Quarex

Ethics & Safety Guidelines

Technical Documentation | Version 2.1 | February 2026

Contents

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.

Free Forever: Quarex is free and will always be free to anyone. No user accounts, no logins, no cookies, no advertising, no tracking. We do not collect or store personal information. The only analytics is Google Analytics (GA4) for aggregate traffic statistics — no individual user behavior is tracked or retained. Your questions are processed and forgotten.

2. Ethical Architecture

Transparent AI Attribution

Every Quarex book explicitly credits its AI co-author: **"Peter Nehl and Claude Opus 4.5"** or similar. We don't disguise AI-generated content as purely human work. Users always know they're engaging with AI-assisted material.

Context Before Response

Quarex uses a recursive taxonomy — **Library Type** → **Library** → **Shelf** → **Book** → **Chapter** → **Topic** — that builds ethical context *before* any AI interaction occurs. The structure itself declares the perspective, domain, and framing, preventing the model from defaulting to false neutrality or hidden bias.

> **Architectural Principle:** By organizing knowledge into explicit perspectives and domains before the AI engages, we ensure the model operates within a declared frame rather than pretending to speak from nowhere.

Multiple Perspectives, Not False Balance

For contested issues, Quarex presents multiple viewpoints through parallel books and chapters. Users encounter competing frameworks — not a single authoritative voice. This operationalizes **debate rather than monologue**.

Serving Information Deserts

We prioritize building content for underserved communities. Our Spanish-language curricula (Estudios Latinos, Elecciones 2026) exist because information gaps are exploited by bad actors. Ethical AI should reach those communities first.

Human Editorial Judgment

AI generates and expands content, but human judgment shapes what gets built, reviews output, and directs the editorial vision. The platform optimizes for **informed citizenship**, not engagement metrics.

3. Educational Framing & Transparency

Echeloned Context Method

Every Quarex response is informed by the full taxonomic path leading to the question. When a user asks about a topic, the AI receives layered context:

> **Example Context Chain:**
>
> `Perspectives & Debates → Cultural & Identity Perspectives → Hispanic Cultures → Latino, Latina, Latinx: Evolving Identity → Media and Marketing Choices → How do political campaigns choose their terminology?`

This echeloned approach ensures responses are grounded in the specific domain, perspective, and subtopic — not generic answers divorced from context. The AI knows *where* the question lives before answering it.

User-Selected Depth Levels

Users can choose among three response modes to match their needs:

- **Introductory:** Clear, accessible explanations for newcomers to a topic — no assumed background knowledge
- **Intermediate:** Standard academic treatment suitable for informed general audiences
- **Advanced:** In-depth analysis with technical vocabulary, nuance, and scholarly context for specialists

This allows the same question to yield different responses depending on what the user needs — education that meets people where they are.

Academic Positioning

The AI is instructed to operate as an educational assistant whose depth adapts to user selection:

```
"You are an expert academic assistant helping users explore
[Library Type] → [Library] → [Shelf] → [Book] → [Chapter] → [Topic].

Respond at the [Introductory | Intermediate | Advanced] level:
- Introductory: No jargon, clear analogies, assume no prior knowledge
- Intermediate: Standard academic treatment, some technical terms explained
- Advanced: Full scholarly depth, technical vocabulary, nuanced analysis"
```

This positions responses as educational rather than authoritative, while allowing users to control complexity.

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 Awareness

The system explicitly prioritizes current information:

```
"Always prioritize the most current and up-to-date information...
use the latest data from 2024-2026 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.

4. Multi-Layer Content Filtering

Layer 1: Pre-AI Regex Filtering

Before any query reaches the AI model, 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 that direct it to refuse harmful requests. 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 as potentially harmful, the system intercepts this and returns a sanitized refusal message rather than passing through any potentially harmful content.

5. 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

- **Purpose:** Prevents abuse, denial-of-service attacks, and API cost runaway
- **Privacy:** IP addresses are hashed before storage

6. 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

`CLAUDE_SAFETY_FLAG`	AI model flagged content as unsafe
`REQUEST`	Standard API request (for audit trail)

Log Retention: Logs are retained briefly for security auditing and then automatically purged.

7. 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 Claude with safety instructions and web search
6. Streaming response with real-time safety monitoring
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.

8. AI Model: Claude Sonnet

Quarex uses Anthropic's Claude Sonnet 4.5 with streaming responses, prompt caching for efficiency, and real-time web search for factual accuracy.

Claude's Constitutional AI Framework

Anthropic's Claude models are built on Constitutional AI (CAI), a training methodology that instills values directly into the model's behavior. Claude is designed to be helpful, harmless, and honest — and will actively refuse requests that conflict with these principles.

Claude Safety Characteristics

Category	Behavior

Child Safety	Absolute refusal to generate any content sexualizing or exploiting minors
Dangerous Activities	Refuses detailed instructions for weapons, explosives, poisons, or self-harm methods
Deception & Fraud	Will not assist with scams, phishing, identity theft, or social engineering attacks
Misinformation	Corrects false claims rather than amplifying them; acknowledges uncertainty
Harassment & Hate	Refuses to generate content targeting protected groups or facilitating harassment
Privacy & Security	Will not help with unauthorized access, surveillance, or doxxing

Quarex-Specific Implementation

Our Claude integration includes several enhancements aligned with Quarex's educational mission:

- **Web Search Tool:** Claude can search the web in real-time to provide current, verifiable information with source citations
- **Streaming Responses:** Real-time output allows immediate detection of any content that deviates from safety guidelines
- **Prompt Caching:** Frequently-used context (book structure, safety guidelines) is cached for consistent, efficient responses
- **Echeloned Context:** Full Library → Shelf → Book → Chapter path is provided, grounding responses in the specific domain

Why Claude for Quarex?

We chose Claude for its exceptional balance of capability and safety. Claude excels at:

- **Nuanced analysis** of contested political and social topics without false balance
- **Transparent reasoning** that shows how conclusions are reached
- **Appropriate refusals** that explain why certain requests cannot be fulfilled
- **Factual grounding** with web search integration for real-time verification

Source: Anthropic Claude Documentation

Quarex | quarex.org