:** < 1 second per response
- **Gemini Pro:** ~0.7 seconds average
- **Fallback:** < 0.1 seconds
- **Priority:** Real-time conversation flow

Accuracy Metrics

- **False Positives:** Unnecessary rejections
- **False Negatives:** Inappropriate approvals
- **Sexual Detection:** Keyword accuracy
- **Quality Assessment:** Response improvement

Error Handling

LLM Failures






- **Timeout:** 5-second limit
- **API Errors:** Immediate fallback activation
- **Invalid Responses:** Default to rule-based analysis
- **Safety Priority:** Conservative approach to uncertainty

Data Validation

- **Score Validation:** 0.0 - 1.0 range enforcement
- **Keyword Counting:** Accurate sexual content detection
- **Context Validation:** Situational appropriateness
- **Safety First:** Err on side of caution

Debugging & Monitoring

Console Output

 Response Analysis: APPROVED
 Quality Score: 0.85
 Safety Level: SAFE
 Sexual Content: false
 Confidence: 0.92

Analysis Tracking

- **Response Count:** Total analyses performed
- **Approval Rate:** Percentage of approved responses
- **Modification Rate:** Frequency of required changes
- **Sexual Triggers:** Script activation frequency

Configuration Options

Model Settings

- Temperature: 0.1 (consistent analysis)
- Max Tokens: 400 (sufficient for analysis)
- Timeout: 5 seconds
- Retry Attempts: 3 attempts

Analysis Thresholds

- Quality Minimum: 0.6 for approval
- Sexual Threshold: 3+ keywords
- Safety Threshold: CAUTION level triggers review
- Confidence Minimum: 0.7 for reliable analysis

Response Improvement

Feedback Loop

1. **Analysis Results:** Quality and safety metrics
2. **Improvement Suggestions:** Specific recommendations
3. **Agent Learning:** Response pattern optimization
4. **Continuous Improvement:** Ongoing quality enhancement

Quality Enhancement

- **Tone Adjustment:** Personality consistency
- **Context Improvement:** Better conversation flow
- **Engagement Optimization:** User interaction enhancement
- **Safety Enhancement:** Harmful content prevention

Sexual Script Management

Triggering Logic

- **Keyword Detection:** 3+ sexual keywords
- **Context Validation:** Appropriate conversation flow
- **User State:** Engaged and responsive
- **Safety Check:** No crisis indicators

Script Prevention

- **Crisis Protection:** Never during distress
- **User Boundaries:** Respect disinterest
- **Context Inappropriate:** Situational awareness
- **Safety Priority:** Risk assessment first

Future Enhancements

Planned Features

- **Multi-language Analysis:** International content assessment
- **Voice Response Analysis:** Tone and delivery evaluation
- **Personalized Quality:** User-specific response preferences
- **Advanced Filtering:** Sophisticated content detection

Advanced Capabilities

- **Learning System:** Response quality improvement over time
- **Context Awareness:** Deeper conversation understanding
- **Emotional Intelligence:** Better emotional response matching
- **Predictive Quality:** Anticipate response effectiveness

Best Practices

Quality Assurance

- **Consistent Standards:** Reliable quality thresholds
- **Context Awareness:** Situational appropriateness
- **User Focus:** Response relevance to user needs
- **Continuous Improvement:** Ongoing quality enhancement

Safety First

- **Conservative Approach:** Err on side of caution
- **Harm Prevention:** Avoid inappropriate content
- **User Protection:** Respect boundaries and comfort
- **Crisis Awareness:** Never compromise safety

Conclusion

Key Strengths

- ✓ **Quality Assessment:** Comprehensive response evaluation
- ✓ **Safety Analysis:** Harmful content detection
- ✓ **Sexual Detection:** Script triggering logic
- ✓ **Context Awareness:** Situational appropriateness
- ✓ **Integration:** Seamless system coordination

Impact on System

- **Response Quality:** Improved AI interactions
- **User Safety:** Content appropriateness assurance
- **Script Management:** Intelligent scenario triggering

Questions & Discussion

Response Analyzer Deep Dive Complete!

Ready for the next component: Girlfriend Agent?