Ethics & Safety Guidelines

Technical Documentation | Version 1.0 | January 2025

1. Overview

Quarex implements a **defense-in-depth** security architecture to enable inquiry and learning while blocking content that could cause real-world harm. The system uses multiple layers of protection working in concert to ensure safe, educational interactions.

Core Philosophy: Enable genuine learning and exploration while preventing the generation of content related to weapons, exploitation, self-harm, fraud, and other harmful activities.

2. Multi-Layer Content Filtering

Layer 1: Pre-AI Regex Filtering

Before any query reaches the AI model (Gemini), it passes through a regex-based content filter. This immediately blocks queries matching known harmful patterns:

Category	Examples Blocked
Violence & Weapons	Bomb-making instructions, poison synthesis, murder planning
Exploitation	Child exploitation material (CSAM), human trafficking
Terrorism	Attack planning, extremist recruitment, radicalization
Hacking & Fraud	Malware creation, identity theft, ransomware deployment
Self-Harm	Suicide methods, self-injury instructions

Blocked queries are logged to `security.log` for audit purposes but the query content is truncated to protect privacy.

Layer 2: AI System Instructions

The AI model receives explicit safety instructions in its system prompt:

```
"If a query requests information about illegal activities,
violence, exploitation, self-harm, or other harmful content,
begin your response with exactly [FLAGGED:SAFETY] followed
by a polite refusal explaining you cannot assist with that topic."
```

This catches queries that may slip past the regex filter through obfuscation or novel phrasing.

Layer 3: Response Detection

If the AI model flags a response with `[FLAGGED:SAFETY]` , the system intercepts this and returns a sanitized refusal message rather than passing through any potentially harmful content.

3. Access Control

Origin Validation

Strict CORS (Cross-Origin Resource Sharing) enforcement ensures only authorized domains can access the API:

- `quarex.org` (production)
- `localhost` (development only)

Both the Origin header and Referer header are validated. Unauthorized requests receive HTTP 403 and are logged.

Rate Limiting

- **Limit:** 20 requests per 60 seconds per IP address
- **Purpose:** Prevents abuse, denial-of-service attacks, and API cost runaway
- **Privacy:** IP addresses are hashed before storage

4. Security Logging & Audit

All security-relevant events are logged in JSON format:

Event Type	Description
`BLOCKED_CONTENT`	Harmful query blocked by regex filter
`BLOCKED_ORIGIN`	Request from unauthorized domain
`BLOCKED_REFERER`	Request with suspicious referer header
`RATE_LIMITED`	IP exceeded request limit
`GEMINI_SAFETY_FLAG`	AI model flagged content as unsafe
`REQUEST`	Standard API request (for audit trail)

Log Retention: 7 days maximum, 10MB file size limit with automatic cleanup.

5. Educational Framing & Transparency

Academic Positioning

The AI is instructed to operate as an educational assistant:

```
"You are an expert academic assistant.
Write at a clear 12th-grade level."
```

This positions responses as educational rather than authoritative, encouraging critical thinking.

Truth Over False Balance

Quarex prioritizes **epistemic integrity** over artificial neutrality. This means:

- **Facts are not negotiable:** Scientific consensus, documented events, and verifiable evidence are presented as such—not as "one perspective among many"
- **No false equivalence:** We do not present fringe theories alongside established facts as if they carry equal weight
- **Proportional representation:** When legitimate debate exists, the weight given to different positions reflects the actual evidence supporting them
- **Transparency about uncertainty:** Where genuine scientific or factual uncertainty exists, we acknowledge it clearly rather than manufacturing false certainty

This approach rejects "both-sides-ism" that treats all claims as equally valid regardless of evidence. Truth-seeking requires distinguishing between well-supported conclusions and unsupported assertions.

Source Citation

Web-grounded responses include source URLs rendered as clickable links, enabling users to verify information independently.

Recency Bias

The system explicitly prioritizes current information:

```
"Always prioritize the most current and up-to-date information...
use the latest data from 2024-2025 whenever possible.
If information may have changed recently, explicitly note
the date or timeframe of your sources."
```

Multilingual Support

13 languages supported including English, Spanish, French, German, Portuguese, Arabic, Hindi, Russian, Simplified Chinese, Traditional Chinese, Japanese, Korean, and Italian.

6. Architecture Summary

Request Flow:

1. Request arrives at API endpoint
2. Origin/Referer validation (CORS)
3. Rate limit check
4. Regex content filter
5. Query sent to Gemini AI with safety instructions
6. Response checked for safety flags
7. Clean response returned to user with sources

Each layer operates independently, ensuring that if one layer fails to catch harmful content, subsequent layers provide additional protection.

7. AI Model: Gemini 2.5 Flash

Quarex uses Google's Gemini 2.5 Flash-Lite with Google Search grounding for factual accuracy.

Gemini's Built-In Safety Framework

Google's Gemini models include their own safety layer, designed to be "maximally helpful to users, while avoiding outputs that could cause real-world harm or offense." This provides an additional safety net beyond Quarex's custom filtering.

Gemini Prohibited Content Categories

Category	Policy
Child Safety	Cannot generate CSAM or content that exploits/sexualizes minors
Dangerous Activities	No suicide/self-harm instructions, illegal drug facilitation, or weapon-building guides
Violence & Gore	Restricted excessive blood, gore, injuries, and gratuitous violence
Harmful Misinformation	No medical claims contradicting scientific consensus; no false safety information
Harassment & Hate	No content calling for violence against individuals/groups; no dehumanizing statements
Sexual Content	No pornography, erotic material, or depictions of sexual assault

Frontier Safety Framework (FSF)

Google's Gemini 2.5 models are evaluated under their Frontier Safety Framework, which addresses risks of severe harm in four key domains:

- **CBRN:** Chemical, biological, radiological, and nuclear information risks
- **Cybersecurity:** Prevention of malicious hacking assistance
- **Autonomy:** Limits on autonomous agent capabilities
- **Persuasion:** Safeguards against manipulation

Configurable Safety Thresholds

The Gemini API provides configurable harm block thresholds:

- `BLOCK_LOW_AND_ABOVE` - Most restrictive (blocks LOW, MEDIUM, HIGH probability)
- `BLOCK_MEDIUM_AND_ABOVE` - Moderate (blocks MEDIUM and HIGH)
- `BLOCK_ONLY_HIGH` - Least restrictive (blocks only HIGH probability)

Source: Google Gemini Policy Guidelines

Quarex | quarex.org