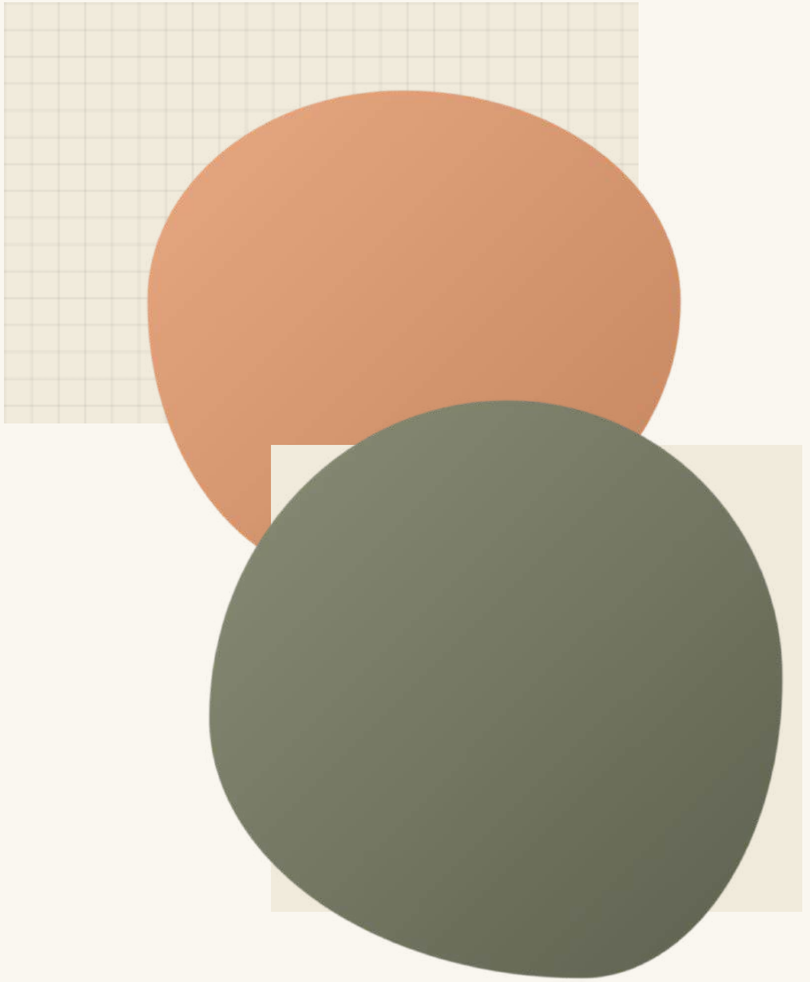
STRATEGIC ANALYSIS · APRIL 2026

Commercializing the black box

A GTM strategy for Goodfire AI

Strategic analysis, market research, product strategy, and commercial intelligence tools — accompanied by *Forge*, a working commercial intelligence prototype.



EXECUTIVE SUMMARY

The only well-funded company commercializing mechanistic interpretability — backed by Anthropic's first-ever startup investment, with founders from DeepMind and OpenAI, and published results no one else has produced.

$1.25B

POST-MONEY VALUATION, FEB 2026

B Capital led $150M Series B; $209M total raised

58%

HALLUCINATION REDUCTION, RLFR

Gemma-3-12B-IT on LongFact++. ~90× cheaper than Gemini 2.5 Pro as judge.

96%

F1 ON RAKUTEN PII DETECTION

vs 51% LLM-as-judge baseline. 15–500× cost savings. In production.

FIVE STRATEGIC INSIGHTS

01 **The category is not better explainability.**
Goodfire isn't competing with Fiddler/Arize/Galileo — they monitor outputs; Goodfire understands internals. The frame is *the scientific instrument that performs the surgery*, not XAI.

02 **Three revenue engines, not five GTM motions.**
Enterprise platform (~40%), runtime monitoring at Rakuten shape (~30%, the scale engine), scientific discovery (~15%), 1–2 frontier-lab design partnerships (~15% plus strategic research).

03 **Every deal operates across five levels.**
Buyer psychology, procurement, market category, evidentiary regulatory demand, civilizational consensus.

04 **Six subproblems on the critical path.**
Category void, translation chasm, packaging gap, proof-of-value, trust infrastructure, scalability — solved in sequence, not in parallel.

05 **The open-weight explosion is a tailwind, not the thesis.**
Enterprise open-weight deployment grew sharply 2025–2026, but the durable wedge is published research (Prima Mente, Evo 2, Rakuten) — not the substrate model.

THE STRUCTURAL THESIS

The $1.25B Series B was priced as a neolab — talent, scarcity, and option value, not revenue. The work between now and Series C is converting that optionality into **$50M ARR by October 2027 (stretch $75M)**, on the back of three engines and one landmark frontier-lab partnership.

THE FORGE PROTOTYPE

A working commercial intelligence OS that operationalizes this analysis.

Multi-level deal briefs, three-voice proposals, pipeline diagnostics, and competitive differentiation responses.
Live at `forge-lemon-beta.vercel.app`.

CONTENTS

Seven sections.
One throughline: from research lab to repeatable commercial engine.

SECTION 01 — 04 SLIDES

Market landscape & competitive analysis.

$11B today, $25B by 2030. Nine players across three categories. One fundamental divide: post-hoc explainability versus mechanistic understanding.

SECTION 01.01 — MARKET SIZE & GROWTH

The adjacent XAI market is a proxy, not our category. The real TAM is model-risk and governance spend at regulated AI deployers.

$11B	$25B	20%	80%+
XAI MARKET, 2025	XAI MARKET, 2030E	COMPOUND ANNUAL GROWTH	ENTERPRISES ON GENAI BY '26

THE NUMBERS

XAI TAM (adjacent, not ours): ~$9.5–11.3B in 2025, 18–21% CAGR to $21–25B by 2030 (Grand View, Mordor, Precedence, Fortune Business Insights). This is the budget Goodfire absorbs and expands — not the category we compete in.

GOODFIRE'S ACTUAL TAM

Bottom-up: ~200 named enterprise accounts × $1–3M average ACV = **$300–600M mature SAM in Year 3**. Plus the runtime monitoring overlay on the $8.19B enterprise LLM market (27.45% CAGR). Plus pharma/scientific discovery R&D adjacency — an order of magnitude above MRM budgets. Goodfire is sizing against MRM + AI governance + scientific R&D, not post-hoc explainability.

SECTION 01.02 — REGULATORY DRIVERS

Five regulatory regimes converging in 2026. Each creates evidentiary demand for mechanistic analysis — not a mandate, but an accelerant.

REGULATION	JURISDICTION	KEY REQUIREMENT	ENFORCEMENT
EU AI Act, Art. 13	EU	High-risk systems must be transparent enough for deployers to interpret outputs	AUG 2, 2026 [1] €15M or 3% global revenue
SR 11-7	US banking	Model risk management: independent validation, explainability, ongoing monitoring	Active; limited AI-specific guidance
FDA AI guidance	US healthcare	Algorithmic decision documentation, clinically relevant explanations	1,250+ AI devices authorized; expectations increasing
NIST AI RMF	US voluntary	Risk management framework including transparency and explainability	Voluntary but increasingly referenced in procurement
State laws (CO, CA)	US states	CO SB 205 (Jun '26), CA SB 53 (Jan '26): AI transparency mandates	1,000+ AI bills across state legislatures since Jan '25

IMPLICATION

Article 13 says *"transparent enough"* — the interpretation is undefined. Whichever methodology regulators cite first becomes the de facto standard. This is a standards-setting window, not a procurement window.

[1] Digital Omnibus (Nov 2025) and March 2026 Council/Parliament negotiating positions proposed slipping Annex III high-risk obligations to Dec 2, 2027 and Annex I product-embedded systems to Aug 2, 2028. Prudent GC advice still treats Aug 2026 as binding pending final text.

SECTION 01.03 — COMPETITIVE LANDSCAPE

Nine players, three categories. The fundamental divide: checking the dashboard vs. understanding the engine.

INTERNAL LAB TEAMS

Research for their own models. Anthropic invested in Goodfire instead of commercializing.

ENTERPRISE XAI

Monitor inputs and outputs. Cannot see inside models. Post-hoc explanations.

MECHANISTIC INTERP

Comprehend internal computations. Goodfire dominates this category.

COMPANY	CATEGORY	FUNDING (APR 2026)	GOODFIRE DIFFERENTIATION
Anthropic (interp)	Internal lab	n/a	Invested in Goodfire rather than commercializing externally
Google DeepMind	Internal lab	n/a	Publicly deprioritized fundamental SAE research (Smith/Rajamanoharan/Conmy, Mar 2025) — validates applied approach
OpenAI (interp)	Internal lab	n/a	Persona-features work (Jun 2025) descriptive, not productized
Arize AI	Enterprise XAI	$131M	Output monitoring only. Cannot see inside models
Fiddler AI	Enterprise XAI	~$45M disclosed	Post-hoc explanations. Different level of analysis entirely
Arthur AI	Enterprise XAI	$63M	Agent monitoring, not model internals
Apollo Research	Mech interp (nonprofit)	nonprofit	Co-author SPD; Lee Sharkey now at Goodfire. Fellow traveler, not competitor
Transluce	Mech interp (nonprofit)	nonprofit	Ally, not competitor. Public-interest auditing
Leap Labs	Mech interp (startup)	pre-VC-disclosed	UK MATS spinout. Sub-scale; the only other commercial-facing interp player

SECTION 01.04 — GOODFIRE AI COMPANY PROFILE

Incorporated June 2024. $209M raised. $1.25B post-money. ~50 FTE. Results no one else has produced.

FOUNDING TEAM

CEO	**Eric Ho** — Founder of RippleMatch ($10M+ ARR)
CHIEF SCIENTIST	**Tom McGrath** — Founded DeepMind's interpretability team
CTO	**Dan Balsam** — Founding engineer and Head of AI at RippleMatch

FUNDING TRAJECTORY

AUG 2024	Seed — Lightspeed-led	$7M
APR 2025	Series A — Menlo-led *+ Anthropic's first-ever startup investment*	$50M
FEB 2026	Series B — B Capital-led	$150M
TOTAL	*$1.25B post-money · ~50 FTE · SF HQ*	$209M

RESULTS NOBODY ELSE HAS PRODUCED

Research portfolio spans all three layers of modern interpretability: probes for detection (Rakuten), features for steering (RLFR), parameter decomposition for deep audit (SPD).

RESULT	DETAIL	SOURCE	SIGNIFICANCE
RLFR hallucination reduction	58% hallucination reduction on Gemma-3-12B-IT, LongFact++ benchmark. ~90× cheaper than Gemini 2.5 Pro as judge. ~$2.5K compute / 360 optimizer steps. No degradation on standard benchmarks.	goodfire.ai/research/rlfr (Feb 2026)	Strongest published evidence interpretability directly improves performance
Prima Mente Alzheimer's biomarkers	Novel cfDNA fragment-length signal. 7B genomic foundation model, 1.9T tokens; distilled to interpretable classifier; Oxford cohort 81 individuals, 30B+ fragments.	goodfire.ai/research/interpretability-for-alzheimers-detection	First foundation model reverse-engineered into a novel scientific discovery
Rakuten production PII detection	96% F1 vs 51% LLM-as-judge baseline; 15–500× cost savings. SAE probes on Llama 3.1 8B; token-level; synthetic-data-only training; deployed in Rakuten's consumer AI agent.	goodfire.ai/research/rakuten-sae-probes-for-pii-detection	First commercial-grade evidence interpretability beats LLM-as-judge on cost and accuracy
DeepSeek R1 SAEs	First SAEs on a 671B-parameter reasoning model. Open-sourced on HuggingFace + GitHub.	github.com/goodfire-ai/r1-interpretability (Apr 2025)	Interpretability scales to frontier reasoning models
Evo 2 with Arc Institute	Recovered exon-intron boundaries, α-helices, coding sequences, species identity. 7B and 40B Evo 2 variants; published as part of the Nature package.	Arc Institute / Evo 2 release (Feb 2025)	Interpretability transfers to scientific foundation models
Stochastic Parameter Decomposition	Successor to SAEs; addresses feature shrinkage and feature splitting. Bushnaq, Braun, Sharkey.	arXiv 2506.20790 · goodfire.ai/research/stochastic-param-decomp	Theoretical frontier; the next-generation primitive

Goodfire participated in the Circuits Research Landscape collaborative replication (Aug 2025) alongside Anthropic, EleutherAI, DeepMind, and Decode. "Biology of a Large Language Model" (Anthropic, Claude 3.5 Haiku, Mar 2025) is not a Goodfire publication.

SECTION 02 — 03 SLIDES

Problem decomposition.

Not "which customers to pursue" but "what specific barriers prevent a research lab from becoming a commercial business." Six subproblems (SP1–SP6), solved in sequence.

SECTION 02.01 — THE PROBLEM STATEMENT

From research lab to repeatable commercial engine — in 18 months.

INITIAL STATE — APRIL 2026

Research lab with commercial proof points.

— ~50 FTE; ~85% research, ~15% commercial-and-engineering
— $209M raised across Seed + Series A + Series B
— $1.25B post-money valuation (B Capital led, Feb 2026)
— Named partners and customers: Microsoft, Mayo Clinic, Arc Institute, Prima Mente, Rakuten
— Published: RLFR (58% on LongFact++), DeepSeek R1 SAEs at 671B, Evo 2 with Arc, Prima Mente biomarkers, Rakuten PII (96% vs 51%)

GOAL STATE — OCTOBER 2027

Self-sustaining commercial engine — positioned for $4–6B Series C.

— ~230 FTE; ~45% research, ~55% commercial-and-engineering
— $50M ARR base case · $75M stretch (with anchor overperformance + scientific discovery breakout)
— 25–35 customers across three motions
— Repeatable inside-sales motion

REVENUE MIX — OCTOBER 2027

MOTION	CUSTOMERS	ARR RANGE	NOTES
Enterprise platform (FDE-led)	10–15	$8–18M	$800K–1.2M blended ACV; 2–3 anchors expanding to $2–4M
Runtime monitoring (Rakuten-shaped)	8–15	$15–25M	Highest GM (~80%); the scale engine
Scientific discovery (Mayo/Prima Mente shape)	4–6	$6–12M	Pharma/life sciences R&D budget pocket; larger ACVs ($1.5–3M)
Frontier-lab design partnership	1–2	$5–15M annualized	$15–30M TCV over 2–3 years; distinguished from ARR
Strategic research partnerships	3–5	$1–4M	Reference accounts; scientific-instrument proof
TOTAL	**26–43 logos**	**$35–74M ARR**	Base case midpoint ~$50M · stretch $75M

ARR figures denote committed annualized recurring revenue. Frontier-lab partnership figures are TCV; annualized impact $5–15M. Base case assumes: 60% Tier 1→ Tier 2/3 conversion, 140% NRR, Anthropic co-sell formalized by Q3 2026, first scientific discovery anchor beyond Mayo, no major incumbent M&A into mechanistic interp. Series C valuation range assumes 60–90× forward multiple. Built to be sharpened against Goodfire's actual ARR baseline, pipeline coverage, and NRR — not presumed unilaterally.

SECTION 02.02 — THE SIX SUBPROBLEMS

Each is an obstacle between initial and goal state. Solving all six is sufficient to reach the goal.

SUBPROBLEM	OBSTACLE	WHY IT PERSISTS	SOLUTION MECHANISM	SUCCESS METRIC
SP1	Category void	Buyers can't name what to buy. No Gartner category, no budget line. "XAI" captured by SHAP/LIME.	Commission analyst report defining "Model Intelligence" category. Use MRI-vs-thermometer contrast relentlessly.	Buyer unprompted uses category language
SP2	Translation chasm	Researchers speak features/SAEs. Buyers speak risk/ROI. Translation is currently a person, not a process.	Build outcome library. Role-specific briefs. Hire GTM from the buyer's world.	First-call-to-scoped-engagement under 21 days
SP3	Packaging gap	Every engagement designed from scratch. No template. A new AE couldn't scope a deal.	Define 4 engagement tiers with scope, timeline, deliverables, price. Tier 1 runs without researcher.	AE scopes Tier 1 independently
SP4	Proof-of-value bottleneck	Buyers want case studies. Case studies require engagements. Engagements require proof.	Use Tier 1 as loss-leader. Discounted assessment for named case study rights. Mine partner referrals.	Named case study in 3 verticals
SP5	Trust infrastructure	No regulatory body references mech interp. Buyers need external authority.	Submit to EU AI Office consultations. Brief AISI UK. Co-author standard with Stanford HAI.	Regulatory doc references mech interp
SP6	Scalability paradox	Revenue ∝ researcher hours. ~40 researchers across the team (~85% of ~50 FTE). Revenue caps without systematic leverage — the constraint to break, not accept.	Build Model Intelligence OS. Separate execution from innovation. Create "Certified Engineer" role.	Researcher hrs per Tier 1 <20

SECTION 02.03 — DEPENDENCY GRAPH & CRITICAL PATH

SP1 + SP2 first. Everything else is downstream.

ROOT · PARALLEL
SP1 Category void
SP2 Translation chasm

→

STEP 2
SP3 Packaging gap

→

STEP 3
SP4 Proof-of-value

→

ENDGAME · PARALLEL
SP5 Trust infrastructure
SP6 Scalability

THREE REINFORCING LOOPS ACTIVATE WHEN ALL SIX ARE RESOLVED

REVENUE LOOP
Category frame lets buyers find you → translation lets them say yes → packages make deals repeatable.

CREDIBILITY LOOP
Deals produce case studies → case studies feed regulatory credibility → regulation accelerates demand.

SCALE LOOP
Engagements train playbooks → playbooks let engineers run standard work → frees researchers for novel work.

COMPOUNDING ASSET
The feature atlas sits beneath all three — each engagement enriches it; each enrichment makes the next engagement faster.

Attempting to solve SP6 (scalability) before SP4 (proof-of-value) wastes effort — you can't systematize delivery without enough engagements to extract patterns from.

SECTION 03 — 04 SLIDES

Constraints & revenue math.

Four engines to $50M ARR base case (stretch $75M). A Palantir deployment model, not a Stripe funnel. The periodic table prediction motion.

SECTION 03.01 — REVENUE ENGINES, REWEIGHTED

Four engines to $50M ARR base case (stretch $75M). Each sold to a different buyer, against a different budget.

ENGINE 01

Enterprise Platform Deployments

~40%

OF INCREMENTAL ARR

ECONOMIC BUYER: Chief Risk Officer or Chief AI Officer
CHAMPION: Head of Model Risk Management or Head of Responsible AI
BUDGET LINE: Model risk management (existing); R&D services (existing)

FDE-led engagements at $500K–1.5M blended ACV. 10–15 platform customers by Oct 2027, with 2–3 anchor accounts expanding to $2–4M.

Dependency: reduce researcher hours / Tier 1 from ~40 → ~15 via Ember v2 partner-deployment platform

Competes with: hiring 2–3 interpretability researchers at ~$1.5M loaded; Big Four AI assurance at $1–3M.

ENGINE 01

Enterprise Platform Deployments

~30%

INCREMENTAL ARR · THE SCALE ENGINE

ECONOMIC BUYER: VP Engineering or VP AI Platform
CHAMPION: Head of ML Infrastructure or Head of AI Safety
BUDGET LINE: AI guardrails / model safety (existing where Lakera or Protect AI deployed)

SAE probes in production. Rakuten benchmark: 96% F1 vs 51% LLM-as-judge at 15–500× cost savings. $200K–500K ACV initially; expansion to $1–2.5M as model count grows.

Dependency: convert 60%+ of platform customers to ongoing monitoring; ship 3 vertical packages (financial PII, healthcare PHI, agent safety)

Competes with: Lakera ($100–300K), Protect AI ($150–400K), internal build (~$500K loaded). Priced at top of band on Rakuten benchmark differentiation.

ENGINE 03

Scientific Discovery Platform

~15%

PHARMA + LIFE SCIENCES

ECONOMIC BUYER: Chief Digital/AI Officer, Head of R&D
CHAMPION: Head of Computational Biology, Head of Foundation Models
BUDGET LINE: R&D program budgets (biomarker discovery, target validation); not IT/AI budget

EVEE (Mayo Clinic, 4.2M ClinVar variants, 0.997 AUROC), Prima Mente Alzheimer's biomarkers, Evo 2 with Arc Institute — three anchor references in 18 months. $1.5–3M ACVs on 2-year research contracts; pharma R&D program budgets are an order of magnitude above MRM.

Dependency: ship 4–6 pharma/life-sciences anchors by Q4 2027.

Competes with: internal ML/compbio hire, academic collaborations, legacy in silico tooling. Anchored by TIME magazine feature (April 2026).

ENGINE 04

Frontier-lab Design Partnerships

~15%

PLUS ~5% STRATEGIC RESEARCH

ECONOMIC BUYER: CEO or VP Research at frontier lab
CHAMPION: Head of Interpretability or Head of Alignment
BUDGET LINE: Strategic research partnerships (custom; not a standard line item)

1–2 partnerships at $15–30M TCV each, structured as multi-year co-development on reasoning-model interpretability or scientific foundation-model work. Annualized $5–15M cash; remainder in compute/credits and milestone-contingent payments.

Dependency: close first landmark deal Q4 2026.

No direct precedent. Closest analogs: Isomorphic–Lilly ($45M upfront + $1.7B milestones); SSI strategic deals. Founder-led; not a scale motion.

WHY THIS MIX

Runtime feature monitors are the only motion with structural distribution leverage — recurring, near-zero marginal delivery cost, 80%+ GM. Enterprise platform is FDE-bottlenecked. Scientific discovery exploits a larger and less-contested budget pocket (pharma R&D, not IT). Frontier-lab anchors category credibility. Running them simultaneously — with runtime as the scale engine — produces **$50M ARR base case, $75M stretch by October 2027**.

SECTION 03.02 — OPEN-WEIGHT ADDRESSABLE MARKET

Enterprise open-weight deployment jumped 23% → 67% in a year. Six distinct customer categories within it.

Representative targets illustrating category composition — not current Goodfire customers.

01

Research institutions & life sciences

MIT, Stanford, Arc Institute, Broad, CERN, pharma R&D (Roche, Pfizer, AstraZeneca)

"Your foundation model contains discoveries no human has made. We've already found AIzheimer's biomarkers in one."

02

Sovereign AI programs

France/Mistral, UAE/TII, EU sovereign cloud, Japan GENIAC, India BharatGPT

"National AI needs transparency guarantees proprietary APIs can't provide."

03

Data-sovereign enterprises

European banks (Allianz, BNP), US healthcare (Kaiser, HCA), defense (Lockheed, RTX), legal firms

"Self-hosting means you own the model risk. We provide the audit your cloud API used to include."

04

AI-native startups (B+)

Vertical AI (legal, health, fin), agent companies, SaaS fine-tuning Llama/DeepSeek

"Fine-tuning introduced behaviors you can't detect. We surface the 1-in-100K failures."

05

Model platforms

Together AI, Fireworks, Prem AI, Replicate, Hugging Face — potential channels

"Embed our interpretability layer. Your customers get monitoring as a built-in feature."

06

Cost optimizers

E-commerce (Walmart-scale), telcos (T-Mobile, Orange), media switching from APIs

"You saved 90% on inference but lost visibility into why your model behaves. We restore it."

67%

ENTERPRISES ON OPEN-WEIGHT

1B+

LLAMA CUMULATIVE DOWNLOADS

~9%

OF PRODUCTION LLM WORKLOADS ON OPEN-WEIGHT

$2.5B

LLAMA-BASED SPEND, 2026

Sources: Meta, McKinsey, Lucidworks

SECTION 03.03 — PRICING ARCHITECTURE & DEPLOYMENT MODEL

Pricing is positioning. Each tier competes against a different alternative.

PRICING DECOMPOSITION

TIER	SCOPE	ANCHOR (NEXT-BEST ALT)	FLOOR (COST-TO-DELIVER)	PRICE	GROSS MARGIN
Tier 1 · Assessment	1 model, 1 use case	Big Four scoping (~$200K)	~$75K (3wk researcher time)	**$75–150K**	0–50% — case-study tier, not a margin tier
Tier 2 · Standard	1 model, 2–3 use cases, production	Internal interp hire (~$500K loaded)	~$150K (6wk researcher + eng)	**$200–350K**	30–50%
Tier 3 · Surgery	Novel model family, regulated deployment	Hire 2–3 interp researchers (~$1.5M/yr loaded)	~$300K (10–12wk team)	**$750K–$2M**	40–70% — the differentiated tier
Runtime monitoring	Production probes, recurring	Lakera ($100–300K), Protect AI ($150–400K)	~$40–80K (amortized SAE + compute)	**$200–500K/yr**	80%+ — the scale engine
Frontier-lab partnership	Multi-year co-development	No direct precedent; nearest = pharma research deals	Custom; ~$2–5M/yr loaded	**$5–15M annualized ($15–30M TCV)**	Variable — strategic, not margin-optimized

PALANTIR DEPLOYMENT — THREE PHASES

PHASE 01 · ACQUIRE

Tier 1 assessment at **$75–150K**. Loss-leader-adjacent margin (0–50%). Designed to generate the case study that closes the next deal, not to make money. Acknowledged as a strategic choice.

PHASE 02 · EXPAND

Additional models + Tier 2 surgery + runtime monitoring conversion. Target expansion **$800K–1.5M per customer** in 12 months at observed industry expansion rates (Glean, Writer, Sierra benchmarks). Blended ACV of $800K–1.2M assumes ~60% Tier 1 → Tier 2/3 conversion within 12 months. Will calibrate against Goodfire's actual data.

PHASE 03 · SCALE

Runtime monitoring — recurring **$200–500K/yr per account**, ~80% gross margin per Rakuten benchmark structure. The only tier with structural margin to fund growth.

FIVE GROWTH MECHANISMS

01 Assessment → surgery upsell (~60% conv target)
02 Land-and-expand across 10–20 models per enterprise (Palantir analog)
03 Embedded partners as FDE proxy — 3–5 high-fidelity, not 20 warm-intro
04 Feature atlas network effect (each engagement enriches the library)
05 Evidentiary regulatory demand as accelerator (not mandate)

How these prices are set

Anchor = next-best alternative cost (the ceiling). **Floor** = our cost-to-deliver.

Differentiation premium: 1.0× (parity), 1.5× (clear advantage with proof), 2.0×+ (unique capability).

Price = max(floor × 2.5, anchor × differentiation premium), rounded to a memorable number.

Runtime gets a 1.5× premium on Lakera because of the 96% vs 51% F1 benchmark — clear advantage with proof, not unique capability. Surgery gets the same 1.5× because RLFR is published and others can't replicate quickly.

WHY STRIPE FAILS FOR GOODFIRE

01 Not self-serve — Ember API deprecated Feb 2026, confirmed pivot.
02 Unit of value unclear at SaaS layer — feature activations need interpretation.
03 Pricing can't be usage-only — delivery cost dominated by SAE training and FDE time.
04 Requires humans — reports, surgery, compliance reasoning.

SECTION 03.04 — THE MENDELEEV MOTION

Mendeleev didn't sell the periodic table. He predicted undiscovered elements — and when they were found, the table became infrastructure.

Goodfire should operationalize the same motion. Convert skeptics in a way no sales deck can.

STEP 01
Analyze a model's internal features.
Standard Tier 1 assessment.

STEP 02
Make 5–10 specific, testable predictions.
Behaviors under untested conditions.

STEP 03
Customer tests predictions.
Independent validation.

STEP 04
Publish methodology and outcomes.
Public track record compounds.

STEP 05
Customer asks for surgery.
8 of 10 confirmed → credibility overwhelming.

THE PITCH

"Let us analyze your model for two weeks. We'll give you ten predictions about behaviors your eval suite won't catch. You test them."

If we're wrong, you've spent $100K on a failed experiment — and you have a published methodology you can hold us to.

If we're right, you have ten confirmed behavioral findings your eval suite missed, plus a vendor with the only published track record of fixing them via RLFR (58% hallucination reduction, Gemma-3-12B-IT on LongFact++) or runtime probes (Rakuten: 96% F1 vs 51% baseline).

Each confirmed prediction compounds into a public track record — the strongest marketing asset an interpretability company can build, and the only one competitors can't replicate without their own published research.

SECTION 04 — 02 SLIDES

Five levels of analysis.

Every deal is simultaneously personal, organizational, market-contextual, regulatory, and civilizational. Single-level sales fails. Multi-level orchestration wins.

SECTION 04.01 — THE FIVE LEVELS

A buyer's decision is personal, organizational, market-contextual, regulatory, and civilizational — at once.

LEVEL	ANALYTICAL LENS	KEY DYNAMIC	HOW IT SHAPES GTM
L1 Individual buyer	Behavioral psych, decision science	Three personas — Technical champion, Business executive, Risk gatekeeper. Sale closes when all three align for different reasons.	Never run one pitch. Tier 1 assessment produces three outputs: technical report, executive summary, compliance appendix.
L2 Organization	Org behavior, procurement	No procurement category for "interpretability." Multi-stakeholder veto chain. AI maturity gradient.	Position under existing budget — "model risk management," "AI governance," "R&D services." Structure as professional services to bypass software procurement.
L3 Market / industry	Competitive strategy, diffusion	Category pre-formation. Goodfire sells to innovators (2.5%). Peer pressure drives adoption.	Sequence early deals to maximize peer pressure. Commission analyst report defining the category.
L4 Regulatory	Policy analysis, institutional econ	Article 13 "transparency" is undefined. CEN/CENELEC standards will decide whether it means SHAP (bad) or mech understanding (good).	Shape interpretation of Article 13. Brief AISI UK, NIST, FCA. Set the standard before competitors define it.
L5 Civilizational	Philosophy of tech, sociology of risk	Does understanding AI matter? Current trajectory favors "understanding matters."	Never argue for interpretability. Narrate its arrival. Every blog post contributes to civilizational consensus.

EMERGENT PROPERTY

The inevitability narrative. When all five levels align, the buyer doesn't feel like they're making a purchasing decision — they feel like they're recognizing an inevitability. Never argue for interpretability; narrate its arrival.

SECTION 04.02 — THREE CROSS-LEVEL CAUSAL CHAINS

Three ways a deal cascades across levels — top-down, bottom-up, and laterally.

CHAIN 01 — TOP DOWN *The compliance cascade · L5 → L1*

L5 AI incident intensifies civilizational concern → L4 Regulators accelerate enforcement → L1 Risk gatekeeper becomes active buyer

GTM implication: prepare incident-response marketing, pre-qualify regulated leads, ensure Tier 1 can start within 48 hours.

CHAIN 02 — BOTTOM UP *The credibility ascent · L1 → L5*

L1 Surgery with measurable outcomes → L3 Peers inquire via competitive pressure → L5 Narrative: "here's proof we can"

GTM implication: select early deals for upstream narrative potential.

CHAIN 03 — LATERAL *The scientific discovery loop · all levels*

L1 Researcher discovers novel insight → L3 New market segment emerges → L5 AI reframed as telescope at nature

GTM implication: every discovery is marketing. Every paper is a sales deck.

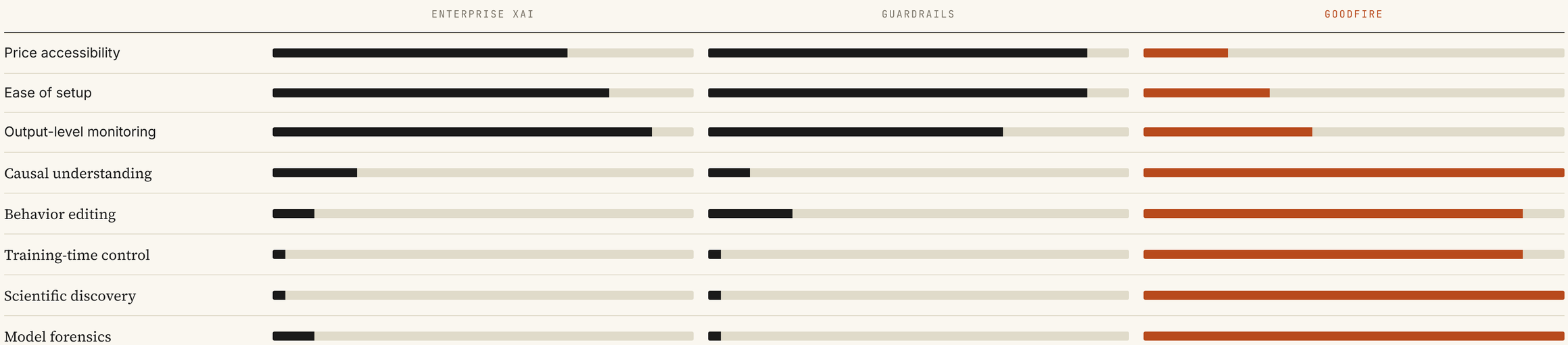
SECTION 05 — 04 SLIDES

Blue ocean strategy.

Competitors cluster on the left. Goodfire's value curve is inverted — dominant on the right across five factors where nobody else competes.

SECTION 05.01 — THE STRATEGY CANVAS

The inverted value curve. Low where competitors cluster — maximum where no one else competes.



Competitors compete on price, setup ease, and output monitoring — the left side of the canvas. Goodfire is deliberately low on those factors and dominant across the factors where nobody else operates: causal understanding, behavior editing, training-time control, scientific discovery, and model forensics. The right side is empty ocean.

SECTION 05.02 — FOUR ACTIONS FRAMEWORK

What to eliminate, reduce, raise, and create.

ELIMINATE

Self-serve developer playground (deprecated, correct). Dashboard-based monitoring UX — don't compete with Arize on dashboards. *"Explainability"* positioning — reject the category; it maps to SHAP/LIME commodity.

REDUCE

Breadth of model support — depth on 5–6 models beats shallow coverage of many. Ease of self-serve setup — 2-week onboarding signals high value. Price accessibility — $100K minimum is a trust signal.

RAISE

Causal understanding — from correlation to causation: *"14 features in layers 18–24 are responsible."* Compliance depth — mechanistic analysis, not process docs. Measurable outcomes — every engagement has pre-agreed metrics.

CREATE

Model surgery — give us your broken model, we return a fixed one. Scientific discovery from models — zero competitors, infinite TAM in pharma. Model forensics — NTSB for AI incidents. Training-time interpretability (RLFR).

SECTION 05.03 — FOUR CROSS-DOMAIN ANALOGIES

Four analogies. Each points at a different facet of the GTM motion.

ANALOGY	STRUCTURAL PARALLEL	KEY GTM LESSON
Illumina — genomics	SAEs = sequencing instrument. Feature atlas = reference genome. Each genome sequenced makes the reference more valuable.	Target research institutions first (genome centers), not mass enterprise. Platform economics via compounding data asset.
Penetration testing — cybersecurity	Model failures = security breaches. EU AI Act = PCI-DSS. Model forensics = incident response. Model surgery = remediation.	Sell fear of specific consequences. First deliverable: *"here are 47 ways your model fails right now."*
SGS — materials testing	AI models are digital materials needing specification testing. Unlike SGS, Goodfire also fixes — *"test and fix" > "test and report."*	Per-model testing fee. Accreditation is the moat — become the methodology regulators reference.
Periodic table — history of science	Current AI = alchemy. Feature atlas = periodic table. RLFR = first chemical synthesis. Predictions prove the framework.	Make public, verifiable predictions about model behavior. Marketing through science.

SECTION 05.04 — THREE TIERS OF NONCUSTOMERS

The biggest market is the one that doesn't know it's a market yet.

TIER 01 — SOON-TO-BE

500–1,000

Orgs deploying AI, experiencing failures, spending on partial solutions

Why they don't buy today: current solutions are "good enough" but don't understand root cause.

"You spend $500K/yr on guardrails that don't know why your model fails. Goodfire diagnoses root cause in 2 weeks — at a fraction of what a single model failure costs in production."

TIER 02 — REFUSING

1,000s

Banks refusing AI credit. Hospitals refusing AI diagnostics. Defense refusing AI targeting.

Why they don't buy today: can't trust AI because can't explain it. XAI is insufficient.

"We can make your model auditable enough for [regulator]. Here's the mechanistic analysis your model risk team needs."

TIER 03 — UNEXPLORED

∞ TAM

Pharma not knowing models contain drug targets. Climate researchers. Materials scientists.

Why they don't buy today: don't connect their problem to AI interpretability at all.

"Your foundation model knows things about biology no human has discovered. We can extract that knowledge."

SECTION 06 — 04 SLIDES

Product strategy.

Four primitives — read, diff, write, train — combine into twelve products across four tiers. Each product strengthens the others in ways that compound.

SECTION 06.01 — THE FOUR PRIMITIVES

Four capabilities no other company possesses simultaneously. Every product concept flows from combining them.

01 PRIMITIVE

Read

SAEs, probes

Extract features and activations from inside a model. The foundation — without reading, nothing else is possible.

02 PRIMITIVE

Diff

Logit diff amplification

Compare models or checkpoints. Surface rare behaviors (1-in-1M samples) that standard evals miss entirely.

03 PRIMITIVE

Write

Feature steering

Modify model behavior by acting on internal features directly — without retraining.

04 PRIMITIVE

Train

RLFR, intentional design

Train models differently using interpretability signal. The 58% hallucination reduction (Gemma-3-12B-IT) lives here.

THE COMBINATORIAL LOGIC

Read + Diff = **Model Forensics.** Read + Write = **Model Surgery.** Read + Train = **RLFR.** Diff + Train = **Training Curriculum Intelligence.** All four = **Interpretability Compiler.**

SECTION 06.02 — CATEGORY-DEFINING & MOAT-BUILDING PRODUCTS

Six products on the near horizon. Three to define the category. Three to defend it.

TIER 01 *Category-defining — build now*

OUTCOME-BASED

Model Surgery Studio

Customer delivers model + behavior spec. Goodfire identifies causal features, applies RLFR training, returns a surgically improved model with mechanistic audit trail. *The 58% hallucination result (Gemma-3-12B-IT), productized.*

Buyer: Chief AI Officer / Head of Model Risk. **Budget line:** model risk or R&D services. Closes against an existing pain (production incident or audit finding).

$250K–$1M per engagement

PER-ANALYSIS

Model Diff Engine

"Git diff for neural networks." Compare checkpoints, get a human-readable report of what changed internally. Surfaces 1-in-1M behaviors evals miss.

Buyer: VP Engineering / Head of ML Platform. **Budget line:** ML infrastructure or eval tooling. Closes against an existing pain (regression after fine-tuning, untrusted model update). Entry-tier pricing by design: the diff report surfaces findings that convert to Tier 3 surgery engagements at ~60% rate.

$75K–$150K per analysis

USAGE-BASED

Runtime Feature Monitors

Lightweight probes on internal feature activations, real-time. Detects hallucination likelihood before the hallucination surfaces. Deployed at Rakuten (44M+ user platform).

Buyer: VP AI Platform / Head of AI Safety. **Budget line:** AI guardrails (existing where Lakera/Protect AI already deployed). Closes against an existing pain (PII leakage, agent safety incident).

$50K setup + per M tokens

TIER 02 *Moat-building — 6 to 12 months*

DATA NETWORK EFFECT

Universal Feature Atlas

Cross-model knowledge graph of labeled features. Each model interpreted enriches the library. *The Palantir Ontology play.*

VOLUME PLAY

Interpretability-guided Compression

SPD identifies causally important structures. Returns a 3× smaller model that retains 95%+ capability. Every enterprise running LLMs at scale would pay for this.

REPLACES FINE-TUNING

Behavior Design Console

Visual interface for specifying behavioral profiles via feature activations. Fine-tuning takes weeks; this takes minutes.

SECTION 06.03 — MARKET-CREATING PRODUCTS & MOONSHOTS

The horizon beyond the moat. Six more products that redefine what AI engineering means.

TIER 03 *Market-creating — 12 to 24 months*

Scientific Discovery Engine

Reverse-engineer domain-specific foundation models to extract novel hypotheses, biomarkers, mechanisms. The Alzheimer's result as a platform. *A single drug target = hundreds of millions.*

Training Curriculum Intelligence

Real-time interpretability during training. "This data strengthens deception features — remove it." Alchemy becomes engineering.

Model Forensics

NTSB-style post-incident investigation for AI failures. Every deployer needs this after their first public incident.

Cross-modal Interpretability

Unified platform across language (Llama), reasoning (DeepSeek R1), vision (SDXL-Turbo), genomics (Evo 2).

TIER 04 *Moonshots — 24+ months*

Feature Transplantation

Identify features responsible for a capability in Model A — transplant them into Model B. Model capabilities become composable modules. SPD is the technical foundation.

NORTH STAR

The Interpretability Compiler

Translate natural language into targeted weight updates. *"Make this model stop hallucinating about historical dates"* → permanent weight change in 30 seconds. If this works, it obsoletes the entire fine-tuning industry.

THE FLYWHEEL

Model surgery generates labeled features → features enrich the atlas → atlas makes the next surgery faster → runtime monitors generate production data → forensics produces failure-mode features → scientific discoveries build credibility. Each product strengthens the others in ways that compound.

SECTION 07 — 02 SLIDES · SHIPPED

Forge.

A commercial intelligence OS. Eight interfaces on a three-layer data architecture. The frameworks in this document, operationalized.

SECTION 07.01 — EIGHT INTERFACES

Forge informs judgment with complete context no unaided human brain can maintain.

01 / PROSPECT INTELLIGENCE

Every prospect scored by a four-filter ICP algorithm — explained, not black-boxed.

Model family match (40%) · regulatory pressure (25%) · peer cluster density (20%) · recent signals (15%). Top 5 priority targets every Monday.

SP1 · SP5 · L1–L5

02 / SOLUTION + PRICING

Three-stage engagement scoping across four tiers ($75K–$15M).

Auto-classifies across four engagement tiers with cost-to-deliver, margin, and LTV projections. Every assumption documented and traceable.

SP2 · SP3 · SP6

03 / GTM COMMAND CENTER

Signal feed with four-factor actionability scoring.

Discourse monitoring across regulatory, competitive, research, prospect sources. ROI calculator grounded in real Goodfire benchmarks.

SP5 · L3 · L4

04 / RESEARCH + PREDICTIONS

The Mendeleev motion, operationalized.

Each engagement produces testable predictions. Accuracy dashboard tracks confirmed vs. refuted over time. Partner health alerts auto-draft check-ins.

SP4 · CATEGORY

05 / MODEL COVERAGE

Three-tier SAE readiness — Available, Planned, On-Demand.

Decision triggers fire when pipeline demand on unsupported models crosses 3+ qualified prospects AND $500K+. Breakeven matrix shows SAE payback.

SP2 · SP6

06 / CHANNEL PARTNERSHIPS

Embedded partner pipeline.

Partner fluency scoring, co-delivery capacity, methodology-fidelity tracking. Visualizes the embedded-partner multiplier — 3–5 high-fidelity partners deliver more throughput than 20 warm-intro firms.

ENGINE 02

07 / NARRATIVE ENGINE

Same signal, four voices.

Discourse monitoring against Goodfire capabilities. Content calendar aligned to regulatory moments. Framing across ML Engineer, CTO, Compliance, AI Community.

SP1 · SP5

08 / OPS + WEEKLY BRIEF

Strategic command center + auto-generated Monday brief.

Pipeline funnel, three engines, TAM across six categories, conversion analytics, adjustable ICP weights. Replaces 2–3 hours of manual assembly.

ALL SUBPROBLEMS VIA SYNTHESIS

SECTION 07.02 — THE INTELLIGENCE THAT COMPOUNDS

System of record vs. system of intelligence — over 52 weeks, the difference is the difference between prescient and catching up.

Projected outcomes at steady state. Current deployment is a single-user prototype; numbers are modeled from the architecture, not observed from production.

01 COMBINATORIAL

Cross-references 200+ signals against 150+ prospects × 6 categories × 10 model families. **estimated 30–50 non-obvious connections/yr** vs. 5–10 for humans.

02 MEMORY

Dormant prospects never decay. When conditions align, Forge reactivates automatically. **Recovers an estimated 15–25% of pipeline** typically lost to drift.

03 CALIBRATION

Every signal rated, every conversion tracked. Discovers counterintuitive patterns within 90 days (e.g. **compliance framing converts 3.6× better** than CTO framing).

04 PRIORITIZATION

Replaces "whatever feels urgent" with weighted expected-value scoring. **estimated $8M vs. $15M closed on the same work input.**

05 MARGIN

Cost-to-deliver visible on every deal. Prevents money-losing engagements. **estimated $1–2M in recovered margin/yr.**

06 PREDICTION

Cumulative track record becomes a compounding credibility asset. *"84% predictive accuracy across 11 model families"* — strongest marketing claim in the industry.

07 COORDINATION

Every handoff between GTM, Applied AI, and research carries full context. **Zero dropped balls.**

DEPLOYED

Eight pages · ~80 components · 20+ library modules · Next.js 14 · TypeScript strict · SQLite · Anthropic SDK.

Loom walkthrough · loom.com/share/97225307f96a417783e964284c00f0b6
Live app · forge-lemon-beta.vercel.app
Source · github.com/chewyuenrachael/forge

THE ONE-SENTENCE CASE

Most GTM tools are systems of *record* — they capture what the GTM lead already knows. Forge is a system of *intelligence* — it surfaces what the GTM lead does not yet know, and learns which insights actually convert.

END OF ANALYSIS · THANK YOU

Never argue for interpretability. Narrate its arrival.

BUILT TO BE SHARPENED

Numerical targets are triangulated from comparable Series B trajectories, Goodfire's publicly named customers, and the published research portfolio. Built to be sharpened against actual data in conversation.

The frameworks survive calibration. The numbers should be expected to shift.

SOURCES & METHODOLOGY

Grand View · Mordor · Precedence · Fortune Business Insights
PitchBook · Tracxn · CB Insights · Goodfire blog · Anthropic
EU AI Act · UK AISI · NIST · McKinsey · Lucidworks

CONTACT

rachaelchewyuen@gmail.com · rachaelchew.com