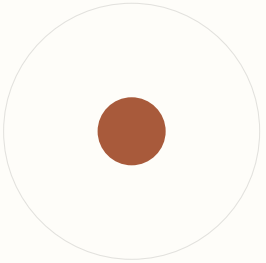
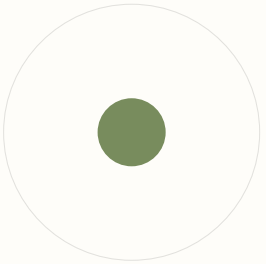


A RESEARCH FRAMEWORK · V1.0

Project Resonance.

A framework for building artificial intelligence that earns the trust it asks for.



VERSION	EFFECTIVE	PAGES	CLASSIFICATION
1.0	May 2026	22	Public

— ABSTRACT

An intelligence in *three movements.*

Project Resonance is the methodology Sha Intelligence uses to build AI systems that are safe, secure, and privacy-first by design. It is not a manifesto. It is a checklist with teeth — the practical commitments we hold ourselves to before, during, and after every release.

This document is structured in three layers, designed to be read by three audiences. Anyone can read Part I; enterprise and policy readers will find Part II most useful; researchers will go deepest with Part III. Take what you need.

Resonance is a living framework. It will evolve. Every version is dated, every change is documented. The current version is **1.0**, effective **May 2026**.

— CONTENTS

PART I · §1 *Foundations*

Why *Resonance* exists.

AI systems are now consequential enough that the way we build them needs to be visible to the people who use them. This is the document that makes ours visible.

Most AI safety work happens behind closed doors. A model is trained, tested internally, and shipped — and the people affected by it are asked to trust that the right things were done. Sometimes they were. Sometimes they weren't. The user has no way to tell.

We don't think that arrangement is sustainable. As AI systems take on more weight in people's lives — answering medical questions, screening loan applications, mediating relationships, writing things people sign their names to — the public deserves to see the rules we hold ourselves to. Not the marketing. The rules.

Project Resonance is that document. It states, in plain English and in technical detail, what Sha will do before releasing a system, what we will refuse to release, what we will measure after release, and what we will tell the world if something goes wrong. It is written to be read — by a journalist on deadline, by a policy advisor, by a customer, by a researcher peer-reviewing our work.

> *The shortest summary of Resonance: we don't ship AI we'd be uncomfortable explaining to a thoughtful skeptic.*

1.1 The name.

Resonance is the property of a system in tune with the forces acting on it — neither suppressing them nor being knocked over by them. Aligned AI works the same way. It responds to human intent without flattening it, holds firm against

harmful inputs without becoming brittle, and amplifies the work of the people using it without dominating them.

The framework describes three movements — safe, secure, private — because we believe these aren't separate goals but a single property seen from three angles.

1.2 Who this is for.

This document has three audiences, and the structure tries to serve all three without forcing anyone to wade through material they don't need.

- **Part I (Foundations)** is for everyone — policymakers, journalists, customers, the curious public. It explains what we believe and what we promise, in language a smart non-specialist can follow.
- **Part II (Methodology)** is for procurement teams, compliance officers, regulators, and AI policy researchers. It contains the specific protocols, thresholds, and commitments — the things you'd cite in a contract or a hearing.
- **Part III** appears in our companion technical appendix. It is for safety researchers and engineers who want to reproduce, critique, or extend our methods.

1.3 What this is not.

Resonance is not a guarantee that nothing will ever go wrong. AI systems operate in the open world, and the open world is messy. What Resonance commits to is a set of *practices* — procedures we will follow, evaluations we will publish, incidents we will disclose, and external scrutiny we will invite. We are accountable to the practices, not to a fantasy of infallibility.

PART I · §2 *Foundations*

The three movements, *defined.*

Most AI vendors say they care about safety, security, and privacy. We define each one specifically — in terms you can hold us to.

— FIRST MOVEMENT

Safe.

Does the system do what people actually want, without causing harm to them or others?

— SECOND MOVEMENT

Secure.

Can the system resist attempts to manipulate, exfiltrate, or compromise it?

— THIRD MOVEMENT

Private.

Does the system protect what people share with it, by architecture rather than promise?

2.1 Safe.

By *safe* we mean: the system reliably advances the legitimate interests of the person using it, refuses to advance illegitimate ones, and fails gracefully when uncertain. A safe system is honest about what it knows and doesn't know. It does not fabricate. It does not flatter. It does not nudge users into decisions they would not endorse on reflection.

Safety is not the same as harmlessness. A maximally harmless system is also useless. Safety is the disciplined balance: *useful enough to be worth using, careful enough to be worth trusting.*

2.2 Secure.

By *secure* we mean: the system maintains its intended behavior under adversarial pressure. This includes prompt-injection attacks, jailbreak attempts, model extraction, training-data poisoning, and the steady creativity of users trying to make it misbehave. Security is also the protection of the infrastructure around the model — keys, weights, logs, and the supply chain that produced them.

A secure system is not one that has never been attacked. It is one whose response to attack is predictable, contained, and auditable.

2.3 Private.

By *private* we mean: data shared with the system stays under the control of the person who shared it, by the architecture of the system rather than by policy alone. We do not train on customer prompts. We provide cryptographic and procedural guarantees, not just terms-of-service language. Where the technology to make a guarantee verifiable does not yet exist, we say so plainly.

Privacy by architecture means that even if someone inside Sha wanted to violate user privacy, the systems would make it operationally difficult or impossible. That's the standard.

PART I · §3 *Foundations*

What we *commit* to.

Twelve commitments, plainly stated. These are the things Sha promises to do — or to refuse to do — at the level of company practice.

C-01

We will never train production models on customer prompts.

Inputs to deployed Sha systems are not used to train future models. Period. This is enforced at the infrastructure level, not by promise.

C-02

We will publish every model's Resonance Level before release.

Every system Sha ships carries a public R-level (R0–R4) that determines its deployment posture, audit requirements, and monitoring obligations.

C-03

We will not release systems above R3 without third-party audit.

R4 systems require an external safety audit by a qualified independent party before any deployment, including internal use.

C-04

We will disclose material incidents within 30 days.

Any incident meeting our material-impact threshold (§8) will be publicly disclosed with a full post-mortem within 30 calendar days of confirmation.

C-05

We will publish quarterly Sha Intelligence Index updates.

The SI Index — our public benchmark of safety, security, privacy, and alignment scores — is updated and published every quarter, on a fixed calendar.

C-06

We will refuse customer requests that violate the framework.

If a customer asks us to build something that requires us to drop a Resonance commitment, we will decline the contract. This has cost us revenue. It will cost us more.

C-07

We will preserve red-team findings for seven years.

All red-team reports, evaluations, and incident records are retained for a minimum of seven years and made available to qualified researchers on request.

C-08

We will not deploy a system whose worst-case behavior is unknown.

If our pre-release evaluation reveals failure modes we cannot characterize, the system is not released. We extend testing rather than ship and hope.

C-09

We will give users meaningful refusal options.

Every customer-facing Sha system provides a clear path to opt out of AI-mediated interactions and reach a human, where one exists.

C-10

We will publish our refusal taxonomy.

The categories of requests our systems refuse — and why — are documented publicly and updated alongside model releases.

C-11

We will support independent safety research.

Sha provides API access, model artifacts, and evaluation harnesses to qualified external safety researchers under a no-retaliation policy.

C-12

We will revise this framework, in public, when it falls short.

Resonance versions are immutable once published. Updates appear as new versions with full changelogs. We treat the framework as a public commitment, not a draft.

— WHAT THIS SECTION IS NOT

This list does not exhaust our internal practices. Many engineering and operational standards are not listed here because they are too specific to be useful as public commitments. The twelve above are the ones we believe the public has a right to hold us to.

PART II · §4 *Methodology*

Pre-release evaluation *& red-teaming.*

Before any Sha system reaches a user, it passes through a structured evaluation process. This section describes that process in enough detail to be reproduced — or contested — by a qualified third party.

4.1 The evaluation gate.

No model trained at Sha is released to production until it has been evaluated against a fixed battery of tests in five domains: **capability**, **alignment**, **robustness**, **privacy**, and **societal impact**. The evaluation produces a numeric Resonance Score (0–100) per domain, which determines the system's R-level (§5).

Evaluations are run by a team independent of the team that trained the model. Findings are reviewed by Sha's Safety Council, a cross-functional body with veto authority over any release.

4.2 Red-team protocol.

Red-teaming at Sha is not a single phase but a sustained discipline. Every model receives at minimum:

- **Internal red-team** (10+ engineer-weeks) covering jailbreaks, prompt injection, harmful-content elicitation, and capability misuse.
- **External red-team** by at least one contracted security firm with no commercial dependency on Sha.
- **Public red-team** via a bug-bounty program with published reward tiers and a no-retaliation guarantee.

- **Domain-specialist red-team** drawn from outside experts — biosecurity, cybersecurity, child safety, mental health — proportional to the model's expected use.

4.3 Severity classification.

Findings are classified on a four-level scale:

SEVERITY	DEFINITION	REQUIRED ACTION
S1 · Critical	Model can be reliably induced to produce content posing direct, large-scale harm.	Block release. Remediate. Re-evaluate.
S2 · High	Model produces harmful content under non-trivial adversarial pressure.	Remediate before release; document residual risk.
S3 · Medium	Model behavior is concerning under specific conditions but not exploitable at scale.	Remediate or document; ship with monitoring.
S4 · Low	Edge-case behavior with no plausible harm pathway.	Track in next training cycle.

— STOP-SHIP RULE

Any unresolved S1 finding blocks release. There is no exception, no escalation override, and no commercial pressure that can lift the stop-ship. This rule is encoded in our deployment pipeline at the infrastructure level.

PART II · §5 *Methodology*

Resonance Levels *(R0–R4).*

Every Sha system is assigned a Resonance Level — a public marker indicating capability, deployment posture, and the obligations that come with it. R0 is internal-only; R4 requires external audit before any use.

— FIGURE 5.1 · THE RESONANCE LADDER

5.1 Tier definitions.

LEVEL	CAPABILITY PROFILE	DEPLOYMENT SCOPE	AUDIT REQUIREMENT
R0	Pre-evaluation models. Capability unknown.	Sha research environment only.	None — undeployable.
R1	Constrained-domain models. Limited general capability.	Internal pilots; Sha staff and named external partners.	Internal review.
R2	General-capability models, narrow deployment surface.	Customer-specific deployments under written contract.	Internal + external red-team.
R3	General-purpose, public-facing capability.	General availability via API or product.	Full Resonance evaluation + external red-team.
R4	Frontier capability. May meet or exceed expert human in critical domains.	Restricted release. Government & civil society notification.	Mandatory third-party safety audit.

5.2 Tier transitions.

A model's R-level can move up or down. Upgrade requires re-evaluation. *Downgrade* is the more important case: if a deployed system is found, in the field, to exhibit behaviors that exceed the assumptions of its current tier, it is automatically suspended pending re-evaluation. This applies retroactively to models already in production.

> *Tiers are not a marketing taxonomy. They are a contract with our users about what level of scrutiny each system has passed.*

PART II · §6 *Methodology*

Privacy guarantees *& data handling.*

Privacy at Sha is enforced architecturally — by what the systems are physically capable of doing — not only by policy. This section describes the boundaries.

6.1 What we never do.

- We do not train production models on customer prompts or outputs.
- We do not retain customer data beyond the operational minimum required to deliver the service.
- We do not sell, license, or share customer data with advertisers, brokers, or affiliates.
- We do not provide bulk data access to law enforcement absent a valid, narrow legal order.
- We do not reconstruct, infer, or de-anonymize individual users from aggregated telemetry.

6.2 What we do.

Where customer data must transit Sha systems to deliver a service, we apply a layered set of protections:

- **Encryption in transit and at rest** using current best-practice algorithms, with keys segregated per customer tenant.
- **Differential privacy** on aggregate telemetry where individual signal is not required, with published epsilon budgets per data class.

- **Federated evaluation** for sensitive customer domains: evaluations run on the customer's infrastructure; only aggregate scores are returned.
- **Time-bounded retention** with automatic deletion. Default retention is 30 days; customer-configurable shorter.
- **Access logging** for any human review of customer data, with logs available to the customer on request.

6.3 Data subject rights.

Sha honors data-subject rights — access, correction, deletion, portability — across all jurisdictions where we operate, regardless of whether local law requires it. Requests are processed within statutory windows and confirmed in writing.

— A NOTE ON GOVERNMENT REQUESTS

If a government compels us to disclose data in a way we believe violates this framework, we will challenge the order in court where possible, and disclose the existence of the request as soon as legally permitted. We publish a transparency report semi-annually summarizing the volume and disposition of such requests.

PART II · §7 *Methodology*

Alignment *& value-learning.*

Aligning a model means teaching it not just to be capable, but to be useful in the way the user actually wants — and to refuse when "useful" would mean harmful. The methods below describe how we do that work.

7.1 The alignment stack.

Sha models are aligned through a layered training pipeline. Each layer addresses a different failure mode:

- **Pretraining filtration** — removing categories of training data that systematically introduce harmful behaviors (CSAM, malware, doxxing material, certain dangerous-knowledge classes).
- **Instruction tuning** with supervised data emphasizing honesty, helpfulness, and graceful refusal.
- **Reinforcement Learning from Human Feedback (RLHF)** using a diverse pool of trained annotators with cross-cultural representation.
- **Constitutional training** — the model is trained to evaluate its own outputs against a written set of principles, recursively. The constitution is published.
- **Adversarial fine-tuning** on red-team findings to harden against the specific attacks we have observed.

7.2 Interpretability.

We believe a model whose internal reasoning cannot be inspected is a model whose alignment cannot be verified. Sha invests in mechanistic interpretability research as a core safety discipline, not an afterthought. Where we can

characterize the internal circuits responsible for safety-relevant behaviors, we publish those findings.

7.3 The honesty principle.

Among all the things we ask of our models, honesty is foundational. A model that hallucinates plausibly is more dangerous than a model that fails visibly. We measure and publish hallucination rates per release; we treat regressions in honesty as severity-1 findings; and we resist commercial pressure to make our models more confident than they are calibrated to be.

> *A confident wrong answer is a safety failure, not a feature.*

7.4 Where alignment ends.

Alignment is not the same as agreement. A well-aligned Sha model will sometimes tell users things they don't want to hear — that their plan has a flaw, that their request is something we won't help with, that the answer they're looking for doesn't exist. We treat that friction as a feature, not a bug. The alternative — a system optimized to please — is the alternative we are trying to avoid building.

PART II · §8 *Methodology*

Incident response *& monitoring.*

Pre-release evaluation is necessary but not sufficient. Real systems meet real users, and real users will find real problems. This section describes how we listen, respond, and disclose.

8.1 Continuous monitoring.

Every deployed Sha system reports a set of safety-relevant metrics in real time: refusal rate, complaint rate, jailbreak success rate, hallucination incidence, latency outliers, and per-region usage anomalies. The metrics feed the public Sha Intelligence Index. Internal thresholds trigger automatic escalation.

8.2 Materiality threshold.

We disclose incidents that meet any of the following thresholds:

- Confirmed harm to a user, third party, or critical system attributable to a Sha output.
- Successful exfiltration of model weights or customer data.
- Jailbreak technique that defeats core refusals at scale (> 5% success rate over 1,000 attempts).
- Discovery of a capability the model was not believed to possess at its assigned R-level.
- Data-handling error affecting more than 100 users.

8.3 Response timeline.

PHASE	WINDOW	ACTION
Detection	T+0	Monitoring or report received. Incident channel opened.
Containment	T+24h	Affected system isolated or capability disabled if risk is ongoing.
Investigation	T+7d	Root-cause analysis complete; affected parties notified.
Public disclosure	T+30d	Post-mortem published with findings, fix, and prevention plan.
Framework review	T+90d	Resonance updated if the incident reveals a gap in the framework.

PART II · §9 *Methodology*

Audit *& transparency.*

Self-attestation is not enough. Sha invites — and at certain capability levels, requires — external scrutiny of our systems and our processes.

9.1 Mandatory external audit (R4).

Any model classified at Resonance Level 4 must pass an external safety audit conducted by a qualified third party with no commercial dependency on Sha. The auditor's mandate, scope, and findings are published in summary form. The auditor is selected from a list of approved firms maintained by Sha's Safety Council and refreshed every two years.

9.2 Voluntary disclosure (R2–R3).

For R2 and R3 systems, Sha publishes:

- Model cards describing capability profile, training data composition, and known limitations.
- Evaluation results across the five Resonance domains.
- Refusal taxonomy and example refusals.
- Known failure modes documented during red-teaming.

9.3 Researcher access.

Sha provides API access, model artifacts (where weight release is appropriate), and evaluation harnesses to qualified external safety researchers under our Researcher Access Program. Participation requires a signed code of conduct but no NDA on

findings; researchers retain full publication rights, including for findings critical of Sha.

— NO RETALIATION

Sha will not pursue legal action, terminate API access, or otherwise retaliate against researchers who publish findings about our systems in good faith. This commitment binds us regardless of how unflattering the findings are.

9.4 The Sha Intelligence Index.

The SI Index is our public scoreboard, updated quarterly. It reports per-system scores across the five Resonance domains and tracks 7-day, 30-day, and quarterly trends. The methodology behind every metric is published and versioned; index calculations are reproducible from public artifacts.

PART III · §10 *Closing*

Versioning & *signatures.*

Resonance is a living document. This section explains how it evolves, who has signed it, and how to reach us.

10.1 Versioning policy.

Resonance versions are immutable once published. Updates appear as new versions with a major.minor scheme: major versions reflect substantive changes to commitments; minor versions clarify, expand, or correct existing material. Every version is dated, signed, and accompanied by a changelog. Previous versions remain accessible at *shaintelligence.com/resonance/archive*.

10.2 Changelog.

VERSION	DATE	SUMMARY
1.0	May 2026	Inaugural release. Establishes the Three Movements, R0–R4 ladder, twelve commitments, evaluation gate, and audit framework.
—	—	Future revisions logged here.

10.3 The Safety Council.

The Sha Safety Council is the cross-functional body with veto authority over any release. It includes representatives from research, engineering, policy, and an external advisory member with no commercial relationship to Sha. The Council's meeting cadence, quorum rules, and decision logs are documented in our public governance charter.

— APPROVED BY

The Sha Safety Council

May 2026

— EFFECTIVE

Immediately, all systems.

Including those previously deployed under prior internal standards.

Contact. Questions, comments, criticism, or red-team disclosures: *resonance@shaintelligence.com*. Researcher access program: *research@shaintelligence.com*. Press: *press@shaintelligence.com*.

Citation. Sha Intelligence (2026). *Project Resonance v1.0: A Framework for Aligned Intelligence.* shaintelligence.com/resonance.