RAGMETRICS.AI

The GenAI Evaluation Playbook

Building Measurable Trust in Enterprise LLMs

RagMetrics

The Author

HELLO THERE! I'M OLIVIER.

With over 15 years of experience building and scaling global SaaS and AI companies, I've dedicated my career to bridging innovation and reliability.

As Founder and CEO of RagMetrics, I lead a team working to make Generative AI systems measurable, auditable, and trustworthy for enterprises worldwide.

I've helped technology companies expand across Europe, the U.S., and Israel—driving growth, fundraising, and partnerships that turn emerging technologies into real business results. Before RagMetrics, I co-founded Libertify, an AI startup for financial intelligence, and held leadership roles at Chyron, leading global sales, marketing, and channel development.

I'm passionate about helping organizations take control of their AI systems to build trust, transparency, and compliance into every model they deploy.

email: olivier@ragmetrics.ai

Foreword: From Probabilities to Proof

BUILD THE SYSTEMS FOR AI TRUST

The most powerful LLMs today can generate poetry, code, and insight — but not certainty. They reason, summarize, and decide while remaining fundamentally probabilistic. The smarter the model, the harder it becomes to verify.

Enterprises racing to integrate GenAI face a new bottleneck: measurement. We know how to track latency and throughput, but we're flying blind on truthfulness, reasoning quality, and contextual adherence.

"You can't trust what you can't measure."

That statement defines the frontier of enterprise AI.

A 2024 study by Yuxia Wang who wrote Factuality of Large Language Models in 2024 found that as prompt complexity increases, **factuality in even top-tier models can decline by 30%** — especially in long-context settings. DeepMind's FACTS benchmark showed hallucinations persist in RAG pipelines unless continuously monitored.

This mirrors a larger enterprise truth. McKinsey's State of AI in 2025 reports that **88% of organizations now use AI in at least one business function,** yet two-thirds are still stuck in pilot mode. The barrier isn't adoption, it's accountability.

And the consequences are visible. McKinsey found **over half of enterprises have experienced at least one negative outcome from AI,** most commonly inaccuracy or explainability gaps.

Enter **RagMetrics** — a framework that embeds trust directly into GenAI workflows. It's not about grading outputs after the fact; it's about engineering **measurable reliability** into every stage of AI deployment.

This playbook outlines that foundation: the systems, standards, and infrastructure that make trust in AI quantifiable.

CHAPTER 1

The Evaluation Imperative

1.1 From Prototype to Production

By 2025, enterprises had discovered that deployment without evaluation equals risk.

McDonald's learned this firsthand when its AI-powered drive-thru ordering system, built with IBM, was suspended after customers reported bizarre misorders and unfiltered mistakes. The problem wasn't creativity, it was calibration.

These failures aren't isolated. Hallucinations erode customer trust. Schema violations break downstream integrations. Context drift compromises factuality. And opacity limits compliance.

McKinsey calls this the "trust bottleneck": only one in three organizations has successfully scaled AI, and those that did built formal governance and validation into their operating model.

OpenAI (2025) observed that current evaluation setups often incentivize LLMs to "guess confidently," amplifying hallucinations instead of suppressing them.

65% of business leaders say AI hallucinations directly undermine customer trust. Unverified AI isn't just a technical risk — it's a brand liability.

1.2 The Evaluation Gap

Most enterprises still evaluate AI like software — too little, too late. Human review doesn't scale. Benchmark tests miss domain nuances. Observability tells you *how fast* your model replied, not *whether it was right*.

Offline QA misses drift while black-box scoring breaks traceability. As Mondorf & Plank (2024) note, in Beyond Accuracy: Evaluating the Reasoning Behavior of LLMs, most benchmarks "miss the reasoning process entirely," leaving no audit trail for decisions.

With no audit trail for decisions, AI trust becomes anecdotal and enterprises can't prove reliability when it matters most.

CHAPTER 2

The Case for Evaluation Infrastructure

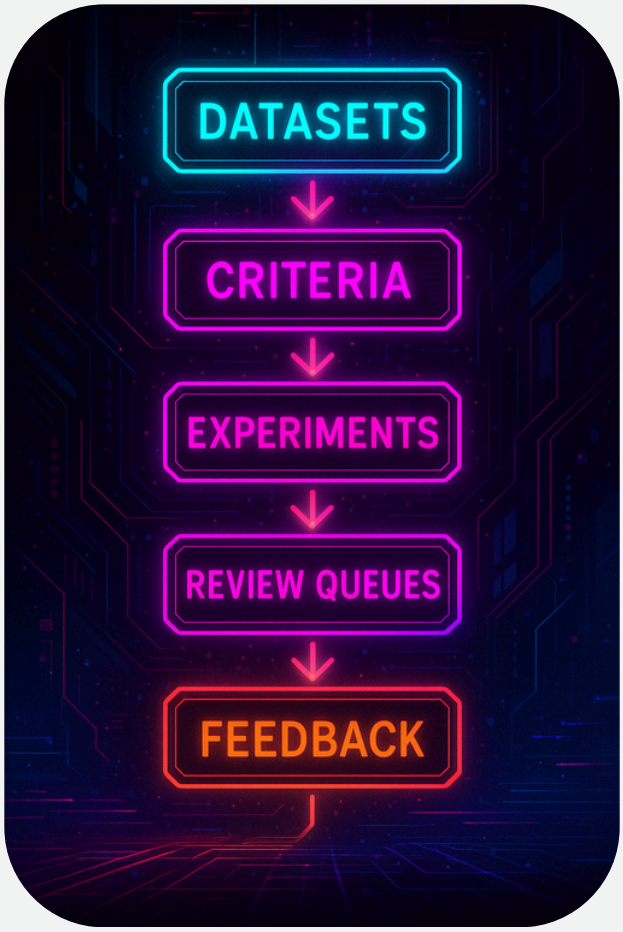
2.1 What Is Evaluation Infrastructure?

Evaluation infrastructure is the next evolution of DevOps. Just as DevOps brought rigor to software delivery, EvaluationOps brought rigor to AI reliability.

It's not a single tool, it's a system that connects datasets, criteria, experiments, and real-time monitoring into one feedback loop.

RagMetrics operationalizes that loop with APIs and automation, unifying labeled data, programmable evaluation rubrics, and scoring workflows under one traceable model.

Companies that embed AI directly into workflows — supported by governance and measurable validation — **capture significantly more value and innovation impact** than those that don't.



Building feedback loops into datasets is key.

2.2 The Role of LLM-as-a-Judge

The **LLM-as-a-Judge** paradigm marks a breakthrough in scalable evaluation. Instead of relying on human reviewers alone, AI systems can now assess one another using defined rubrics for factuality, tone, or compliance.

AWS (2025) found **98% alignment between automated evaluations and human scorers**, while cutting costs and latency dramatically. DeepMind's FACTS benchmark showed even greater reliability when using multi-model ensembles — combining Gemini, GPT-4, and Claude to balance bias.

McKinsey Senior Fellow Michael Chui put it bluntly:

"***When it comes to agents, it takes hard work to do it well.***"

Despite this, only **23% of organizations report scaling agentic systems beyond pilots**. LLM-as-a-Judge doesn't mean letting AI police itself. It means embedding consistent, auditable judgment into workflows, enabling explainable, repeatable measurement at scale.

CHAPTER 3

Designing Evaluation Loops

3.1 Criteria: The DNA of Quality

Anthropic (2024) found rubric-based evaluation reduced hallucinations by 40% during prompt iteration cycles.

Every evaluation starts by defining "good." RagMetrics ships with 200+ criteria covering accuracy, coherence, structure, and tone. Every score includes a rationale, turning evaluation from guesswork into governance.

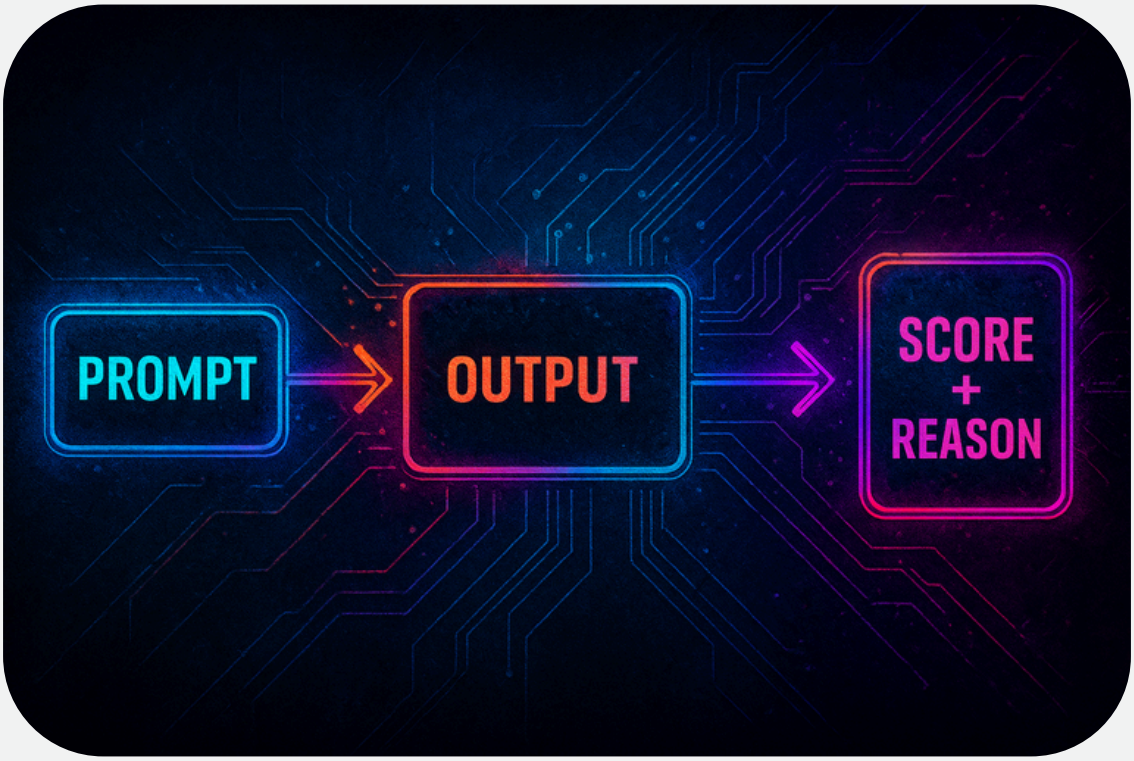
Based on our research, **AI high performers are three times more likely to redesign workflows around measurable evaluation loops.**

These criteria become programmable trust definitions — customized for each domain, department, or compliance framework.

3.2 From A/B to Always-On

Traditional A/B tests can't keep up with evolving models. Prompts change, data shifts, context expands.

Continuous experimentation makes quality a moving target, and RagMetrics tracks it. Compare GPT-4 to a fine-tuned Llama model, monitor score drift across accuracy and tone, and preserve every run in a historical record.



Continuous experimentation and scoring is the key to success.

CHAPTER 4

From Experiments to Enterprise Monitoring

4.1 APIs that Integrate Everywhere

Evaluation must live where your models live. RagMetrics provides both Push and Pull APIs to log model traces and evaluate them continuously.

All evaluations feed into a **Trace Store,** a versioned ledger of every input, output, and score, spanning dev to production.

Based on our testing, companies that scale AI successfully do so by **embedding evaluation and compliance into production pipelines**, not running them as separate checklists.

4.2 Review Queues: Live Guardrails

Evaluation doesn't end at testing. RagMetrics' Review Queues enable human-in-the-loop governance for live systems. If an output fails criteria, it's flagged, routed, and reviewed — closing the feedback loop.

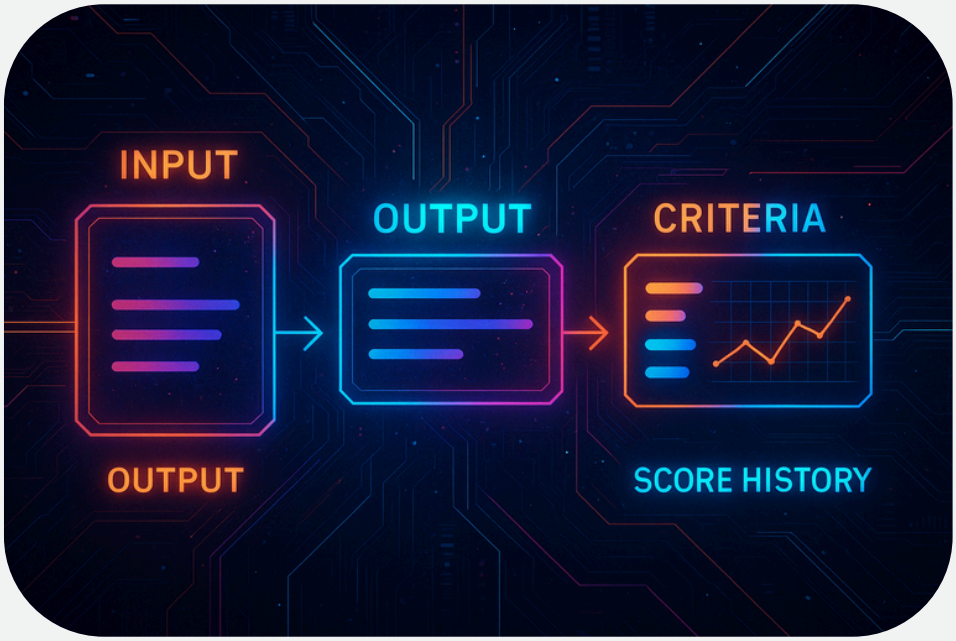
Over time, reviewed data becomes new training fuel, creating a self-reinforcing safety system that evolves with each model release.

> *"The combination of AI systems with human judgment creates the hybrid intelligence that delivers real value."*
> *Bryce Hall - McKinsey*

4.3 Auditable by Design

Every evaluation is fully traceable. Each entry logs context, version, criteria, and reasoning — providing replayable transparency.

For auditors, compliance officers, or internal QA, RagMetrics produces verifiable evidence that your models behave as intended.



"Review Queues helped us catch a 10% prompt regression before it reached users."
-RagMetrics Customer.

CHAPTER 5

Continuous Evaluation as Governance

5.1 Why Static QA Fails

LLMs evolve too quickly for one-time validation. Model updates, data changes, and prompt drift quietly erode quality.

Continuous Evaluation turns QA into a control loop. Models are scored continuously — each new output feeding back into reliability analytics.

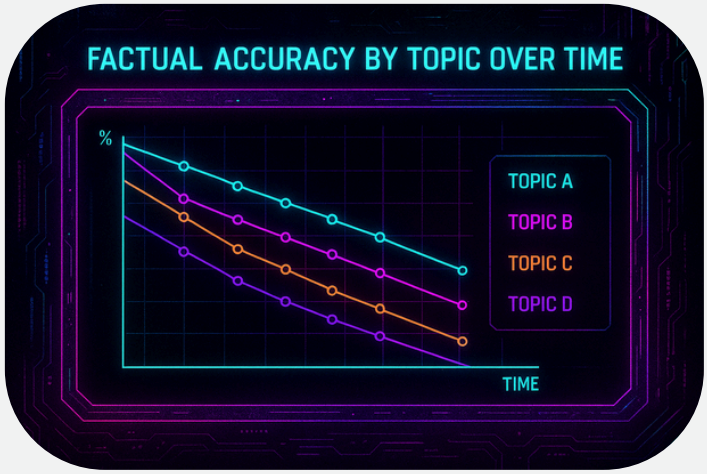
Imagine an LLM-as-a-Judge agent automatically auditing live model responses in real time. That's evaluation as infrastructure.

5.2 Detecting Drift

Drift — subtle degradation over time — is the silent killer of GenAI performance.

A tone change can undermine helpfulness. New documentation can corrupt retrieval results. Minor model updates can spike hallucination rates.

RagMetrics continuously tracks these shifts. Trend dashboards flag anomalies before they hit production.



5.3 Evaluation Becomes Policy

Once evaluation becomes continuous, it evolves into policy.

- **Product teams** justify changes with evidence.
- **Compliance teams** verify fairness and consistency.
- **Executives** track reliability through dashboards.

This turns AI governance from a checkmark into a competitive advantage.

Leading enterprises integrate RagMetrics directly into CI/CD pipelines, gating releases on quality thresholds.

CHAPTER 6

The RagMetrics Advantage

MORE ON RAGMETRICS.AI

6.1 Why This Matters Now

LLMs have moved beyond labs and prototypes. They're generating product recommendations, screening applicants, assisting with legal summaries, and making critical financial decisions. The margin for error is shrinking, and the demand for explainability is growing.

Trust isn't just about accuracy. It's about knowing when something changed, why it changed, and whether it's still safe to ship. That requires infrastructure. RagMetrics provides that infrastructure: not just tools for scoring, but a system for learning from those scores.

Without this layer, teams are left guessing. With it, they're managing risk, accelerating iteration, and building systems that improve with every output.

The future of AI belongs to those who can prove their models are right. It's not about the fastest model — it's about the most verifiable one.

6.2 Defensibility

RagMetrics is purpose-built for enterprises that can't afford uncertainty. Its architecture integrates testing and monitoring into a single data model. Its criteria engine is fully programmable and explainable. Every score is traceable, and every evaluation contributes to the system's improvement.

Unlike toolkits or spreadsheets, RagMetrics scales with your ambitions—and evolves alongside the models you deploy.

Conclusion: Evaluation Is the Foundation of Trust

You can't trust what you can't measure. RagMetrics transforms uncertainty into structured feedback, and experiments into governed pipelines. In a world where AI can do more than ever, confidence is the new currency.

Next Steps:
Define what "good" means for your GenAI use case. Label a dataset. Select your criteria. Run an experiment. Set up monitoring. Track trust in real time.

Get Started → ragmetrics.ai

.

Appendix: Reference Highlights

- Wang et al. (2024) – Factuality decline in long-context prompts
- DeepMind FACTS (2024) – LLM-as-a-Judge with ensemble scorers
- AWS Bedrock (2025) – Scaled evaluation + compliance with LLM judges
- OpenAI (2025) – Incentive structures encourage hallucination
- Mondorf & Plank (2024) – Current benchmarks miss reasoning quality

.

Ready to get started?

Start testing now!

CLICK FOR A DEMO