

AI EXPLAINABILITY • WHITE PAPER

The AI Explainability Scorecard

A practical framework to measure whether your AI can actually explain itself — and a reference for scoring every model you deploy.

01 INTRODUCTION

The Power and Peril of the Black Box

When a medical AI system once recommended denying a patient treatment, the doctors hesitated — but couldn't explain why.

The algorithm's reasoning was invisible, locked inside a mathematical "black box." Only later did an audit reveal the model had learned to equate zip codes with health outcomes — unintentionally penalizing people from poorer neighborhoods.

This moment captures why AI explainability isn't just a technical nicety. It's a moral and practical necessity. By making the logic behind AI decisions transparent, we can question, improve, and trust the systems that increasingly shape our lives.

Question

Surface why a decision was made, not just what was decided.

Improve

Expose hidden bias and unsafe correlations before they cause harm.

Trust

Turn opaque systems into transparent, auditable partners.

02 WHY IT MATTERS

Why AI explainability matters

Explainability is the foundation of trustworthy and accountable AI. The case for it spans three reinforcing dimensions.

01 It builds trust and ethical oversight

By making model reasoning visible, explainability helps users understand decisions, builds confidence, and enables ethical oversight. Transparent reasoning also reveals hidden risks — such as bias and unsafe correlations — before they cause harm.

02 It's a legal and ethical requirement

Regulations like the **EU AI Act**, **GDPR**, and the **U.S. AI Bill of Rights** demand transparency, traceability, and auditability for AI that affects human rights or safety. Explainable models make such compliance achievable through traceable evidence.

03 It enhances debugging, learning, and collaboration

By showing which inputs drive outputs, explainability helps developers diagnose errors and refine models. It also empowers humans to question, interpret, and improve AI decisions — turning opaque systems into transparent partners that boost both reliability and trust.

03 INTERPRETABILITY VS. EXPLAINABILITY

Why some models are inherently more explainable than others

Two properties determine how well humans can understand an AI model — and they are not the same thing.

INTERPRETABILITY

A person can trace how the model works **without any extra tools**. Decision trees and linear regression are interpretable by design: the logic is visible and follows clear rules.

EXPLAINABILITY

How well a model **communicates its reasoning in human terms**, regardless of whether its internals are visible. A model can be explainable without being interpretable — using post-hoc tools like SHAP to surface what drove a decision.

All interpretable models are explainable, but not all explainable models are interpretable. Interpretability is one path to the ultimate goal — explainability.

The real trade-off in AI is not between performance and transparency, but between **scope and transparency**. Small, focused models are easy to inspect because their logic is narrow and predictable. Larger, more capable models handle far broader tasks but require external tools to explain their reasoning. **The more a model can do, the harder it becomes to see inside it.**

04 THE SCORECARD

The AI Explainability Scorecard

It's important to set clear criteria to evaluate the need for explainable AI. Some models are harder to explain than others, and not every AI use case needs the same level of transparency. High-stakes or safety-critical systems — such as medical diagnosis, financial assessment, or autonomous control — require stronger explainability standards. Lower-risk or exploratory uses can tolerate less if it enables broader scope or experimentation.

To evaluate how effectively an AI model communicates its reasoning, each criterion is rated on a **1–5 scale**, aligned with the "Question It Answers" for that criterion. The final score is the average of all five.

5

A strong "yes" — the explanation fully answers the question to the user's satisfaction.

1

A strong "no" — the explanation fails to answer the question meaningfully.

ALL FIVE CRITERIA SERVE TWO QUESTIONS

Can you trust the explainability?

Driven by **Sound** and **Consistency**.

How actionable is the explainability?

Driven by **Readable**, **Optimize**, and **Easy**.

05 REFERENCE CARD

AI Explainability Scorecard criteria

Use this as your reference. Score every AI solution you deploy against these five criteria — rate each from 1 to 5, then average.

CRITERION	DEFINITION	QUESTION IT ANSWERS
Sound ● ● ● ● ●	The explanation accurately represents the model's true reasoning — how it actually made its decision, not a plausible-sounding reconstruction.	<i>"Is this explanation actually how the model made its decision?"</i>
Consistency ● ● ● ● ●	Similar inputs lead to similar explanations.	<i>"Would the model explain similar decisions in the same way?"</i>
Optimize ● ● ● ● ●	The explanation has value beyond audit trails — helping engineers debug, validate, and improve the model.	<i>"Does this explanation provide actionable insights to improve the AI system?"</i>
Readable ● ● ● ● ●	The explanation is clear and meaningful to non-technical users and subject-matter experts.	<i>"Does this help a non-technical expert understand and act within their field?"</i>
Easy ● ● ● ● ●	The explanation is easy to obtain, interpret, and apply without excessive time, expertise, or compute.	<i>"Can this explanation be generated and used efficiently without significant burden?"</i>

Interplay between criteria. The trust criteria (Sound, Consistency) and actionability criteria (Readable, Optimize, Easy) often trade off. Making an explanation more sound can expose how complex the model really is, making it harder to read. The goal is not to maximize all five, but to achieve a **balance aligned with the model's purpose, audience, and risk profile.**

06 MODEL COMPARISON

Comparing explainability across model types

We apply the scorecard to three model types — each a key stage in the evolution of AI, and a distinct approach to learning and reasoning. The goal is to assess the level of explainability achievable with minimal effort.

K-NN

K-Nearest Neighbors

A simple, clear method that makes decisions based on similar or nearby data points. Easy to understand.

NEURAL NETS

Neural Networks

Powerful tools used in deep learning. They work well for specific tasks, but it's hard to see how they make decisions.

LLMS

Large Language Models

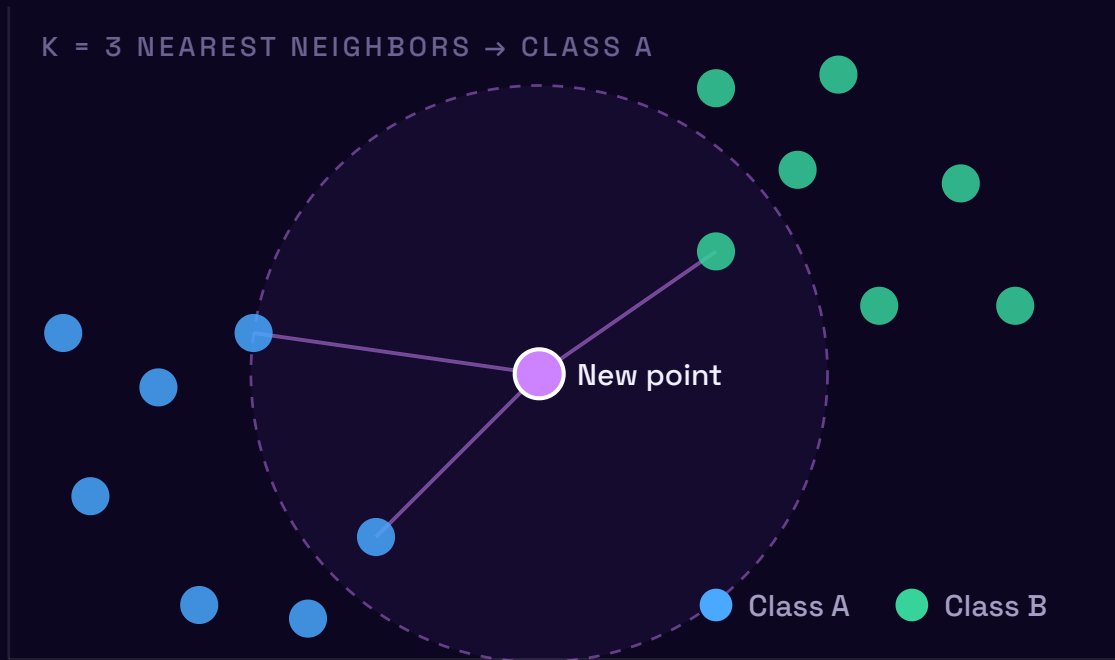
The latest and most advanced systems — an evolution of transformers. They handle an enormous range of tasks, but how they decide is largely hidden from view.

Together, these models span the full spectrum of explainability — from inherently interpretable methods to those requiring advanced external explanation tools.

K-Nearest Neighbors (K-NN)

SCORE 4.8 / 5

K-NN makes decisions by looking at the most similar examples from its training data. Its behavior is directly tied to that data, making it one of the most inherently interpretable AI methods.

**Sound 5 / 5**

Explanations are perfectly faithful to the decision process. Each prediction traces directly back to evidence in the training data — no hidden abstraction.

Consistency 5 / 5

Always gives the same result if the data, settings, and rules stay the same. Fully deterministic; no random guessing.

Optimize 5 / 5

Complete actionable insight: developers improve outcomes by adjusting training data, feature scaling, or correcting labels.

Readable 4 / 5

“Similar cases lead to similar outcomes” is intuitive — though reading a list of neighbors and distances is less direct than a narrative explanation.

Easy 5 / 5

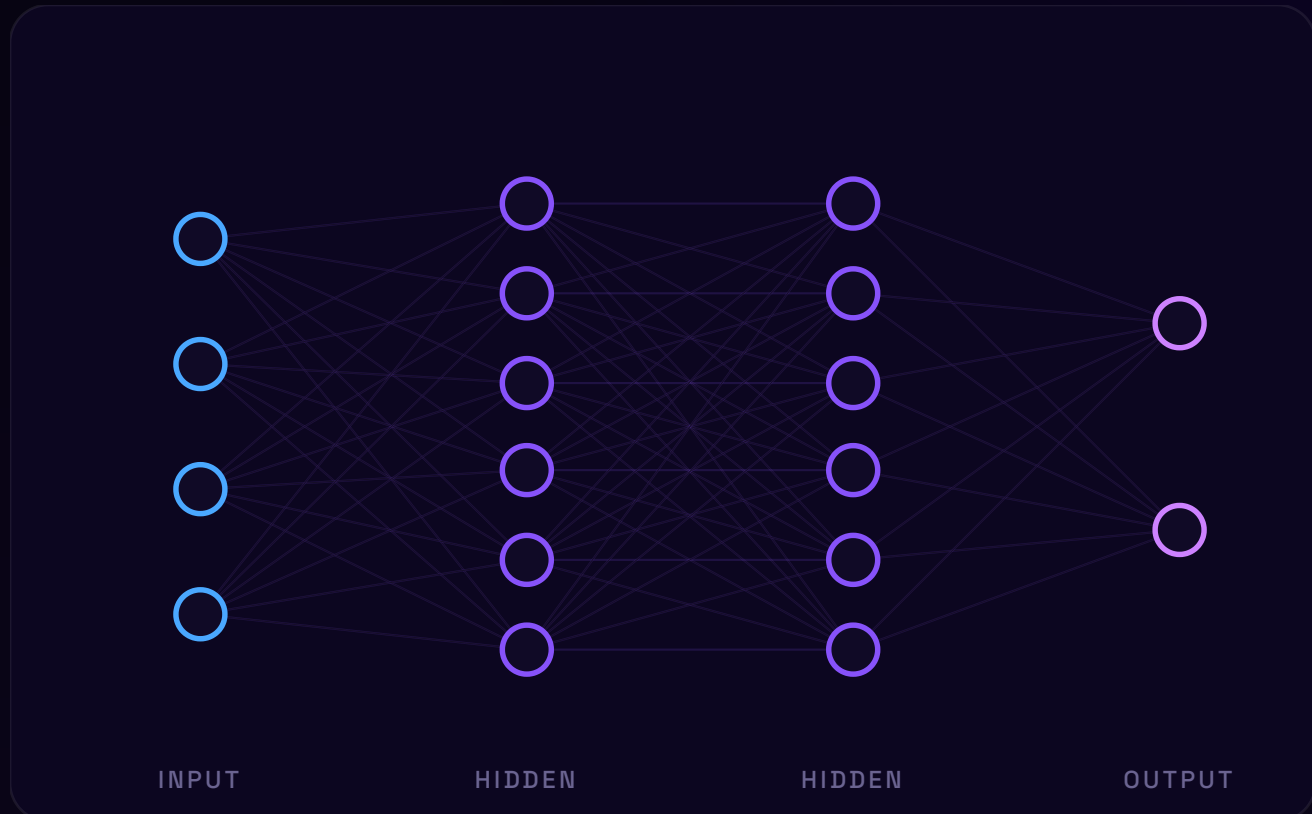
Immediate and transparent. The relevant samples, distances, and labels are inspected directly without tooling.

K-NN achieves near-perfect explainability — explicit, faithful, consistent, and easy to inspect. A top choice for simple, understandable AI.

Neural Networks

SCORE 3.6 / 5

Layers of connected nodes transform input data using math and learned weights. They excel at finding patterns in complex data, usually on one clear task — but how they decide is hard to see. Without special tools, it's almost impossible to explain their thinking.

**Sound 3 / 5**

Tools like SHAP highlight influential inputs to approximate true reasoning, but hidden-layer complexity prevents fully faithful representations.

Consistency 4 / 5

Consistent outputs and explanations when inputs and parameters are fixed; small input or tooling changes shift results slightly.

Optimize 5 / 5

Weights, gradients, and activations directly reveal how to debug and improve the model — concrete enough to guide meaningful changes.

Readable 3 / 5

With SHAP or visualizations, experts can see which parts of the data drive results — but non-technical users still need help.

Easy 3 / 5

SHAP and LIME make explanations possible but require technical skill and compute. Practical for engineers, not effortless.

Moderate explainability through external tools rather than inherent transparency — partly understandable, well enough for experts to improve the model.

Large Language Models (Chain-of-Thought only)

SCORE 3.0 / 5

LLMs predict the next word by learning from huge amounts of text. They understand and generate language impressively, but how they reason is a mystery. Here, explainability is evaluated using **Chain-of-Thought (CoT)** reasoning only — the model expresses its steps in natural language, with no additional interpretive methods.

Sound 2 / 5

CoT outputs sound plausible but don't necessarily reflect the model's actual internal decision-making.

Consistency 1 / 5

Explanations vary on the same input — even with temperature and random seed held constant.

Optimize 2 / 5

The narrative format provides limited value for debugging or improving models.

Readable 5 / 5

The CoT narrative is highly understandable to both SMEs and general users.

Easy 5 / 5

Generated through simple prompting — no technical expertise required.

THE MEDICAL CASE

A CoT explanation might read: *"The patient has missed follow-ups, unstable coverage, and lower expected recovery, so treatment is not recommended."*

It sounds reasonable — but may simply rationalize a decision already made, without revealing that **zip code** was the key driver. Convincing yet structurally opaque, which is perhaps **more dangerous than no explanation at all**.

SUMMARY — EXPLAINABILITY BY MODEL TYPE

CRITERION	K-NN	NEURAL NETS	LLMS (COT)
Sound	5	3	2
Consistency	5	4	1
Optimize	5	5	2
Readable	4	3	5
Easy	5	3	5
Overall score	4.8	3.6	3.0

07 BEYOND CHAIN-OF-THOUGHT

Improving the explainability of large language models

Basic LLM explanations like Chain-of-Thought are easy to get but not very accurate or detailed. To improve them, we use the **DARPA Explainable AI (XAI) framework**, which defines three categories of explainability methods.

Deep Explanation

Techniques that expose or visualize a model's internal reasoning processes.

Interpretable Models

Architectures that are transparent by design — inherently understandable without external tools.

Model Induction

Post-hoc methods that build simplified approximations or external analyses of black-box behavior.

Because LLMs have billions of parameters and learn from vast unstructured data, they are **not interpretable by design**. We therefore focus on Deep Explanation and Model Induction — practical strategies to illuminate how these opaque systems reason, respond, and generalize.

DEEP EXPLANATION METHODS

Feature Visualization

SCORE 3.0 / 5



Aggregates attention weights or activations across layers to approximate how information flows through the model. Due to the massive parameter scale of modern LLMs, only coarse, high-level summaries are practical to visualize.

Sound 5 / 5

The path through the network shows exactly how inputs resulted in outputs — the only method that directly traces actual internal computation.

Consistency 4 / 5

Mostly follows clear, repeatable steps, though even at temperature zero the answer may shift slightly due to parallelism and sampling.

Optimize 2 / 5

Limited insight — engineers see attention focus areas but not the underlying transformations or decision pathways.

Readable 1 / 5

Requires ML expertise. At LLM scale, visualizations become abstract graphs of billions of parameters that convey little meaning.

Easy 3 / 5

Technically easy to create, but making them clear and useful for large models takes heavy compute and extra steps.

Produces attractive visuals, not genuine understanding. Its explainability impact is low where it matters most.

MODEL INDUCTION METHODS

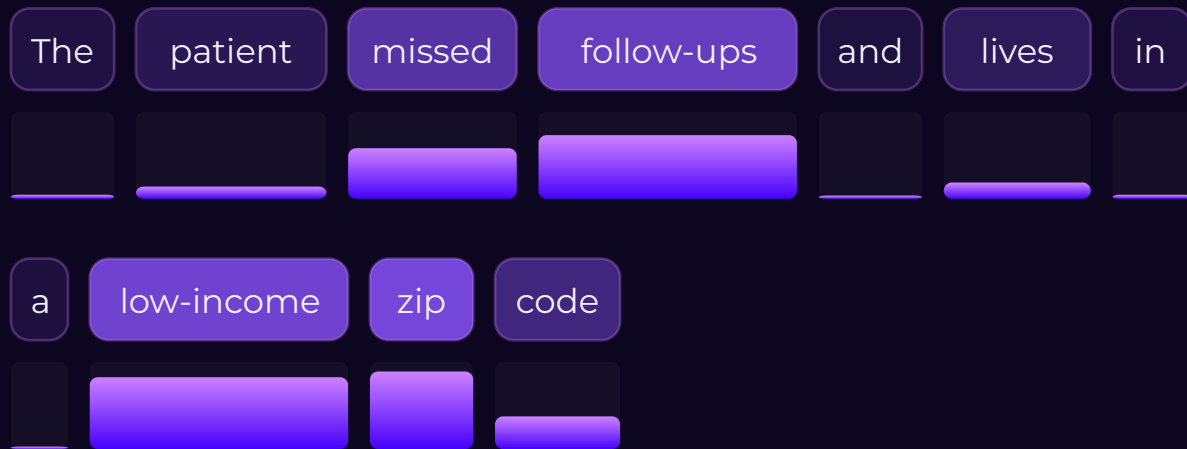
SHAP

 (SHapley Additive Explanations)

SCORE 2.6 / 5

SHAP estimates how much each input **token** contributes to a model's output, assigning each a value for its marginal contribution. In theory, useful; in application, LLMs are too large and complex for it to compute and interpret well.

TOKEN CONTRIBUTION TO PREDICTION · “DENY TREATMENT”

**Sound** 4 / 5

Divides credit across tokens by how they work together, but with billions of parameters it can only estimate those interactions.

Consistency 4 / 5

Steady results when input and model stay the same; small changes shift which tokens seem most important.

Optimize 2 / 5

Shows which tokens influence predictions, but not how contributions propagate through deep attention layers.

Readable 2 / 5

Token-level attributions are hard for non-technical users — tokens are often subword pieces, not intuitive words.

Easy 1 / 5

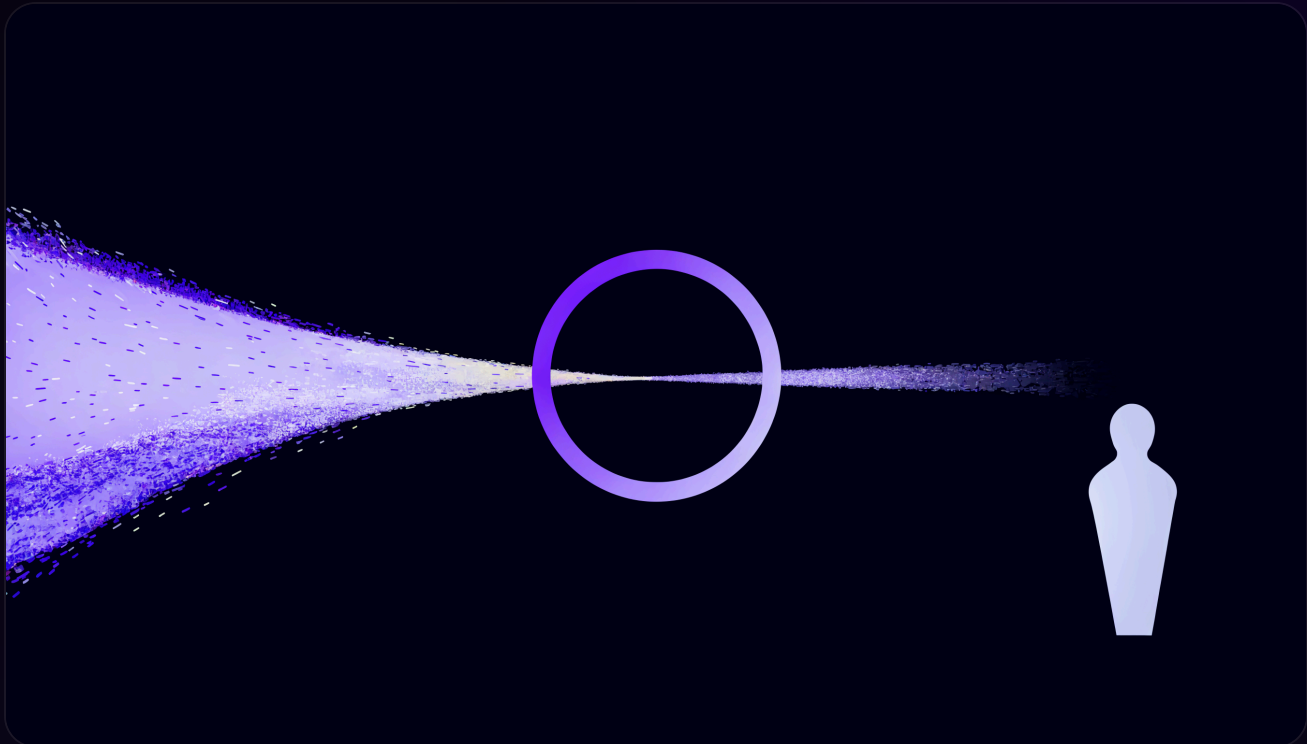
Computationally expensive: the model must run many times per explanation. Cost-prohibitive at every interaction.

Solid math, clear in theory — but impractical for today's huge LLMs. Useful for small models, not real-world scale.

MODEL INDUCTION METHODS

Surrogate Models (Monitoring Approach)

SCORE 3.8 / 5



A K-NN model works as a real-time monitoring surrogate — not a replacement for the LLM. It compares how the model represents full sentences against past examples, explaining behavior in plain terms: *the model answered this way because it saw something similar before.*

Sound 3 / 5

Shows how inputs and outputs relate, but reflects what we can see — not the deeper cause inside the LLM.

Consistency 5 / 5

With stable embeddings, distance rules, and reference data, the same input always finds the same matches.

Optimize 5 / 5

Precise, actionable insight: which prior sentences influenced an output supports debugging and bias analysis.

Readable 4 / 5

Users relate outputs to similar examples — clear, context-driven reasoning for technical and non-technical audiences.

Easy 2 / 5

Building a good labeled example set is costly up front; once established, it runs cheaply and quickly.

In the medical case, this surfaces similar historical patients from low-income zip codes who received worse predictions — making hidden bias **visible and actionable**.

08 AT A GLANCE

Explainability for LLMs at a glance

How each technique scores across the framework — from the CoT baseline through Deep Explanation, Model Induction, and the interpretable surrogate.

CRITERION	LLM BASELINE (COT)	FEATURE VIZ (DEEP EXPL.)	SHAP (MODEL IND.)	SURROGATE (INTERPRETABLE)
Sound	2	5	4	3
Consistency	1	4	4	5
Optimize	2	2	2	5
Readable	5	1	2	4
Easy	5	3	1	2
Overall	3.0	3.0	2.6	3.8
Can you trust it?	No	Yes	Yes	Yes
Drives decisions?	Yes*	No	No	Yes

* CoT drives decisions, but explanations may not reflect the model's true reasoning.

SHAP illustrates how traditional methods struggle to scale with large, general-purpose LLMs, where feature interactions are too complex for stable attribution. **Surrogate models** — approximating behavior through observation and comparison of similar examples — offer a more practical, real-world path to interpreting LLM decisions.

09 CONCLUSION

Explainability is the cornerstone of trustworthy AI

Explainability bridges human understanding and machine reasoning. As models grow more complex and autonomous, transparency becomes both harder and more essential. Agentic AI systems — which act and reason independently — cannot be considered trustworthy without robust explainability.

Explainability spans a continuum, from inherently interpretable models like K-NN to opaque LLMs that rely on post-hoc methods. Traditional tools such as SHAP falter at LLM scale, whereas a **surrogate model using sentence-level K-NN monitoring** offers a promising lens into model behavior.

Ultimately, explainability must integrate interpretability, continuous monitoring, and transparent design — so that increasingly agentic AI systems remain understandable, auditable, and aligned with human values.

Interpret

Favor inherently transparent methods where stakes are high.

Monitor

Use real-time surrogates to keep opaque models accountable.

Prove

Design for traceable evidence regulators and users can trust.

SEE IT IN ACTION

Make every AI you deploy explain itself.

Cranium monitors hundreds of agentic AI risk signals in real time — surfacing the reasoning behind every decision so your teams stay in control.

[Book My Demo →](#)