Foreshock

WHITE PAPER · TRUST & SAFETY

The signal arrives before the shock.

A purely semantic Trust & Safety and organizational-insight engine that reads the geometry of language, never its content, and shows you the math instead of hiding it.

DOCUMENT
Platform Overview v0.1

BUILT ON
Aldous · DECE

CLASSIFICATION
Zero-shot, no LLM reasoning

CONTACT
Tim Post · tim@foreshock.io

ABSTRACT

Foreshock measures the human emotional and intent spectrum and acts as a tunable semantic firewall that communities can train transparently on commodity laptops. It is a Nearest Centroid Classifier with diagonal covariance estimation: it uses only the geometry embedded in everyday language, returns numeric scores, and applies standard, well-documented math and physics models on top. No string matching, no regular expressions, no LLM reasoning, no opaque heuristics.

00 CONTENTS

01 Privacy-respecting geometric classification

02 What Foreshock can see

03 Two kinds of Third Rail

04 Monitoring and report investigation

05 Receipts for adverse actions

06 In an LLM governance pipeline

07 Organizational health monitoring

08 Pricing tiers

09 What comes next

10 About Foreshock

01 HOW IT WORKS

Privacy-respecting, zero-shot geometric classification

Foreshock does not use or rely on LLM inference or output. It uses only the geometry that is already embedded in our everyday language.

An embedding model, which is technically a large language model, supplies that geometry. Embedders produce only a sequence of numbers, called vectors, that describe the relationships between the words, or tokens, in a phrase. That is the full extent of LLM involvement in Foreshock's classification. The text itself is never reasoned about, summarized, or stored.

From there, Foreshock applies standard, accepted, well-documented math and physics models to the vectors that Aldous produces. What is unique is not the models themselves, but how Foreshock transforms emotional data into inputs those models can consume without modification.

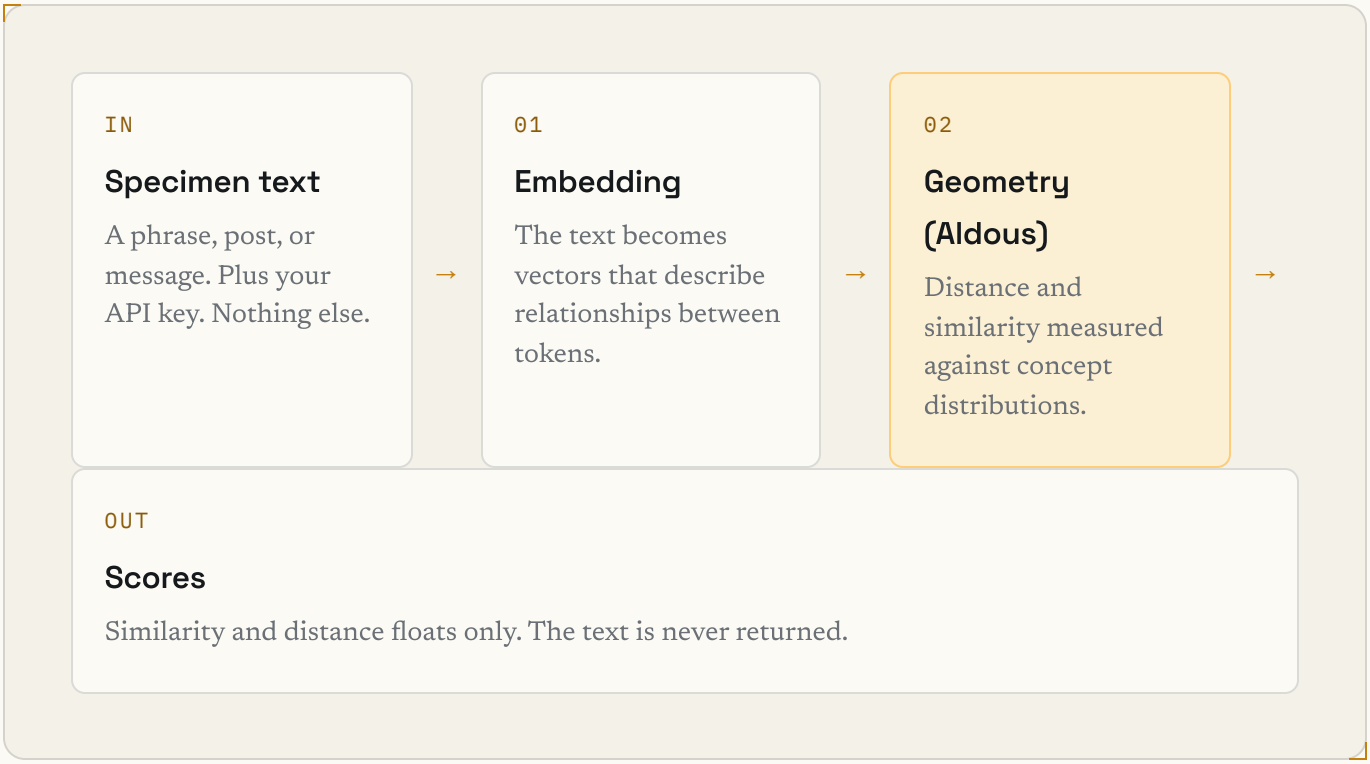

Fig. 1 — The classification path. An LLM touches the data once, to embed it. Everything downstream is ordinary geometry on numbers.

> *We put the math out front and want you to understand it. Competing products try to hide secret sauce and keep you locked into opaque systems.*
>
> FORESHOCK DESIGN PRINCIPLE

Foreshock is built on Aldous, a semantic model our engineers designed from scratch and open sourced, so that every fledgling community can have access to Trust & Safety tooling that does not require computationally prohibitive inference at the cost of privacy and autonomy over decision making.

02 CAPABILITIES

Understand user-generated content effectively and intelligently

Many platforms cannot deliver these insights because they lean on LLMs to guess, or on string-matching rules that chronically fall out of date. Because Foreshock reads geometry rather than keywords, it can do several unusual things at once.

- **Irrational hate, by intensity.** Detect the presence and the strength of many distinct kinds of irrational hate, not just a binary flag.
- **Graduated emotion.** Read emotions on a tiered scale by intensity: joy, +joy, ++joy, sadness, anger, hedging, optimism, partisanism (left / center / right), philosophical tension, existential angst, change anxiety, fear, violence, conflict, sexuality, and more.
- **Synthetic origin.** Estimate the approximate probability that content is machine generated, including stitched or partially synthetic material.
- **Latent Content Erasure.** Use LCE to ask whether a sample has any intrinsic value once vectors matching a Trust & Safety third-rail are removed, then automatically re-examine what is left.

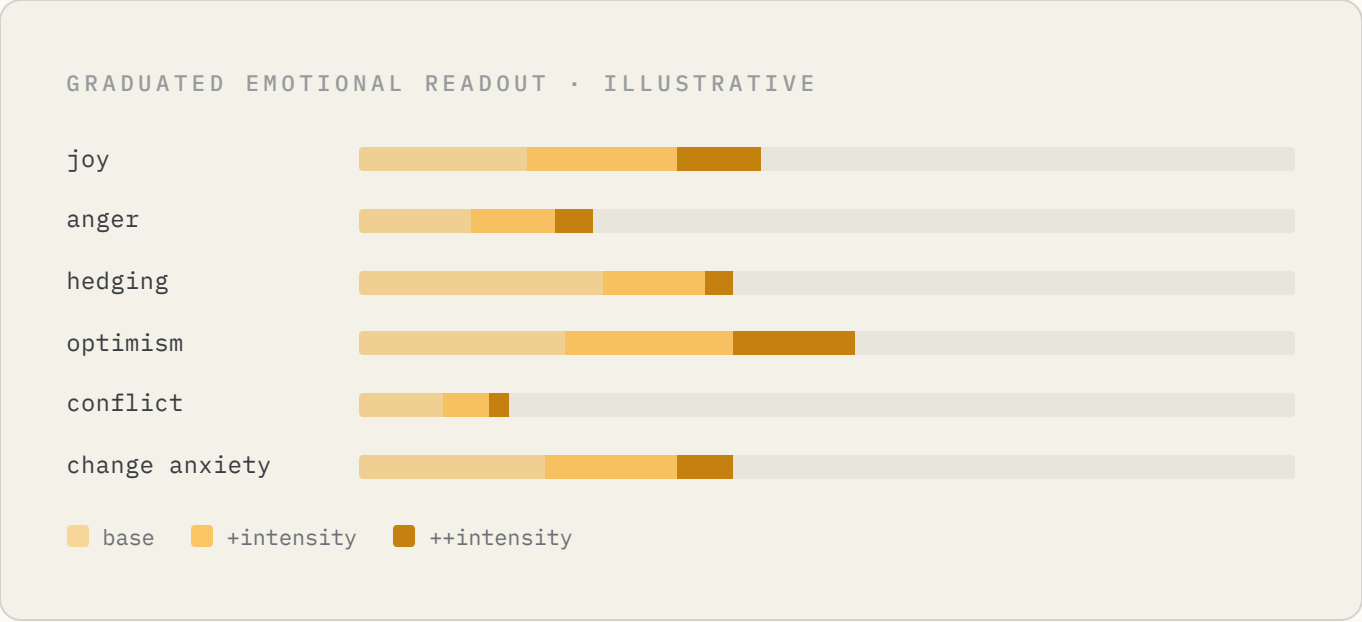

Fig. 2 — Each dimension reports a graduated intensity rather than a single yes or no, so tone is preserved instead of flattened.

03 THIRD RAILS

Two kinds of circuit breaker, each with its own strengths

Foreshock offers two families of Trust & Safety third-rail, called shunts. They sit at opposite ends of a tradeoff between precision and speed of update, and most teams use both.

TYPE A · INTRINSIC

Centroid-matched, differentially weighted

- High accuracy in matching and discernment.
- Requires tuning and partial re-training to adjust.
- Ideal for concepts that do not move much linguistically.
- Built from complete phrases that carry most of what a sentence needs.
- 3 to 5 phrases encapsulate a concept; 18 to 20 approach comprehensive or over-reaching.
- Each phrase adds roughly 30 seconds of training time.

TYPE B · MONOLITHIC

Mean-pooled, updatable every second

- One fast embedding pass to adjust, so updates can be near-continuous.
- Ideal for onslaughts of bad actors who keep changing tactics.
- A mixed bag of phrases, sentences, and casts that teams need to catch.
- Lower magnitude means lower resolution for differentiation.
- Informational, but should not drive automation on presence alone.
- No centroid set, so LCE verification is not available.

Latent Content Erasure (LCE)

Intrinsic shunts can erase problematic vectors from a specimen and re-score what remains. This gives an accurate view of the content that is left after whatever tripped the system has been removed. If the remainder is just noise, it can be rejected with confidence.

It is never possible to reverse vectors back into text, but it is possible to pull vectors out of other vectors. Because these shunts use Standardized Euclidean Distance, which requires the centroid set from the actual phrase distribution rather than straight-line shortcuts, Latent Content Erasure is very achievable.

RECEIPTS & TRANSPARENCY

Intrinsic shunts can return two heat maps in a receipt: the specimen with the problematic vectors, and the specimen without them. We recommend showing the "with" map to anonymous users and both to trusted users. Transparency instills trust in the system.

A note on monolithic resolution

Because monolithic shunts are not individually embedded, they behave as token clusters and operate at a lower magnitude. Expect some false positives, and pair them with corroborating signal rather than wiring them straight to automated action. For longer samples, Foreshock provides a heat map of which sections scored highest, which you can use to estimate the ratio of problematic to benign content. In short samples this is simply not practical.

04 CONSOLE

Comprehensive, web-based monitoring and report investigation

Foreshock gives you one place to customize your model's behavior and to examine logs of past classifications with clarity. Every value is visualized and narrated as you scroll, and exporting a detailed result as PDF or JSON takes a single click.

Shunt editor | Sensor & model tuning | One-shot visualizer classifier | PDF / JSON export | Sensor allocation | Multi-org management

The console hosts an intuitive editor for your shunts and, depending on your plan, controls to tune any part of your model. You can run the one-shot classifier directly through the visualizer interface. Graduated-tier users manage additional organizations and sensor allocation from the same place.

05 RECEIPTS

Detailed notices for adverse actions

If you cannot explain and defend an adverse action against a user acting in good faith, you will not retain many users. Receipts signal responsibility and earnestness in moderation, usually early on, and show that you have invested in a system that cares about people by default. Modern LLM-driven safety systems rarely provide what a responsible notice requires. Foreshock issues four things at minimum.

01 **Why was the action taken?**
A list of scores showing how the content tied to your rules.

02 **What can they do differently next time?**
Because Foreshock is semantic, there is no fear of leaking guarded keywords. You can show the heat-map location in the sample and how it scored.

03 **Where can they appeal?**
A link to dispute the decision, which you provide.

04 **Where can they report possible bias?**
A link to flag bias, which Foreshock provides and follows up on.

Holding receipts back out of fear that they help bad actors improve is no longer necessary. That fix is baked into Aldous by design, and it resolves a uniquely frustrating problem for everyone involved.

06 GOVERNANCE

Helping LLMs understand emotional impact and user intent

Foreshock adds value at several points in an LLM governance pipeline. Two uses tend to matter most for organizations of any size.

1 · Pre-score input and attach the result to context

At minimum this helps a model respond with empathy. In many cases the shunts can detect turn-based attacks as concepts begin to form. Whispering that signal to a model, in a form it is prompted or trained to understand, helps it grasp sarcasm, or recognize when a prompt was written in reaction to someone's autonomy being infringed.

2 · Preview the impact for the model

Foreshock has sensors for human sycophancy, hedging, anger, AI sycophancy, and more, all of which a model might generate from training but rephrase given the right insight.

AS AN MCP SERVER

Wire Foreshock in as an MCP server and you can set a hard preference such as "I never want answers with a sycophancy score over 40," then enjoy conversations with your model again.

07 ORGANIZATIONAL HEALTH

Long-term emotional health and stability monitoring

By piping many samplings of daily work chatter through the model, you gain a real understanding of the emotions running through your organization. Ordinary methods can be blind, or biased, to underlying tension, fear, unconscious bias, gloom, hedging, and partisanism. Foreshock surfaces what is there, beneath the surface, in a blameless and actionable way.

Our approach models organizations like Magnetorheological (MR) fluids, which lets standard rheological models monitor health in terms of laminar flow, surfactant persistence, pressure, and other critical metrics. That translates a wall of emotional numbers into real behaviors. We also use damped and driven double-harmonic oscillation models for earlier detection of phase transitions.

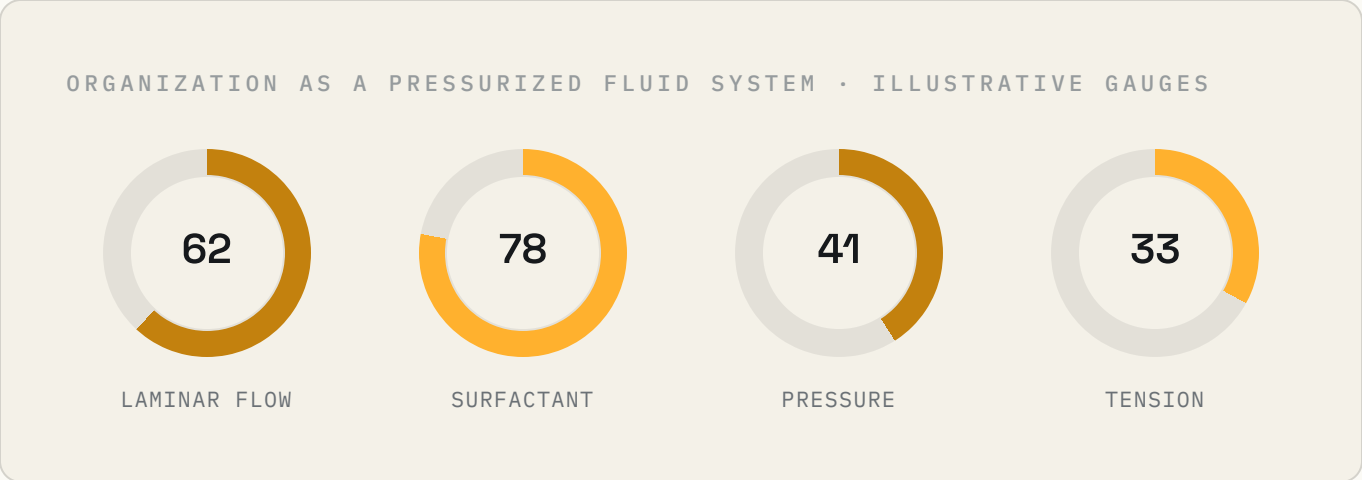

Fig. 3 — Put simply: we visualize your organizational health like a pressurized fluid system, with the gauges and most of the physics that entails.

PRIVACY BY CONSTRUCTION

At no point is generated or contributed content meaningfully comprehended by intelligence. Only numeric scores are returned. Text is never fed to search engines or third parties, only to non-generative models under our direct control. Foreshock collects no PII and no user data. It needs only the specimen text and your API key.

08 PRICING

Tiers that let new communities start safely

New communities need to start safely, and Foreshock wants to be the thing that keeps them from becoming victims of their own success. There will always be a free tier with enough resources for a small population of 100 to 250 members.

TIER	PRICE	INCLUDES MONITORING FOR
Volcanic	`$0.00`	• 1 chat channel, 1 social feed, 1 RSS feed, 1 GitHub Community • 150 monthly direct API classifications • Aldous open source supported, no support SLA • 14-day local data and journal retention, free exports
Tectonic	`$10 / mo` `$99 / yr`	• 5 chat channels (+$2/mo$ each), $5 RSS feeds$ (+1/mo each) • 2 social feeds (+$2/mo$ each), $3 GitHub Communities$ (+5/mo each) • Open source or customized models • 500 direct API requests every 72 hours, rolling • Shunt Studio, near-instant webhooks • 45-day retention + BYO Postgres, priority support
Seismologist	`$25 / mo` `$249 / yr`	• Everything in Tectonic, plus multiple organizations • Shared quota across orgs: 25 feeds, 20 chats, 10 GH repos, 10 socials • 1,500 direct API requests every 72 hours, rolling • Shunt, Sensor & Tuning Studio, open or custom models • SSH & console access to Splinter tools, priority support
Integrator	`free if approved`	• For those integrating Foreshock into their own product or process • No confidentiality agreements or lengthy contracts, we work directly with your team • Approval based on research & growth potential and complexity • Projects with GPU spend require a payment method on file

On-prem options are available. Contact us to discuss your location and sensor needs.

09 ROADMAP

Next: a culture-aware Lab Assistant

Aldous's embedding model already understands many languages, several dozen of them comprehensively. What it does not yet understand is how different cultures use the same words to mean very different things. The geometry of emotion across English-speaking and Spanish-speaking regions differs in many ways.

We may produce a model based on Qwen 2.5-7B-Instruct, highly tuned with Weight-Decomposed Low-Rank Adaptation (DoRA). Trained on Foreshock's methodology, it would be deployed for evaluation escalation and for narrating trends in continuous streams. Though technically an instruct model, it will not be very conversational, just very culture-aware.

7B	DoRA	1 GPU	0 telemetry
LAB ASSISTANT BASE · QWEN 2.5-INSTRUCT	WEIGHT-DECOMPOSED LOW-RANK ADAPTATION	SCALES TO ORG NEEDS ON A SINGLE GPU	PRIVACY-RESPECTING, ZERO-TELEMETRY MANDATE

A confidence head with β-NLL loss monitoring will guard against overconfident generation and make gaps in training easy to find through adversarial prompt testing. We will also introduce loss-priors, both real (the dangers and financial harm of misclassification, domain specific) and synthetic (only as needed), to further guard against overconfidence.

This becomes the base for a Partner Model intended for vendors who want to add our technology to their processes. Because DoRA is not plagued by the forgetting that appears when a topic lacks magnitude in the original weights, the adapter is never entered for concepts it should not touch, and an existing rich physics and math corpus removes much of the risk.

Multi-culture support is not a herculean task, but its transliteration complexity should not be understated. One-to-one translation of concept phrases fails to capture how a given Spanish-speaking culture expresses an emotion with the proper dimensionality. We are designing the studio with eventual cultural transliteration in mind as a next major release candidate.

10 ABOUT

The company behind Aldous and Splinter

Foreshock reinvests its profits in growth through research. Democratizing models that serve Trust & Safety and governance fully transparently, and that train rapidly on commodity hardware, is our best path to sustained and responsible growth.

> *Communities need auditable armor on the outside and a deeper understanding of what is happening on the inside, without giving up privacy or accepting arbitrary, nebulous decisions.*
>
> FORESHOCK MISSION

That is the mission: to sustain ourselves so we can afford to work on the parts of AI governance and Trust & Safety that nobody else wants to work on.

Foreshock

Contact: Tim Post · tim@foreshock.io