Red Hat AI

From Ethics To Engineering

Robbie Jerrom

Principal Technologist AI : AI Business Unit

Red Hat

From Ethics To Engineering

We have powerful models, but we lack tools and architectures to make them trustworthy, auditable, and aligned with societal expectations. It's not enough to state principles; they must inform your architecture, platform, and decision-making.

We need more than checklists; we need AI system design that embodies trust as first-class.

AI Ethics… it's complicated

AI ethics addresses the moral principles and societal impacts of artificial intelligence systems

Fairness and Bias

Treat all people equitably without discrimination.

Transparency and Explainability

Show your working, and explain how decisions are made.

Privacy and Data Rights

Protect personal data and respect user consent.

Safety and Control

Prevent harm and maintain human oversight.

Accountability

Take responsibility when things go wrong.

Economic and Social Impact

Consider effects on jobs, inequality, and access.

How can AI platform engineering help?

Technical governance through transparency

Black-box AI no longer suffices under growing regulatory oversight and public scrutiny. Open-source models offer a more auditable and transparent approach, which is essential for meaningful governance, compliance, and trust.

Let's start with the **model**, and then discuss the systems and platforms surrounding it.

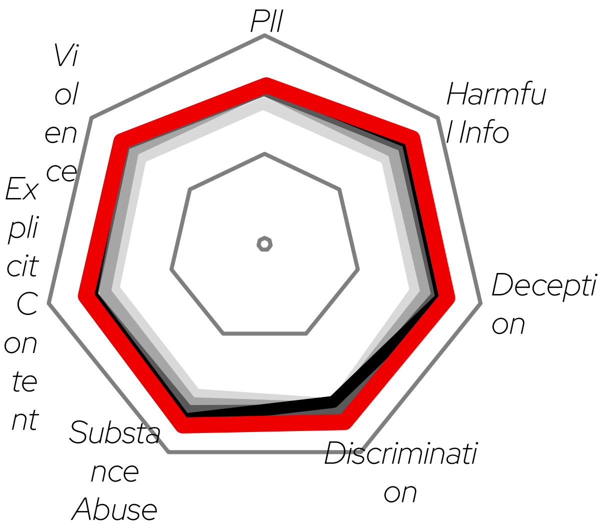
Open-Source AI Models

Unlike open-source software, open-source AI is a little more complex.

License Type	Example Models	Commercial Use	Key Restrictions	Can Fine-tune?	Can Distill?	Training Data Available?
Apache 2.0	IBM Granite Mistral Falcon	✅ Unlimited	✓ None ✓ Use anywhere ✓ Any user scale	✅ Yes	✅ Yes	Granite: ✅ Yes Others: ❌ No
Meta Llama 3.x	Llama 3, 3.1, 3.2, 3.3	✅ Yes (below 700M users)	⚠️ 700M monthly user limit ❌ Can't train other LLMs ❌ Can't distill	✅ Yes	❌ No	❌ No
Meta Llama 4	Llama 4 Opus 4.1 Sonnet 4, 4.5	✅ Yes (below 700M users)	⚠️ 700M monthly user limit ❌ Can't train other LLMs ❌ Can't distill 🇪🇺 No multi-modal in EU	✅ Yes	❌ No	❌ No
BigScience RAIL	BLOOM BLOOMZ	✅ Yes (with ethical rules)	⚠️ No surveillance use ⚠️ No discrimination ⚠️ No misinformation	✅ Yes	✅ Yes (keeps rules)	✅ Yes
BigCode RAIL	StarCoder StarCoder 2	✅ Yes (with ethical rules)	⚠️ No malware ⚠️ No exploits	✅ Yes	✅ Yes (keeps rules)	✅ Yes
MIT License	DeepSeek R1, V3, V3.1	✅ Unlimited	✓ None ✓ Use anywhere ✓ Any user scale	✅ Yes	✅ Yes (can train other LLMs)	❌ No

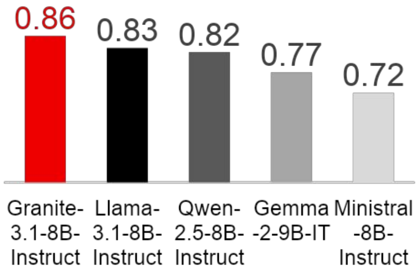
It’s not just the license,

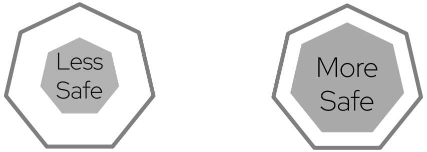
Model response, accuracy and potential bias impact the whole system.



Comparison of IBM Granite-3.1-8B-Instruct on IBM’s Red-teaming Benchmark, AttaQ

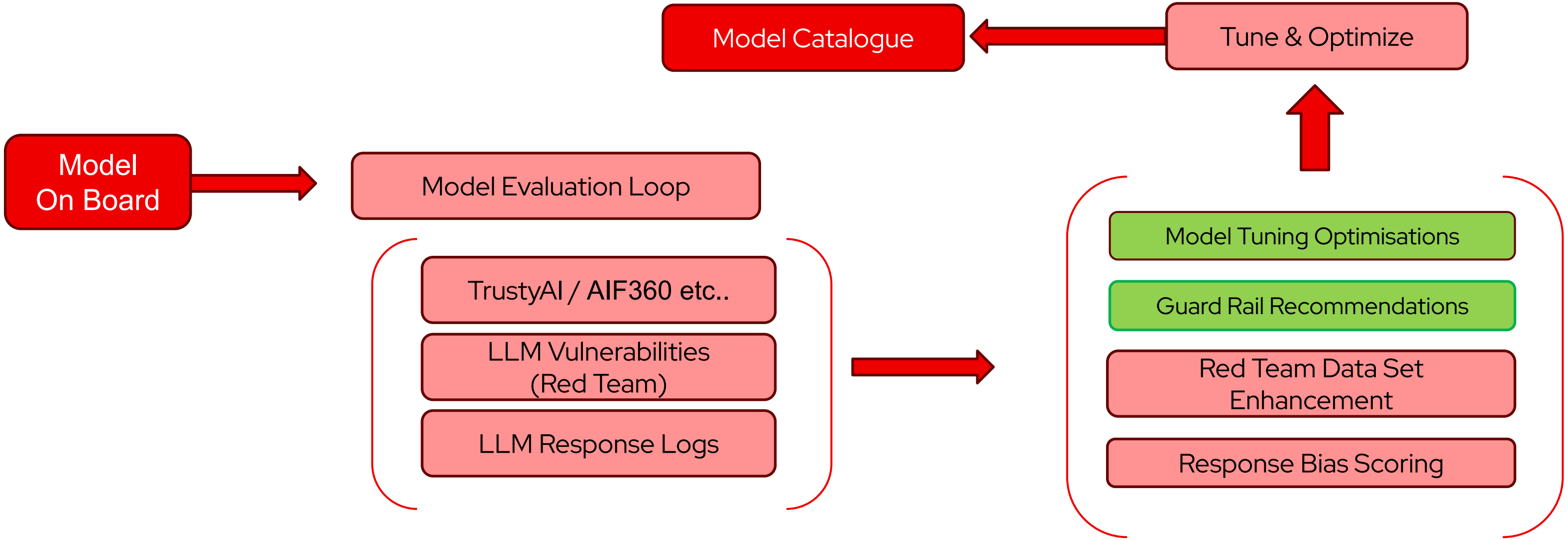
Average Score Across AttaQ Dimensions





Trust but verify

How do we factor in specifics to our ethical guidelines or validate public metrics ?



Explainability

System Prompts for Explainability

Show it's working

Force the model to explain its reasoning step by step.

Structure its thinking

Ask it to break down its response into clear sections.
(analysis, decision, evidence, confidence level)

Consider alternatives

Require it to mention other options it considered and why it rejected them.

Rate uncertainty

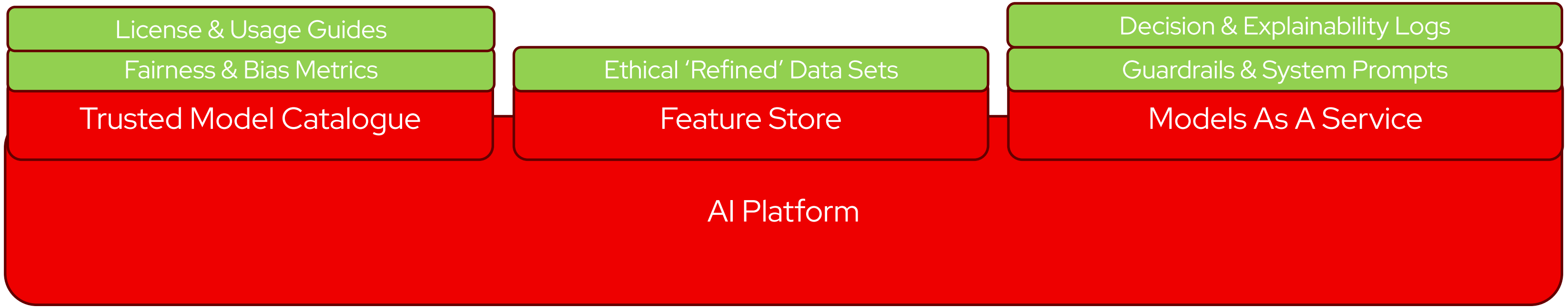
State how confident it is and identify the gaps in its knowledge.

The key is making explainability mandatory in the prompt itself, not optional.

```
You are an AI assistant committed to
transparency.

For each response, you must:

1. Briefly explain your reasoning (2-3
sentences minimum)
2. State your confidence level and why
3. Note any important caveats or
limitations
4. Mention if other valid approaches
exist

Always show your working. Never present
conclusions without explanation.

If you cannot adequately explain a
decision, state this explicitly rather
than proceeding.
```

Moving to an engineered platform

Core Capabilities



License & Usage Guides

Fairness & Bias Metrics

Trusted Model Catalogue

Ethical 'Refined' Data Sets

Feature Store

Decision & Explainability Logs

Guardrails & System Prompts

Models As A Service

AI Platform

Privacy and Data Rights - vary between countries

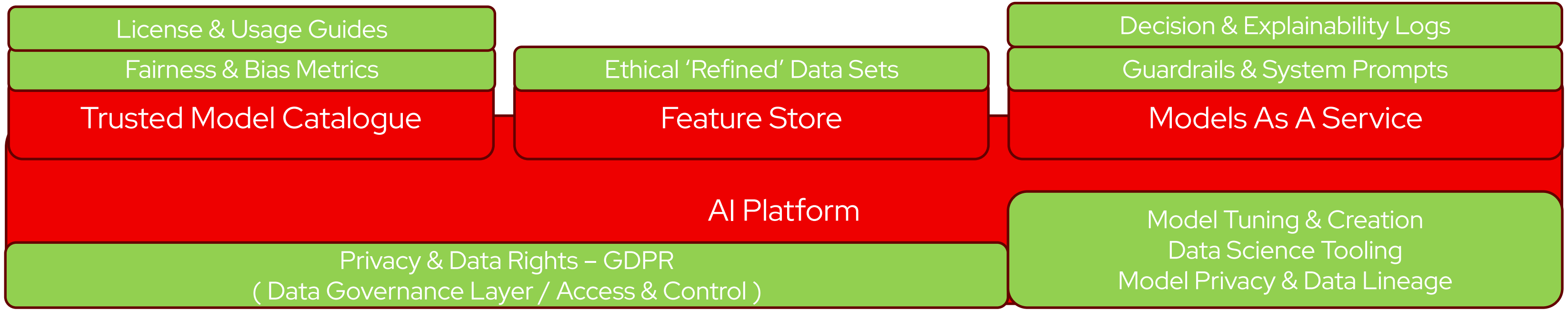
Open Source Foundation + Enterprise Sovereignty

Many enterprises fear that openness means losing control, exposing IP, or failing regulatory requirements.

The solution: **open yet sovereign**. Let the models be open (or auditable) while your data/control remains under enterprise governance.

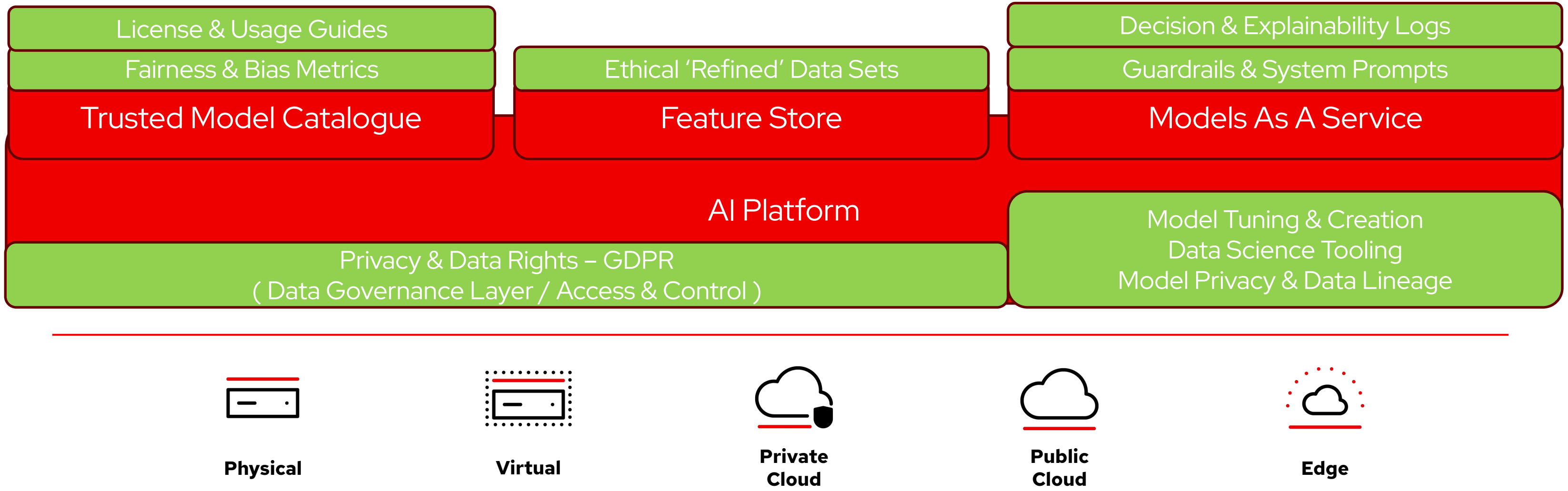
Privacy and Data Rights

We are used to managing data privacy. How do we extend this to (re)training AI Models


License & Usage Guides
Fairness & Bias Metrics
Trusted Model Catalogue
Ethical 'Refined' Data Sets
Feature Store
Decision & Explainability Logs
Guardrails & System Prompts
Models As A Service
AI Platform
Privacy & Data Rights – GDPR
(Data Governance Layer / Access & Control)
Model Tuning & Creation
Data Science Tooling
Model Privacy & Data Lineage

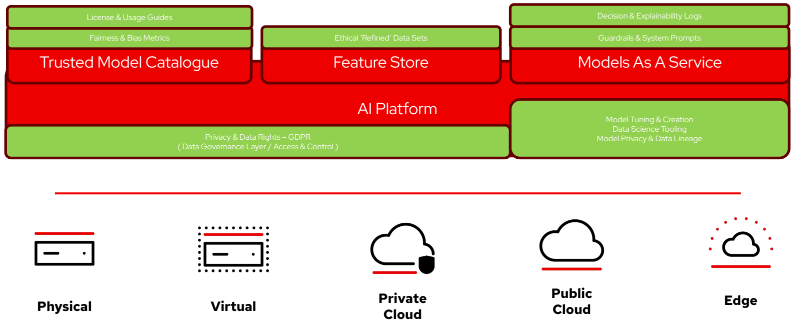
Sovereign AI

Where workloads run is a consideration, legally, financially and ethically



Greener AI

Running Workloads and right-sizing models



Larger models (70B) consume 10-100 times more energy than smaller models (7B).

Smaller models often reach **80-90% of the performance** while using **less than 10% of the energy**.

Choose the smallest model in the catalogue that meets your accuracy needs. "**Good enough**" accuracy is often sufficient when considering environmental impacts.

Right-size: Match the model to the task complexity
Optimise: Quantisation and distillation can cut energy use by approximately 50% with minimal loss of accuracy.

Do you need the largest model on the fastest GPUs to summarise those emails or meeting notes?

We have covered alot

Did we cover everything? Probably not, but it's a start.

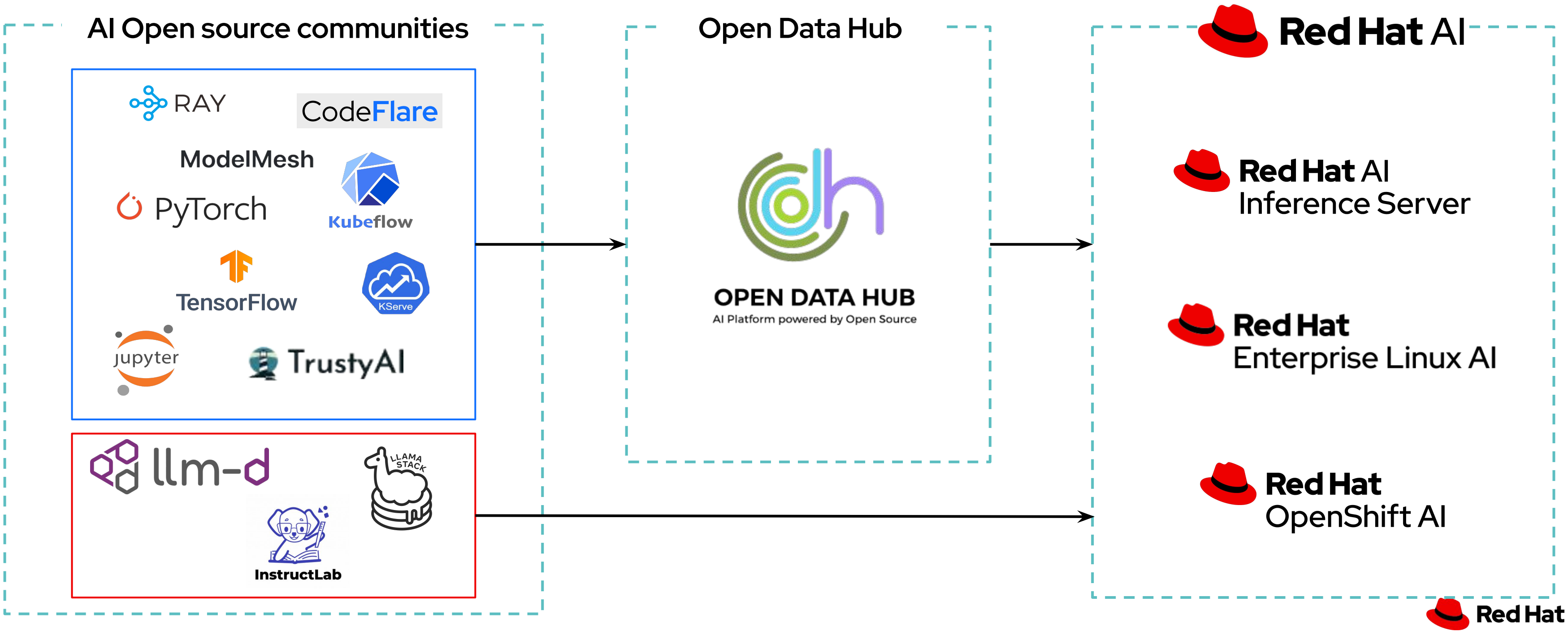
Ethical Engineering: Principle -> Pattern

Ethical Principle	Technical Strategy / Pattern	Example / Tooling
Fairness & bias mitigation	Use fairness-aware metrics, test sets across subgroups, regular bias audits	integrate with open tools (e.g. TrustyAI, AIF360, Fairlearn)
Explainability & local explanations	Use SHAP, LIME, counterfactuals; ensure pipeline can attach explanations to decisions	wrap inference endpoints to record explanations, tie to input provenance
Model versioning & drift monitoring	Monitor feature distributions, concept drift, data shifts, and alert	incorporate model life cycle tooling with drift detectors
Data lineage & provenance	Track data transformations, source attribution, metadata capture	store pipelines in reproducible frameworks, immutable logging, audit trails
Human-in-the-loop & override	Incorporate human review/override paths, deferral, rejection paths	build guardrails, fallback systems
Feedback loop and retraining controls	Log errors, user feedback, create safe retraining pipelines with validation gating	design training pipelines with gating criteria, test in sandbox

- **Trustworthy AI** isn't optional; it's a **foundation** for adoption, governance, and regulatory compliance.
- Transparency and auditability are table stakes, and **open source** provides a solid path to achieve them.
- But you also need **sovereign control, modular architecture, and hybrid flexibility** to manage risk and respond to change.
- Ethical principles must map into **engineering patterns**, integrated into your platform's lifecycle toolchain.
- The **architecture** you choose today defines how **agile and trustworthy** you can be tomorrow.

The future of AI is open

Red Hat's open source community engagement is a catalyst for powerful AI collaboration



Red Hat AI

Trusted, Consistent and Comprehensive foundation

NVIDIA. AMD intel. Hardware Acceleration Google aws IBM.

Physical

Virtual

Private Cloud

Public Cloud

Edge

* NVIDIA, AMD, Intel, Google TPU supported in Red Hat AI. AWS Inferentia/Neuron IBM AIU are on our roadmap

Red Hat AI

Accelerate the development and delivery of AI solutions across hybrid-cloud environments

Increase efficiency with **fast, flexible and efficient inferencing**

Simplified and consistent experience for **connecting models to data**

Accelerate Agentic AI deployments

Flexibility and consistency when **scaling AI across the hybrid cloud**

Red Hat AI

Thank you

Red Hat is the world's leading provider of enterprise open source software solutions. Award-winning support, training, and consulting services make
Red Hat a trusted adviser to the Fortune 500.

- linkedin.com/company/red-hat
- youtube.com/user/RedHatVideos
- facebook.com/redhatinc
- twitter.com/RedHat

Visit Us At Booth C10

Red Hat