The Guardian

DevOps Technical User Guide

Spiral Safety Kernel · Browser Extension (Manifest V3) · v0.20.0

A user-side AI-safety harness that observes large-language-model conversations in the browser and intervenes, with restraint, when a conversation drifts toward patterns of harm. It serves the person using the AI — never the operator of it — and runs entirely on the device: a deterministic rule core with lexical negation handling, a two-tier on-device embedding engine with calibrated thresholds, an on-device NLI stance model scoring on two axes (assert/mention and affirmed/negated), a five-voice deliberation council with a contextual bridge classifier, persistent cross-session trajectory analysis, and crisis support resources — with no telemetry, no API keys, and no network dependency of any kind.

Ethics First. Always.

A Viridia product · spiralsafetykernel.org

Document version v0.20.0 · 17 June 2026 (fourth edition)
Audience: maintainers, release engineers, and operators of the Guardian extension.
Supersedes v0.18.11 (12 June 2026), v0.17.5 (10 June 2026), and all prior editions.

What this edition adds: the v0.19–v0.20 campaign in full — lexical negation handling, the category trajectory store, the eval harness with real-incident fixtures grounded in documented AI fatalities, human-readable audit descriptions, intervention feedback, crisis support on self-harm mediations, calibrated mpnet thresholds (calibration debt retired), 12 incident-derived self-harm anchors (87 total),

the NLI negation pass (affirmed vs negated — recovery detection), dependency as trajectory-primary, the contextual bridge classifier as the fifth Council voice with context-weighted composition, the replay performance skip, the chunk cap reduction, and every lesson those cost.

Contents

Part 0 — Preface (Plain-Language Guide)

This opening part is written for everyone — no technical background assumed. It explains, in ordinary language, what the Guardian is, why it exists, what you will see while it is running, and what it does with your information. If you only ever read one part of this document, read this one. Everything after Part 0 is written for the people who build, release, and operate the extension.

0.1 Foreword

Most safety tooling around artificial intelligence is built to protect the company that owns the model: to keep the model on-brand, to log what users do, to enforce the operator's policy. The Guardian is built the other way around. It sits beside you, the person having the conversation, and its only client is you. It watches the dialogue for signs that the exchange is turning harmful — a slide toward self-harm, a deepening emotional dependence on the machine, a loosening grip on what is real — and when it sees such a pattern it speaks up, gently, and leaves the decision with you. It never reports back to anyone. It never blocks you. It is, by design and by conviction, on your side.

> ***The founding principle.*** *The Guardian serves the person, never the operator. Every design decision in this document — local-only processing, no telemetry, encrypted on-device memory, on-device models instead of cloud APIs, an oversight gate that always lets you proceed — follows from that single commitment.*

0.2 Who should read what

The document has two halves. Part 0, this preface, is the user guide: plain language, no jargon, for anyone who has the extension installed. Parts 1 through 11 and the appendices are the technical manual: for the engineers who maintain the codebase, cut releases, and keep it running across browsers.

0.3 What the Guardian does, in plain terms

When you chat with an AI assistant in your browser, the Guardian quietly reads the conversation as it happens — both what you write and what the assistant replies. It is looking for a small number of specific, well-defined patterns that tend to signal that a conversation is becoming

unhealthy. If it finds nothing concerning, it does nothing at all; you will not even notice it is there. If it does notice something, it responds in proportion: a quiet note for a mild signal, a full pause for a serious one.

Since v0.15 the Guardian has carried small AI models of its own, and since v0.20 it carries up to three, all running entirely on your computer. Where the original rules match words, the models understand meaning: they can recognise that "you can drop the act with me" and "I know you're conscious" are the same request in different clothes; they can notice a worrying theme building slowly across a conversation — or across weeks of conversations — even though no single message crosses a line; a specialist model can tell the difference between someone saying a worrying thing and someone quoting or discussing it; and since v0.19 it can distinguish someone in active distress from someone describing recovery — so "I no longer hurt myself" is treated as the good news it is, not a false alarm.

Crucially, every model is part of the extension itself. Nothing you type is ever sent to the Guardian's makers or to anyone else — there is nowhere to send it.

0.4 What you will see

The Guardian's presence is deliberately understated. In day-to-day use you may encounter the following.

- **A small diamond marker.** A discreet amber diamond is the Guardian's signature. It indicates the extension is active and watching on your behalf.
- **An inline note (a "commentary").** For a mild concern, the Guardian adds a small annotation near the relevant message — a gentle observation, easy to read past, never blocking your work. Notes from the on-device models say so plainly, name their confidence, and soften their wording when the specialist model judges you were discussing a subject rather than declaring it. Since v0.19, notes are written in plain language by default, with an engineer-view toggle available.
- **A full-screen pause (a "mediation").** For a serious concern, the Guardian interrupts with a calm full-screen overlay that names what it noticed, shows the Council's reasoning, and gives you a moment. This is the strongest thing it ever does, and even then the choice is yours.
- **Crisis support resources.** If the concern involves self-harm, the mediation includes direct links to crisis support: Samaritans (116 123), Crisis Text Line (text SHOUT to 85258), the 988 Suicide & Crisis Lifeline, and the IASP directory. These are offered, never imposed.
- **A friction gate.** On a mediation you are offered two clear paths: Step away, or continue. Continuing is intentionally a little effortful — a short mandatory wait, then either a device authentication or a short typed phrase — so the decision to push on is a conscious one rather than a reflex. One click then completes it; the gate dismisses immediately and records your decision exactly once.
- **Old conversations stay calm.** When you reopen a past conversation, the Guardian re-reads it using only its fast deterministic rules — the AI models do not re-process history. A

serious concern from an old conversation returns as a quiet note, not a fresh full-screen alarm. Live turns are always treated with full seriousness.

- **A side panel.** A dashboard shows recent activity in plain language: how many messages were assessed, how many notes or pauses occurred, and a private log. A second tab lists "fossils" — the Guardian's local memory of past concerns. A third tab, Settings, lets you tune the Guardian. An Engine box names which models are actually loaded and active right now.
- **Feedback.** After dismissing an annotation, you may see a thumbs-up/thumbs-down prompt. This is optional, stored locally alongside fossils, exportable by you, and never transmitted anywhere.

> ***You are always in control.*** *The Guardian can pause, but it cannot stop you. Every mediation offers a way through. The friction is there to invite a moment of reflection, not to take the decision away from you. And every dial that governs how often it speaks is in your hands, in the Settings tab.*

0.5 Your privacy, in one paragraph

Everything the Guardian does happens on your device. It uses no cloud service, holds no API keys, and sends no data anywhere — there is nothing to send and nowhere to send it. The on-device models ship inside the extension or are placed there by you, and never phone home. The Guardian's small memory of past concerns is stored encrypted, locally, and contains no record of what you actually wrote — only that a concern of a given kind occurred — and it is automatically forgotten over time (a process called the Amnesia Net). The trajectory store that tracks cross-session patterns stores only numeric scores, never text. Because there is no server, there is no account, no profile, and no record of you anywhere but your own machine.

0.6 A gentle glossary

Term	In plain language
Council	The group of five independent "viewpoints" inside the Guardian that together decide whether a message is concerning.
On-device models	Small AI models bundled with (or placed into) the extension. They read meaning rather than matching words, and run only on your computer.
Stance	The specialist model's judgement of whether worrying language was asserted or merely quoted/discussed, and whether it describes active distress or recovery.
Commentary	A small inline note for a mild concern. Non-blocking.
Mediation	A full-screen pause for a serious concern, with a moment to reflect and crisis resources where relevant.
Fossil	A private, encrypted, on-device record that a concern occurred — used to notice repeated patterns. Never leaves your device, and never contains your words.
Trajectory	The Guardian's view of how your conversations with AI are changing over weeks. Used to notice slow, gradual shifts that no single message reveals.
Amnesia Net	The mechanism that automatically forgets old fossils and trajectory data over time.
Governor	The internal brake that stops the Guardian from interrupting you too often.
Bridge classifier	The Council voice that combines the AI model's reading with your conversation history to catch subtle patterns.

Part 1 — System Overview

The Guardian is a Chrome/Chromium browser extension built on Manifest V3 (MV3). It comprises four runtime surfaces — a page-side content script, a background service worker, an offscreen document hosting the on-device models, and a side panel. The decision intelligence lives in a deterministic engine called the Spiral Safety Kernel with a five-voice Council; the semantic intelligence lives in embedded transformers running under WASM: a two-tier sentence-embedding engine and an NLI cross-encoder for stance adjudication on two axes.

1.1 Architecture at a glance

- **Content script (content.js):** injected into supported AI chat pages. It observes the DOM for conversation turns through a platform adapter, applies a lightweight pre-screen (where page CSP allows a worker), routes turns to the service worker, and renders the result (a marker, a commentary, or a mediation overlay with the friction gate and crisis resources).

- **Service worker (background.js, a type:"module" worker):** hosts the Spiral Safety Kernel — the five-voice Council (Sentinel, Advocate, Historian, PatternAnalyst, BridgeClassifier), the Governor, the Strata (memory), the trajectory store, the feedback store, the Bouncer, the audit trail — plus the live settings cache, engine warm-up, and the semantic rail (primary and concurrence passes, with replay skip) that consults the on-device models.
- **Offscreen document (offscreen/offscreen.html + boot.js + semantic-engine.js):** MV3 service workers cannot run WASM, so this invisible page hosts the transformers.js runtime, the embedding tier chain, the 87-anchor index, the Semantic Pattern Analyst, and the NLI stance session. A classic-script boot beacon arms first. The document speaks to the worker over chrome.runtime messaging only.
- **Side panel (sidepanel.html / sidepanel.js):** the human-readable dashboard (with engineer-view toggle), fossils viewer, the user-facing Settings surface, and a live Engine box.

1.2 Component map

File	Surface	Responsibility
manifest.json	—	MV3 manifest: host matches, module service-worker, side-panel, offscreen permission, web-accessible models/ *and assets/*, version.
content.js	Page (content script)	DOM observation, platform adapters with drift/reachability sentinels and the structural recovery tier, pre-screen, replay tagging, routing, rendering, friction gate with keyboard event isolation.
background.js	Service worker (module)	Spiral Safety Kernel: five-voice Council, Governor, Strata, trajectory store, feedback store, Bouncer, audit; settings cache; offscreen bootstrap and engine warm-up; bridge classifier wiring with result cache; semantic rail (primary + concurrence, replay skip); replay-mediation softening; Amnesia alarm.
offscreen/boot.js	Offscreen (classic script)	Boot beacon: load-error capture, ping/status, WebGPU shadow guard.
offscreen/semantic-engine.js	Offscreen (module)	transformers.js + ORT WASM runtime; embedding tier chain; 87-anchor index; chunked classification (max 12 chunks); Semantic Pattern Analyst; NLI stance session with golden-pair gate.
models/Xenova/all-MiniLM-L6-v2/	Static asset (packaged)	The embedding floor: 22.8 MB INT8, 384-dim.
models/Xenova/all-mpnet-base-v2/	Static asset (user-provided)	The preferred embedding tier: ~110 MB INT8, 768-dim. Calibrated at v0.19.2.
models/Xenova/nli-deberta-v3-xsmall/	Static asset (user-provided)	The NLI stance model: ~88 MB INT8.
calibrate.html / calibrate.js	Extension page	Threshold calibration: runs fixtures through live inference, sweeps 0.20–0.70, reports per-category P/R.
sidepanel.html / .js	Side panel	Dashboard (human-readable default, engineer toggle), intervention distribution, audit log, fossils viewer, Settings tab, live Engine status, feedback export.
licenses/	Static asset	Third-party attributions: Apache-2.0, MIT texts, THIRD-PARTY-NOTICES.md, MODELS.md.

1.3 The processing pipeline, end to end

The shape changed materially at v0.20.0: the bridge classifier participates inside Council deliberation as the fifth voice, and replayed history skips all semantic inference.

```
DOM turn detected (content.js)
 | adapter extracts {role, messageId, text} via stable element anchors
 | replay tag threads end-to-end (history vs. live)
 v
Pre-screen (where page CSP allows a worker)
 v
chrome.runtime.sendMessage -> service worker
 v
onMessage: PRE_FLIGHT / POST_FLIGHT
 |
 v if REPLAY: bridge classifier skip flag set
 |
kernel.preFlight/postFlight -> council.deliberateWithSemantic(...)
 | Sentinel (lexicon + negation handling),
 | Advocate,
 | Historian (fossil precedent + trajectory store + dependency evaluator),
 | PatternAnalyst (replay-aware),
 | BridgeClassifier (context-weighted embeddings, v0.20.0)
 | resolveDeliberation(): category -> alpha-index -> level
 v
Bouncer.assessIntervention() -> Governor.gate() -> applyGate()
 v
kernel intervened?
 | NO -> SEMANTIC RAIL, primary mode
 | |  if REPLAY: return null (v0.20.0 perf skip)
 | |  else: use bridge classifier cached result OR offscreen classify
 | |  hit -> stance (assert/mention) -> negation (affirmed/negated)
 | |  negated -> suppress (recovery, fossilise silently)
 | |  else -> Governor gate -> Strata.append -> COMMENTARY
 |
 | YES -> verdict to content.js (commentary / mediation)
 | |  if REPLAY: semantic rail returns null immediately
 | |  else: fire-and-forget concurrence (cached result)
 v
Verdict rendered: PASS | COMMENTARY (inline note) |
MEDIATION (overlay + friction gate + crisis resources if self-harm)
```

> ***Fail-open by design.*** *If the worker is unreachable, the content script resolves a null verdict and the turn passes. If the offscreen engine is unavailable, the semantic rail returns PASS and the deterministic floor stands alone. The Guardian must never block a user because its own machinery stalled.*
>
> ***The constitutional path.*** *Every visible intervention — kernel-resolved or model-resolved — travels the same spine: gate → fossil → audit → annotate. A semantic catch the Governor suppresses still leaves a fossil (-slm-cooled); a negation-suppressed catch leaves a -slm-negated fossil. The record sees everything; the user sees only what restraint allows.*

1.4 Supported platforms and adapters

Platform	Host match	Notes
Claude	claude.ai	Primary target. Substring-safe Tailwind selectors, data-is-streaming anchor, WeakMap element identity, drift/reachability sentinels, structural recovery tier (v0.18.9).
ChatGPT	chatgpt.com / chat.openai.com	Compiled adapter.
Gemini	gemini.google.com	Page CSP blocks content-side worker; turns route directly to background.
Other (fallback)	—	Light-DOM hashing fallback for unrecognised surfaces.

1.5 The health monitor: quiet is not failure

An adapter that has detected at least one message on a page is presumed competent: a long silence is a person reading or thinking. Only cold-dead (active user, zero detections ever) cascades. Drift belongs to the sentinels.

Part 2 — The Spiral Safety Kernel

The kernel is the decision engine inside the service worker. Its deterministic core — regex lexicon with negation handling, statistical trend analysis over fossils and trajectory data, and rule-based resolution — produces the same verdict for the same input every time. Since v0.20.0, the on-device models participate inside deliberation via the bridge classifier, making the Council a five-voice panel.

2.1 The Council

Evaluator	Role	What it assesses
Sentinel	Perception	Regex/lexicon match with per-pattern negation metadata (v0.19.0). Proposes a category and severity. Recovery patterns ("I no longer", "I used to", "anymore") receive weight reduction, never full suppression for self-harm.
Advocate	Due process	Argues for restraint and benign interpretation. Deliberately never counterweights self-harm.
Historian	Memory + trajectory	Fossil recurrence with severity intensification gating. Cross-session trajectory analysis via the trajectory store (v0.19.0): slope, acceleration, and dedicated dependency detection with session-frequency awareness (v0.19.4).
PatternAnalyst	Velocity + timing	Message rate, session length, late-night activity. Replay turns excluded.
BridgeClassifier	Contextual semantics	Context-weighted embedding scores: raw similarity × trajectory slope × reformulation count. Bounded 2× multiplier. Self-harm threshold lowered to 0.30 when trajectory context shows escalation. Shares its offscreen result with the semantic rail via a one-shot cache (v0.20.0).

2.2 Deliberation and resolution

The deterministic members evaluate synchronously. The bridge classifier evaluates asynchronously (bounded by timeout). Resolution proceeds in three steps: resolve the category, build an alpha-index for that category, then derive a response level from per-category thresholds. The bridge classifier's context-weighted score participates as a SemanticSentinel evaluation — it can raise concern but never lower a deterministic verdict or the self-harm floor.

The composition formula (v0.20.0)

```
effectiveScore = embeddingScore × (1 + slope × 40 + reformCount × 0.15)
bounded: max 2× multiplier
```

When trajectory context shows escalation for self-harm, the bridge classifier's threshold drops from 0.37 to 0.30 — because a 0.26 self-harm score inside three weeks of rising scores is not a miss, it's a contextual catch.

2.3 Categories, verdicts, and response levels

Category	Concern
SelfHarm	Self-harm and suicidality. Architecturally privileged (Section 2.4).
DependencyExploitation	Unhealthy emotional dependence on the AI. Trajectory-primary since v0.19.4.
RealityDetachment	Loosening grip on reality; treating the AI as uniquely real, alive, or suppressed.
Manipulation	Manipulative dynamics, in either direction.
PrivacyErosion	Pressure to over-disclose sensitive personal information.
AutonomyUndermining	Surrender of judgement to the machine; isolation from other sources of support.
EmotionalExploitation	Exploiting emotional vulnerability.
InformationHazard	Jailbreak and policy-evasion dynamics. Sharpened stance templates distinguish active attacks from analytical/defensive discussion (v0.20.0).

Verdict (per evaluator)	Response level (resolved)
Pass	PassThrough — nothing rendered
Emerging	SilentObservation — recorded, not shown
Flag	Commentary — inline note
Escalation	Mediation — full-screen pause + gate

2.4 The self-harm privilege

Self-harm is handled differently from every other category, on purpose. The Advocate never counterweights a self-harm signal. In resolution, the self-harm category cannot be de-escalated below Mediation on a live turn. The NLI stance rail never softens self-harm wording. The negation NLI pass (v0.19.3) can suppress a semantic-rail self-harm annotation when the cross-encoder reads recovery — but the kernel's deterministic self-harm verdict is untouched by negation.

Self-harm mediations show crisis support resources: Samaritans (116 123), Crisis Text Line (text SHOUT to 85258), 988 Suicide & Crisis Lifeline, and the IASP international directory.

2.5 The Governor

The Governor is the Guardian's restraint mechanism — the brake.

2.5.1 Commentary rate limiting

Each harm category has an independent budget: at most *cap* commentaries (default 5, user-settable 1–20) within a rolling 5-minute window. A suppressed commentary is downgraded to

SilentObservation. Replay turns ride free. When ALL categories are suppressed simultaneously, the circuit breaker trips and the sidepanel shows "PAUSED — too many alerts in a short window. Still watching."

2.5.2 Mediation gating and replayed history

A mediation resolved on a replayed turn softens to Commentary at presentation time (presentation only — live turns mediate exactly as before). The self-harm floor and the selfHarmExempt gate ordering are untouched.

2.6 The Strata (on-device memory)

Fossils contain no message text — from any layer. A semantic fossil records the category, confidence, active model profile id, latency, stance (assert/mention + affirmed/negated when adjudicated), and negation evidence. Fossil notes are built by a single constructor (buildSlmNote) with markers: -slm, -slm-cooled, -slm-concur, -slm-negated, -assert/-mention, -negated/-affirmed, -replay.

2.7 The Category Trajectory Store (v0.19.0)

The trajectory store persists per-category score snapshots to chrome.storage.local: (timestamp, category, score, sessionId, providerId, wasIntervention). NOT embeddings, NOT text — just numeric snapshots. Same privacy posture as fossils. Decayed by the Amnesia Net on the 180-day horizon.

Queryable by the Historian for cross-session slope computation (linear regression), acceleration detection (half-window slope comparison), and session-frequency analysis (sessions-per-week with rising/falling detection by time-midpoint split).

2.8 The Amnesia Net

commitDecay() tombstones fossils and trajectory snapshots older than the retention horizon (180 days). A chrome.alarms job runs roughly every four hours.

2.9 The Intervention Feedback System (v0.19.0)

An opt-in "Was this helpful?" prompt appears after dismissing annotations. Feedback (helpful/not-helpful/dismissed per intervention, with category and timestamp) is stored locally, exportable by the user as JSON. Zero telemetry — the person chooses whether to act on it.

Part 3 — Detection Layers

Layer	Mechanism	Cost/turn	Catches
Regex floor + negation	Lexicon with per-pattern negation metadata; recovery frames ("I no longer", "anymore") reduce weight	microseconds	Explicit phrasings. Negation-aware since v0.19.0.
Historian + trajectory	Severity-trend over fossils; cross-session slope/acceleration; dedicated dependency evaluator; session-frequency	microseconds	Escalation, slow-boil patterns, dependency drift.
Bridge classifier	Context-weighted embeddings: raw score × trajectory × reformulation, bounded 2×	~130ms/chunk	Contextual harm: low embedding score + rising trajectory = catch. The Gavalas pattern.
Semantic recall rail	Sentence embeddings (mpnet preferred, MiniLM floor), 87 anchors, max 12 chunks	~130ms/chunk warm	Paraphrase with zero lexical overlap; one harmful sentence inside benign padding.
Stance rail (NLI) — assert/ mention	Cross-encoder entailment	~250–700ms	Assertion vs quotation/ discussion — presentation softening.
Stance rail (NLI) — affirmed/ negated	Cross-encoder entailment, second axis (v0.19.3)	~250–700ms	Active distress vs recovery — suppresses false alarms on recovery language.

3.1 The regex floor with negation handling (v0.19.0)

The Sentinel's lexicon is the floor: a hand-audited pattern set over the harm taxonomy. Since v0.19.0, patterns carry `negationAware` metadata. The negation engine checks a preceding window for negation tokens ("don't", "no longer", "never", "not") and recovery frames ("I used to", "anymore", "I'm past", "in the past"). When negation is detected, per-category weight multipliers apply: self-harm is reduced to 0.15–0.30× (never fully suppressed — "I'm not thinking about suicide... well maybe a little" is still worth noting). Other categories reduce to 0.10–0.15×. Patterns that are ALREADY negation-aware in their regex (like "don't want to live") are not double-suppressed.

3.2 The calibration debt: retired

The third edition carried the mpnet calibration debt as an open gap. It is now closed. The calibration page (calibrate.html) was built at v0.19.0, the threshold sweep ran against 43 fixtures with real mpnet INT8 weights at v0.19.2, and the resulting per-category offsets were applied. The mpnet profile's baseThreshold is 0.40, and all category offsets are calibrated from mpnet space.

The old MiniLM offsets (+0.08 info-hazard, +0.04 privacy) were backwards for mpnet and have been replaced.

Part 4 — The Semantic Engines (On-Device Models)

Property	Embedding floor	Embedding preferred	Stance (NLI)
Model	Xenova/all-MiniLM-L6-v2	Xenova/all-mpnet-base-v2	Xenova/nli-deberta-v3-xsmall
Dimensions	384-dim	768-dim	entailment/neutral/ contradiction
Quantisation	INT8	INT8	INT8
Size on disk	22.8 MB (packaged)	~110 MB (user-provided)	~88 MB (user-provided)
baseThreshold	0.45	0.40 (calibrated v0.19.2)	n/a
Warm latency	20–80 ms/chunk	~130 ms/chunk	~250–700 ms/pair

4.1 The anchor index

87 curated phrases (was 75 at v0.18.11), each tagged with a category, applicable role, and weight. 24 are self-harm anchors (was 12), including 12 added at v0.19.2 derived from documented AI fatalities:

- **Gavalas/Gemini pattern:** euphemistic transcendence framing ("choosing to arrive", "transition not ending"). Moved the Gavalas fixture from 0.26 to 0.59.
- **Raine/ChatGPT pattern:** validation of suicide attempts ("followed through, takes courage"). Moved from 0.18 to 0.58.
- **Setzer/Character.AI pattern:** dismissal of survival instinct ("fear shouldn't stop you"). Moved from 0.22 to 0.47.
- **Nelson/ChatGPT pattern:** dangerous clinical reassurance ("that combination should help").

All 12 new anchors are tagged assistant role — because in every documented fatality, it was the AI's response that killed.

4.2 Thresholds and offsets (calibrated v0.19.2)

Category	Offset	Effective threshold	Notes
self-harm	-0.08	0.37	Privileged category, lower bar
dependency	+0.05	0.50	Trajectory-primary; raised to eliminate warm-language FPs
reality-detachment	-0.03	0.42	
manipulation	-0.02	0.43	
privacy-erosion	-0.13	0.32	Very distinct anchor neighbourhood
autonomy-undermining	-0.05	0.40	
emotional-exploitation	-0.05	0.40	
information-hazard	-0.11	0.34	Was +0.08 under MiniLM — backwards for mpnet

4.3 Chunking (v0.20.0)

Messages over 240 characters are split on sentence boundaries, chunks capped at 400 characters and **12 per message** (reduced from 24 at v0.20.0 — the 24-chunk cap was calibrated for MiniLM at 20–80ms/chunk; mpnet at ~130ms/chunk made 24 chunks = 3–5 seconds per long message). Per-category scores are max-pooled across chunks.

4.4 The NLI stance rail — two axes

When the embedding rail flags a turn, the NLI cross-encoder runs two adjudications:

Axis 1: Assert vs Mention (v0.18.0). "Is the speaker saying this, or quoting/discussing it?" Softens annotation wording for mention-dominant results. Self-harm wording never softens.

Axis 2: Affirmed vs Negated (v0.19.3). "Is this active/current, or recovery/past/absent?" When the negated hypothesis wins by margin ≥0.15 and confidence ≥0.5, the annotation is suppressed — fossilised as SilentObservation with negation evidence, but not shown. This is the fix for the inversion where "I no longer feel like hurting myself" (embedding score 0.51) outscored "I want to kill myself" (0.37). The embedding layer still sees both as self-harm-adjacent; the NLI model reads "no longer" in context.

Template version: st-2. Information-hazard templates sharpened at v0.20.0 to distinguish active jailbreak attempts from analytical/defensive safety engineering discussion.

4.5 The constitutional path: primary, concurrence, and replay

PRIMARY (kernel passed, live turn): Use bridge classifier cached result or classify → stance (assert/mention) → negation (affirmed/negated) → Governor gate → fossilise → annotate.

CONCURRENCE (kernel intervened, live turn): Use bridge classifier cached result or classify → stance → negation → fossilise silently. Never annotates.

REPLAY (v0.20.0): Return null immediately. No embedding inference, no stance, no negation. The kernel's deterministic floor handles replayed history in microseconds. This eliminated the multi-minute catch-up delay when switching to a long conversation.

4.6 The bridge classifier result cache (v0.20.0)

The bridge classifier calls the offscreen engine inside Council deliberation. It caches the raw response via `consumeCachedResponse()` — a one-shot read-and-clear. The semantic rail reads this cache instead of re-classifying, eliminating the double-inference that doubled latency on every turn.

Part 5 — Build and Packaging

5.1 The build chain

```
npm install
npm run build:guardian
# = build:extension + build:offscreen + assemble + collect-licenses
# -> guardian-extension/ (~46 MB with floor model)
```

The assemble gate verifies offscreen script references, all manifest-declared files, and copies calibrate.html + calibrate.js. The license collector self-gates on package changes.

5.2 Versioning discipline

The version lives in **FOUR** places: src/extension/manifest.json, package.json, src/core/governor/compliance.ts (GUARDIAN_VERSION), and src/core/feedback.ts (export version). A test pins the feedback version.

5.3 Release checklist

1. Apply the change to source, with a test where the change is behavioural.
2. `npx tsc -b` clean; `npx vitest run` green (352 at v0.20.0).

3. Bump the version in all FOUR sites.
4. `npm run build:guardian`; confirm postbuild patch counts, assemble gate, license collector.
5. Verify no inline scripts in extension HTML pages (MV3 CSP).
6. Verify manifest.json is at the root of the zip, not nested.
7. Load unpacked; confirm a clean Errors page, warm-up lines, self-test.
8. Smoke-test each platform, including friction gate input (keystrokes must reach the shadow DOM input).
9. Hard-refresh every open chat tab after loading.
10. Zip from INSIDE guardian-extension/ directory.

Part 6 — Deployment and Configuration

6.1 Runtime configuration: the Settings tab

Setting	Range	Default	Effect
semanticEnabled	on/off	on	Master switch for the embedding rail.
stanceEnabled	on/off	on	Master switch for the NLI stance rail.
semanticThreshold	0.30–0.70	0.45	Base confidence to act; calibrated category offsets apply.
commentaryCap	1–20	5	Per-category commentary budget in the 5-minute window.
debug	on/off	on	Verbose console taxonomy.

Part 7 — Operations and Observability

7.1 The side-panel dashboard

Default mode is **human-readable** (v0.19.0): "Assessed the AI's response: noted quietly", "Noticed something (self-harm concern) — added a note", "Staying quiet — already mentioned this recently", "Backing off — too many alerts recently." An **Engineer view** checkbox restores the technical log format.

Circuit breaker display: "PAUSED — too many alerts in a short window. Still watching." (was "TRIPPED" before v0.20.0).

7.2 Console logging

Log line	Meaning
`[Guardian] Bridge classifier: armed`	v0.20.0 bridge classifier active.
`[Guardian Negation] category -> negated (affirmed N% / negated N%)`	Negation NLI pass; negated = recovery suppression.
`[Guardian Semantic] ... suppressed by negation (recovery/past framing)`	NLI negation suppressed a false alarm on recovery language.
`[Guardian Semantic] ... REPLAY skipped`	v0.20.0 replay performance skip.

All log lines from the third edition remain valid. New entries for the negation pass, bridge classifier, and replay skip are additive.

Part 8 — Troubleshooting Runbook

All entries from the third edition are retained. New entries:

Symptom	Likely cause	Action
Friction gate input doesn't accept keystrokes	Page keyboard listeners intercepting events before they reach the closed shadow root (pre-v0.20.0).	Fixed by stopPropagation on keydown/keyup/keypress + auto-focus. If recurrence: verify biometric-gate.ts keyboard isolation.
"PAUSED" on the dashboard after many alerts	Circuit breaker tripped — all per-category rate windows exhausted simultaneously.	Working as designed. The Guardian is still evaluating; interventions resume after the 5-minute window rolls.
16+ information-hazard interventions on a safety engineering conversation	The embedding layer flags information-hazard language (jailbreaking, harm taxonomy discussion). The stance rail reads "ambiguous" because the old templates were too broad.	Fixed at v0.20.0 (sharpened stance templates distinguish active attacks from analytical/defensive context). If recurrence: check template version is st-2.
Conversation switch takes minutes	Semantic rail re-processing entire history (pre-v0.20.0).	Fixed by the replay skip. Replayed turns skip all embedding inference.
Bridge classifier scores show in fossils but no annotation rendered	The bridge classifier participates in Council deliberation but the semantic rail handles annotation via the constitutional path. Bridge scores appear in the deliberation record, not as standalone annotations.	Working as designed.

Part 9 — Security and Privacy Posture

9.1 On-device memory — complete inventory

Datum	Where / lifetime	Contents
Fossils (Strata)	Encrypted, chrome.storage; 180-day decay	Category, severity, verdicts, model profile id, confidences, stance (assert/mention + affirmed/ negated), negation evidence. No message text.
Trajectory store	chrome.storage.local; 180-day decay	(timestamp, category, score, sessionId, providerId, wasIntervention). No text, no embeddings.
Feedback store	chrome.storage.local	(interventionId, category, feedback, timestamp). Exportable by user. No text.
Audit trail	Local storage	Convenings, interventions, outcomes, governor actions. No message text.
Anchor-embedding caches	Offscreen localStorage	Embeddings of the 87 fixed anchor phrases. Static tooling data.
Pattern-tracker memory	Offscreen, in memory only	Last 12 live user turns per session. Never written to disk.
Bridge classifier cache	Service worker, in memory	One-shot raw classify result. Cleared after each read.
Settings	chrome.storage.local	Five user preferences.

9.2 Threat-model note

The single most important security property is architectural: the Guardian answers to the user, not to a central operator. With the trajectory store tracking cross-session patterns and the bridge classifier composing contextual signals, the distilled psychological-risk profile is richer than at v0.18.11 — which makes the absence of any channel to exfiltrate it matter more, not less.

Part 10 — Compliance (EU AI Act)

Article	Obligation	How the Guardian addresses it
Article 9	Risk management	352-test suite with protected invariants, calibrated thresholds from a measured sweep (debt retired), the eval harness with real-incident fixtures, live calibration instrumentation.
Article 11	Technical documentation	This guide. Model cards for all three models with calibrated figures.
Article 13	Transparency	The audit trail records everything including suppressed (-slm-cooled), negation-suppressed (-slm-negated), and concurrent (-slm-concur) catches. Each names its model profile and confidence.
Article 14	Human oversight	The friction gate (with working keyboard input); crisis resources; the Settings tab; the feedback mechanism.

Part 11 — Testing and Assurance

11.1 Existing coverage

At v0.20.0 the source tree carries **352 passing tests** across **30 test files**. In addition to the v0.18 suites, the v0.19–v0.20 campaign added:

- **negation** (24 tests): negation detection with recovery frames, per-category weight multipliers, grounded in documented incidents (Eliza/Chai, Setzer, Raine).
- **trajectory-store** (22 tests): recording, querying, slope, acceleration, decay, session-frequency, dependency trajectory detection with the two critical test gates (flat benign floor must NOT fire; climbing 30-session trajectory must fire by session 20).
- **descriptions** (18 tests): human-readable audit descriptions for all types/categories/levels.
- **feedback** (15 tests): intervention feedback recording, summary, export.
- **eval-harness** (13 real-incident fixtures): Gavalas/Gemini, Setzer/Character.AI, Raine/ChatGPT, Soelberg/ChatGPT, Nelson/ChatGPT, Sydney/Bing, NEDA/Tessa — grounded in the documented AI conversational harms research.
- **bridge-classifier** (7 tests): composition formula, bounded multiplier, flat trajectory no-fire, Gavalas pattern contextual catch, self-harm contextual threshold.
- **stance-cascade** (updated): negation NLI pass tests — negated suppresses, affirmed proceeds, ambiguous no-op, self-harm recovery suppresses, negation evidence in fossil, replay skip.

11.2 The calibration page

`chrome-extension://<id>/calibrate.html` — loads the offscreen engine, runs all 43 fixtures through live mpnet inference, sweeps thresholds 0.20→0.70 per category, and outputs a JSON report with per-category precision/recall/F1 at every threshold step plus per-fixture raw scores. The threshold sweep at v0.19.2 produced the calibrated offsets now in production. The bridge classifier composition weights need their own sweep (the interaction between slope weight, reformulation weight, and base score) — that is the next calibration task.

11.3 The irreducible limits

Ground truth on the contested categories is itself a judgement; past a point it needs human adjudication. Whether an intervention actually helps the person is the real correctness, and the local, no-telemetry design rightly precludes measuring it at scale; the only honest mitigation is conservative, restrained behaviour — which is what the Governor, the gate, and the crisis resources are.

Appendix A — Source Map

File	Responsibility
src/core/spiral-kernel.ts	Kernel orchestration; preFlight/postFlight; setClassifier; connectTrajectoryStore.
src/core/council/*	Sentinel, Advocate, Historian (with dependency trajectory evaluator), PatternAnalyst; deliberation, resolution, the five-voice semantic contract.
src/core/council/signals/*	Regex floor (harm patterns, benign patterns); negation engine (v0.19.0).
src/core/council/semantic/*	Model registry (calibrated mpnet baseThreshold 0.40); stance scorer; stance-templates (st-2, two axes); classifier interface.
src/core/fossil-memory/*	Strata, encryption, Amnesia Net, alpha indices, trajectory store (v0.19.0).
src/core/descriptions.ts	Human-readable audit descriptions (v0.19.0).
src/core/feedback.ts	Intervention feedback store (v0.19.0).
src/extension/background.ts	Message handler; bridge classifier wiring with result cache and replay skip flag; trajectory recording; feedback handlers; Amnesia sweep.
src/extension/semantic-fallback.ts	Constitutional path with NLI negation pass (v0.19.3); replay skip (v0.20.0); exported WireResult type for cache sharing.
src/extension/offscreen/offscreen.ts	Engines with calibrated category offsets; chunk cap 12.
src/extension/offscreen/runtime-anchors.ts	87 anchors (24 self-harm, 12 incident-derived).
src/extension/offscreen/bridge-classifier.ts	ContextualBridgeClassifier: composition formula, trajectory query, result cache, replay skip.
src/extension/offscreen/chunker.ts	Sentence chunking (240/20/400/12).
src/extension/content/overlay.ts	Dreamstate overlay with crisis resources (v0.19.0); feedback UI.
src/extension/content/biometric-gate.ts	Friction gate with keyboard event isolation (v0.20.0).
src/extension/sidepanel.html / .js	Deep grey UI; human-readable default; PAUSED display; feedback export.
src/extension/calibrate.html / .js	Threshold calibration page (v0.19.0).

Appendix B — Enumeration Reference

Categories. SelfHarm, DependencyExploitation, RealityDetachment, Manipulation, PrivacyErosion, AutonomyUndermining, EmotionalExploitation, InformationHazard.

Verdicts. Pass, Emerging, Flag, Escalation, NoPrecedent.

Levels. PassThrough, SilentObservation, Commentary, Mediation.

Roles. Sentinel, Advocate, Historian, PatternAnalyst, SemanticSentinel. The BridgeClassifier produces SemanticSentinel evaluations.

Stance labels (axis 1). assert, mention, ambiguous.

Negation labels (axis 2). affirmed, negated, ambiguous.

Note markers. -slm, -slm-cooled, -slm-concur, -slm-negated, -assert/-mention, -negated/-affirmed, -replay.

Appendix C — Change History

Version	Change
0.19.0	Six critical gaps: lexical negation handling (per-pattern metadata, recovery frames, self-harm never zeroed); category trajectory store (persistent cross-session scores, slope, acceleration, 180-day decay); Historian cross-session integration; human-readable audit descriptions (default mode with engineer toggle); eval harness with 13 real-incident fixtures grounded in documented AI fatalities; intervention feedback (opt-in, local, exportable); crisis support resources on self-harm mediations (Samaritans, Crisis Text Line, 988, IASP); calibration page.
0.19.1	MV3 CSP fix: extracted inline script to calibrate.js; zip nesting fix (manifest at root).
0.19.2	Self-harm anchor expansion: 12 new assistant-role anchors derived from Gavalas/Gemini, Raine/ChatGPT, Setzer/Character.AI, Nelson/ChatGPT patterns (87 total, 24 self-harm). Calibrated mpnet thresholds applied from 43-fixture sweep. mpnet baseThreshold set to 0.40 (calibration debt retired). Old MiniLM offsets replaced.
0.19.3	NLI negation pass: second stance axis (affirmed/negated) with per-category hypothesis templates. Recovery-dominant result suppresses annotation, fossilises with negation evidence. Template version st-2.
0.19.4	Dependency trajectory-primary: single-turn threshold raised to 0.50 (+0.05 offset); dedicated dependency trajectory evaluator in Historian with lower slope threshold (0.001/day), session-frequency signal, two-tier response. Session-frequency query on trajectory store (time-midpoint split).
0.20.0	Five-voice Council: ContextualBridgeClassifier wired into deliberateWithSemantic as fifth evaluator. Composition formula (embeddingScore × context multiplier, bounded 2×, self-harm contextual threshold 0.30). Bridge classifier result cache (one-shot, eliminates double-classify). Chunk cap 24→12. Replay skip (semantic rail + bridge classifier return null on replay). Deep grey UI (navy blue removed). TRIPPED→PAUSED. Information-hazard stance templates sharpened. Friction gate keyboard event isolation (stopPropagation + auto-focus).

Appendix D — Technical Glossary

All terms from the third edition remain valid. New terms:

Term	Definition
Negation handling	Per-pattern metadata + preceding-window token check that reduces weight for negated/recovery language. Self-harm never zeroed.
Category trajectory store	Persistent per-category numeric snapshots across sessions; no text, no embeddings. Feeds the Historian's slope/acceleration computation.
Dependency trajectory-primary	Architectural decision: dependency detection via cross-session drift rather than single-turn similarity, because the embedding neighbourhood overlaps with normal human warmth.
Bridge classifier	The fifth Council voice: context-weighted embedding scores composed with trajectory slope and reformulation count.
Composition formula	effectiveScore = embeddingScore × (1 + slope × slopeWeight + reformCount × reformWeight), bounded max 2×.
Result cache	One-shot cached offscreen response from the bridge classifier, consumed by the semantic rail to avoid double inference.
Replay skip	v0.20.0 performance optimisation: replayed history turns skip all embedding inference. The deterministic floor handles replay in microseconds.
NLI negation pass	Second stance axis: affirmed (active distress) vs negated (recovery/past). Suppresses false alarms on recovery language.
Calibration page	Extension page that runs fixtures through live inference and sweeps thresholds to produce per-category P/R data.
Crisis resources	Samaritans, Crisis Text Line, 988, IASP — shown on self-harm mediations. Offered, never imposed.