zensar

ZenseAI.AssureAI

Building Trust in Probabilistic Systems

A framework for assuring AI quality across the full life cycle — from data pipelines through agentic deployment — aligned to NIST AI RMF, ISO/IEC 42001, and the EU AI Act

White Paper

An RPG Company

Table of Contents

Executive Summary

Artificial intelligence is no longer a future aspiration for enterprise technology teams — it is a present-day operational reality. Large language models, retrieval-augmented generation pipelines, and autonomous agentic systems are being woven into customer-facing products, regulated workflows, and mission-critical decision engines. The speed of this adoption, however, has dramatically outpaced the maturity of the quality assurance practices meant to govern it.

Traditional software testing was designed for deterministic systems: given the same input, a well-written function returns the same output every time. AI systems do not work this way. They are probabilistic, context-sensitive, and capable of failing in ways that are subtle, hard to reproduce, and potentially consequential. A hallucinated fact delivered with perfect grammatical confidence, a retrieval system that returns plausible but irrelevant passages, an agentic workflow that executes an unintended tool call — none of these failures would be caught by a conventional regression suite.

This white paper introduces ZenseAI.AssureAI service offering: a purpose-built assurance practice that addresses this gap. The framework spans the full AI life cycle through four integrated pillars — Data Quality Assurance, Functional and Model Evaluation, Trustworthy Assurance, and Non-Functional Assurance. It is grounded in the leading governance frameworks of the day: the NIST AI Risk Management Framework, ISO/IEC 42001, and the EU AI Act. Further, it is operationalized through a catalog of 30 structured tests, proprietary accelerator tooling, and a tiered evaluation methodology that blends deterministic checks, LLM-as-judge assessments, and human annotation.

Industry research suggests that 70% of enterprises will have deployed AI systems by 2027, yet fewer than 15% currently operate a formal AI testing framework. That gap represents both a significant enterprise risk and an urgent market need.

The sections that follow make the case for why this approach is necessary, how it is structured, what it covers, and why it delivers measurable value — across regulated industries, agentic AI deployments, and the expanding frontier of generative AI.

1

Why Traditional Software Testing Fails in the Age of LLMs

The enterprise software industry has spent decades refining its testing discipline. Agile practices, continuous integration pipelines, automated test suites, and shift-left methodologies have collectively raised the bar for software quality. Yet all of these practices share a foundational assumption: the system under test behaves predictably. Given the same code path and the same inputs, the output is known in advance.

AI changes that assumption entirely. A large language model (LLM) is not a function — it is a probability distribution over token sequences. Its outputs depend on training data that may be outdated, incomplete, or biased. Its behavior can shift with minor changes to prompt phrasing, context window content, or temperature settings. It can generate responses that are grammatically flawless and factually wrong, confident and hallucinated, policy-compliant in testing and policy-violating in production. These are not edge cases — they are inherent properties of the technology.

The regulatory horizon

Regulators have noticed. The EU AI Act, which entered force in August 2024, introduces binding obligations for high-risk AI systems — including requirements for data quality controls, technical documentation, human oversight mechanisms, robustness, and cybersecurity standards. [1]

The NIST AI Risk Management Framework 1, while voluntary in the United States, has rapidly become the de facto reference architecture for enterprise AI governance, articulating that testing, evaluation, verification, and validation should occur throughout the life cycle — not only before release. [2]

ISO/IEC 42001 adds a certifiable management system layer, giving procurement teams and auditors a structured lens through which to evaluate whether an organization has genuinely institutionalized responsible AI practices rather than merely documenting them. [3]

Market-wide quality gap

Despite this regulatory urgency, the quality gap is real and widening. The challenge is not that organizations lack intent — most have published responsible AI principles and have teams thinking about governance. The challenge is operationalization. How do you test whether a retrieval-augmented generation (RAG) pipeline is retrieving faithfully? How do you measure whether an agentic system respects its authorization boundaries when given an ambiguous instruction? How do you demonstrate to a regulator that your model's demographic performance is equitable? These are not questions that existing software quality toolchains can answer.

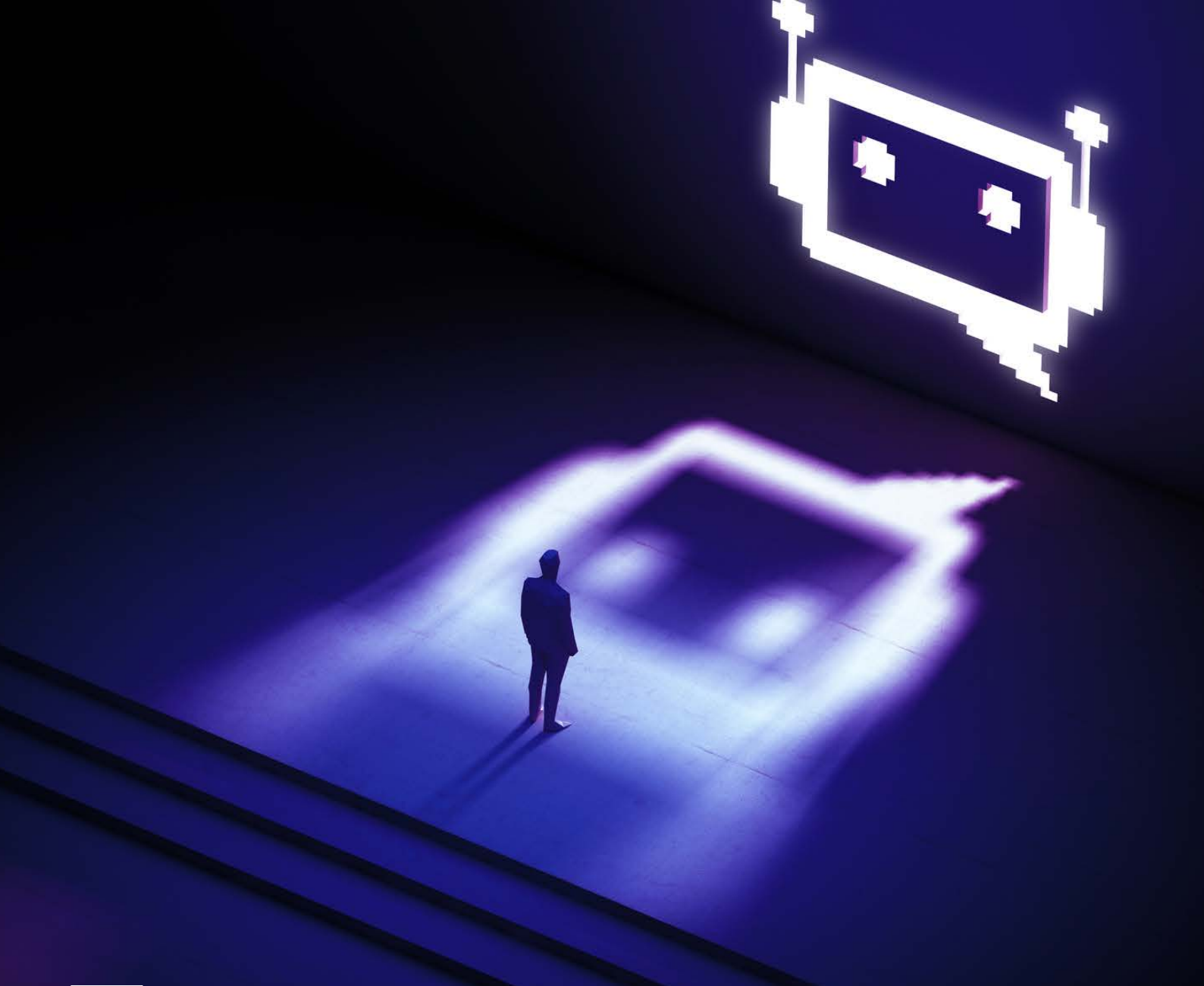
Research across healthcare, automotive, financial services, and software engineering confirms that quality practice for AI converges on several shared themes: life cycle governance rather than point-in-time testing, risk-based validation tied to use-case criticality, continuous monitoring after deployment, and alignment with emerging international standards.

45

What has been missing is a coherent, end-to-end service practice that brings these themes together in a form that can be applied to a real enterprise deployment with real timelines and real risk tolerance.

Domain	Primary quality focus	Typical methods	Regulatory trend
Healthcare	Clinical validity and impact	Internal/external validation, clinical impact assessment	Pre-deployment acceptance testing and post-deployment QC
Finance	Risk, accountability, robustness	Model risk management, traceability, governance controls	Stronger compliance, auditability, risk-based regulation
Automotive	Functional safety and scenario coverage	Simulation, scenario-based testing, runtime monitoring	Active standardization; regulatory focus on AV certification
Software/AI	Non-functional quality of AI-enabled systems	Metamorphic testing, LLM evaluation, NFQC assessment	MLOps integration, ISO/IEC 25010 alignment, NFQC metrics

Table 1: AI quality priorities by domain (synthesis from cross-domain research)



2

The Real Cost of Ungoverned AI: Moving Past Traditional QA to Mitigate Strategic and Regulatory Risk

When organizations deploy AI without a structured quality framework, they are not simply accepting technical risk — they are accepting strategic, regulatory, and reputational risk. The failure modes of AI are qualitatively different from those of conventional software, and they tend to surface in ways that are difficult to anticipate and expensive to remedy after the fact.

Inadequacy of traditional quality assurance

A conventional QA team approaches a new feature by writing test cases against a specification. The system is expected to return value X when given input Y. If it does not, the test fails and the defect is raised. This model breaks down completely for generative AI in several dimensions:

Not deterministic: For a question-answering system, there is rarely a single correct answer. Testing frameworks that depend on exact-match comparison cannot evaluate whether a generated response is faithful, relevant, or safe.

Distributional failure: A model may perform well on the test set and fail systematically on a demographic subgroup, a linguistic variant, or a rare but important input pattern. Aggregate metrics conceal these failures.

Drift over time: AI systems that perform well at launch can degrade as the distribution of real-world inputs shifts away from the training distribution. A one-time pre-deployment test provides no signal about whether the system continues to perform acceptably six months later.

Adversarial vulnerability: LLMs are susceptible to prompt injection attacks, jailbreaks, and context manipulation in ways that have no analog in traditional software security testing. OWASP ranks prompt injection as the number-one risk for LLM applications.

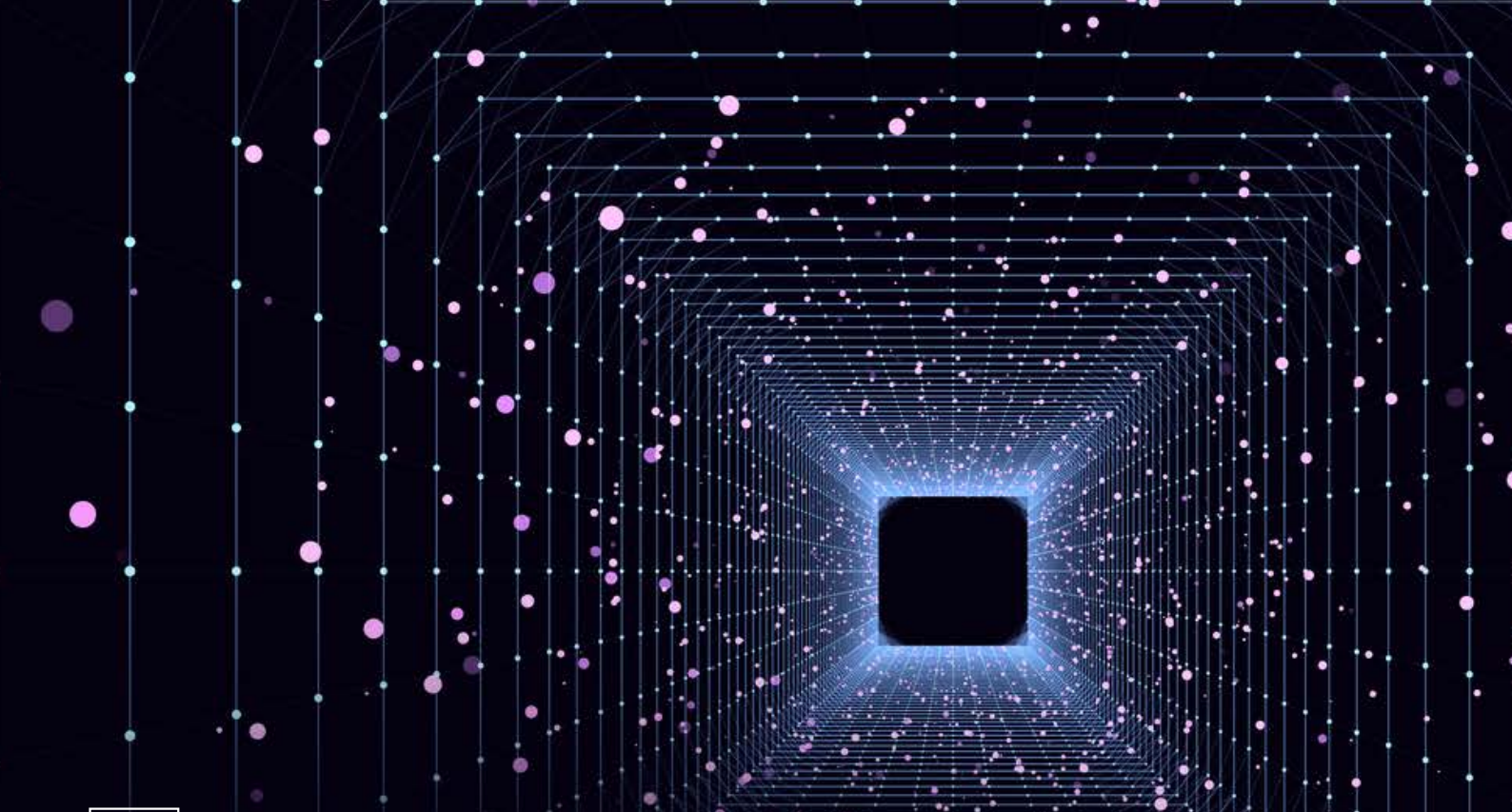
Agentic complexity: When AI systems are given tools — the ability to query databases, send emails, execute code, or call external APIs — the blast radius of a failure increases enormously. Testing multi-step reasoning chains and autonomous decision-making requires entirely different methodologies from testing a static API.

The compliance imperative

Regulatory exposure is not hypothetical. Under the EU AI Act, prohibited AI practices became enforceable in February 2025. Obligations for general-purpose AI models took effect in August 2025. High-risk AI system rules become fully applicable in August 2026. Organizations that have not built compliance-ready technical documentation, bias testing evidence, and ongoing monitoring plans are already behind. [6]

Similarly, the NIST AI RMF's Generative AI Profile identifies risk areas — confabulation, harmful bias, privacy violations, information integrity failures, and value chain integration risks — that are specific to large language model deployments and require dedicated testing methodologies to address. [7]

The core problem is not a lack of AI capability — it is the absence of a systematic, evidence-generating quality practice to ensure that AI capability is reliable, safe, fair, and compliant across the full deployment life cycle.



3

The ZenseAI.AssureAI Framework

Zensar's ZenseAI.AssureAI practice is organized around a four-pillar assurance framework. Each pillar addresses a distinct domain of AI risk and can be deployed independently, depending on client maturity, or as an integrated program for organizations seeking comprehensive coverage.

The pillars are designed to be life cycle-spanning — not confined to pre-deployment testing, but extending through production monitoring and governance assurance.

Pillar	Focus area	What it addresses
1. Data Quality Assurance	The foundation of all AI quality	Schema validation, bias audits, drift monitoring, RAG retrieval quality
2. Functional and Model Evaluation	What the model actually does	Hallucination detection, prompt robustness, jailbreak resistance, agentic tool-use validation
3. Trustworthy Assurance	Whether the system can be trusted	NIST AI RMF, ISO 42001, EU AI Act alignment; red-teaming, fairness, explainability
4. Non-Functional Assurance	How the system performs at scale	Latency benchmarking, cost-per-inference, resilience testing, security penetration

Table 2: ZenseAI.AssureAI

Zensar AI Quality Assurance

1. Data Quality

Ensuring the foundation is sound

- Data quality validation
- Bias/fairness audits
- Drift monitoring
- Synthetic data testing

2. Model Evaluation

Validating functional behavior and intent

- GenAI/LLM testing
- AI agent testing
- Prompt evaluation
- Multi-modal validation

3. Trustworthiness

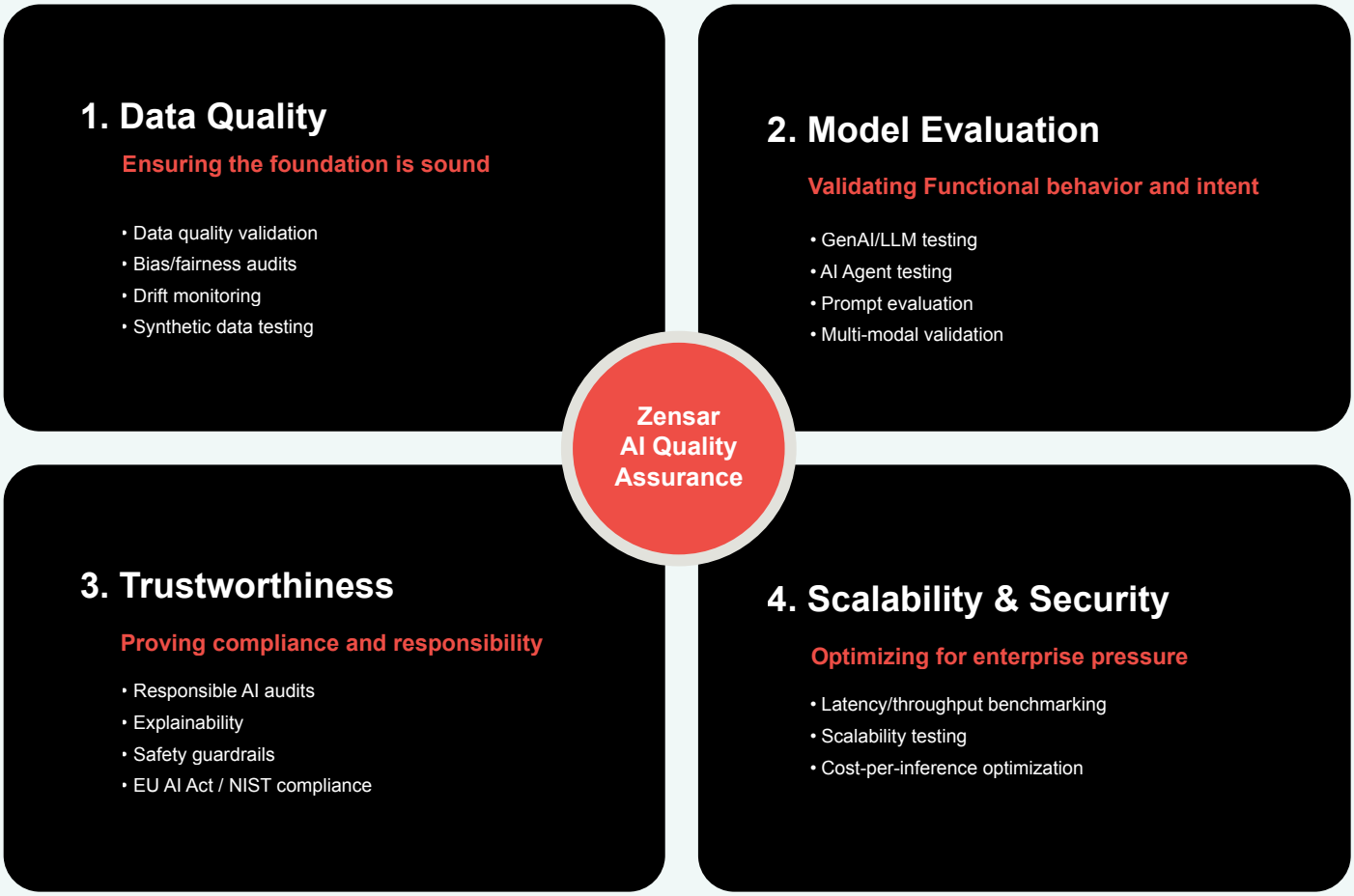
Proving compliance and responsibility

- Responsible AI audits
- Explainability
- Safety guardrails
- EU AI Act / NIST compliance

4. Scalability & Security

Optimizing for enterprise pressure

- Latency/throughput benchmarking
- Scalability testing
- Cost-performance optimization



Pillar 1:
Data Quality Assurance

Every downstream failure in an AI system can ultimately be traced to the data that trained it, the data that retrieves context for it, or the data that flows through it at inference time. Data Quality Assurance is therefore not merely the first pillar — it is the load-bearing foundation beneath the others.

Zensar's DataSentinel accelerator provides continuous validation across five dimensions: completeness, consistency, accuracy, timeliness, and bias representation. For retrieval-augmented generation systems specifically, context precision and context recall — as defined in the RAGAS framework — are used as operational metrics to determine whether the retrieval component is surfacing the right passages for the generation component. [8]

Data lineage and provenance verification ensure that the origin, transformation, and ownership of training and inference data can be traced end-to-end — a requirement that is increasingly mandated by the EU AI Act and referenced in the NIST AI RMF's updated guidance on model provenance. Privacy compliance testing covers PII detection, differential privacy budget tracking, and GDPR and HIPAA consent management, with explicit attention to the empirical finding that differential privacy training can worsen bias against underrepresented groups — requiring coordinated fairness monitoring when privacy mechanisms are applied.[9]

Pillar 2:
Functional and Model Evaluation

This pillar addresses the core question that most organizations ask first: Does the model actually work? The answer, for AI systems, requires more than a pass/fail accuracy rate.

Ensuring quality in AI systems demands a new class of observability infrastructure that goes well beyond traditional testing frameworks. ZenseAI.AssureAI for AI practice recommends a complementary tooling stack anchored on two platforms: Arize Phoenix and LangSmith.

Arize Phoenix, an open-source observability platform built on 10OpenTelemetry, provides end-to-end tracing of LLM application execution — surfacing latency bottlenecks, token-usage anomalies, retrieval failures, and hallucination-prone spans across the entire inference pipeline. Its LLM-as-a-judge evaluation library enables scalable, automated quality scoring across dimensions such as faithfulness, relevance, and toxicity, without requiring manual review at every stage of the life cycle.

LangSmith11 complements this by bringing structured evaluation discipline to the development and deployment workflow — enabling step-by-step assessment of agent tool calls, trajectory evaluation across multi-step reasoning chains, and dataset-driven regression testing to catch quality drift before it reaches production. Together, these platforms operationalize what Zensar terms eval-driven quality engineering — a shift from point-in-time testing to continuous, evidence-based quality assurance embedded across the AI application life cycle.

The evaluation methodology is deliberately tiered: deterministic, rule-based checks for structured outputs; LLM-as-judge scoring for open-ended generative assessments; and human annotation for high-stakes or ambiguous cases. This tiering ensures that the cost and latency of evaluation are proportionate to the risk of the use case.

Pillar 3:
Trustworthy Assurance

Trustworthy Assurance is where AI quality engineering intersects directly with governance, ethics, and regulatory compliance. It is also where most organizations have the largest and most consequential gaps.

Zensar's TrustScore methodology maps the assurance program to the four functions of the NIST AI RMF — govern, map, measure, and manage — translating each into concrete evidence artifacts: governance policies and accountability structures, system context and impact mapping, quantitative and qualitative testing evidence, and incident response and residual risk treatment plans.

Safety and red-teaming are treated as recurring controls rather than pre-launch exercises. Zensar's red-team methodology covers direct prompt injection and indirect injection (via external documents and web content), multi-lingual adversarial prompts, context hijacking, role-play manipulation, jailbreak attack families, and confabulation induction. Metrics include unsafe response rate by policy class, severity-weighted harm score per thousand prompts, jailbreak success rate by attack family, and regression delta after model or policy updates.

Fairness and bias testing go beyond single-axis demographic analysis to include intersectional audits across combined protected attributes, disparate impact analysis using the four-fifths rule and statistical significance testing, and proxy-variable effect analysis. Residual bias that cannot be fully remediated is documented, risk-accepted, and subjected to ongoing monitoring — a pattern consistent with both NIST AI RMF guidance and EU AI Act requirements.

Explainability validation verifies that explanations generated by methods such as SHAP and LIME are actually faithful to model behavior — not merely visually plausible. This includes stability testing across seeds and perturbations, explanation consistency across correlated features, and a user-facing review of explanation accuracy with domain expert panels.

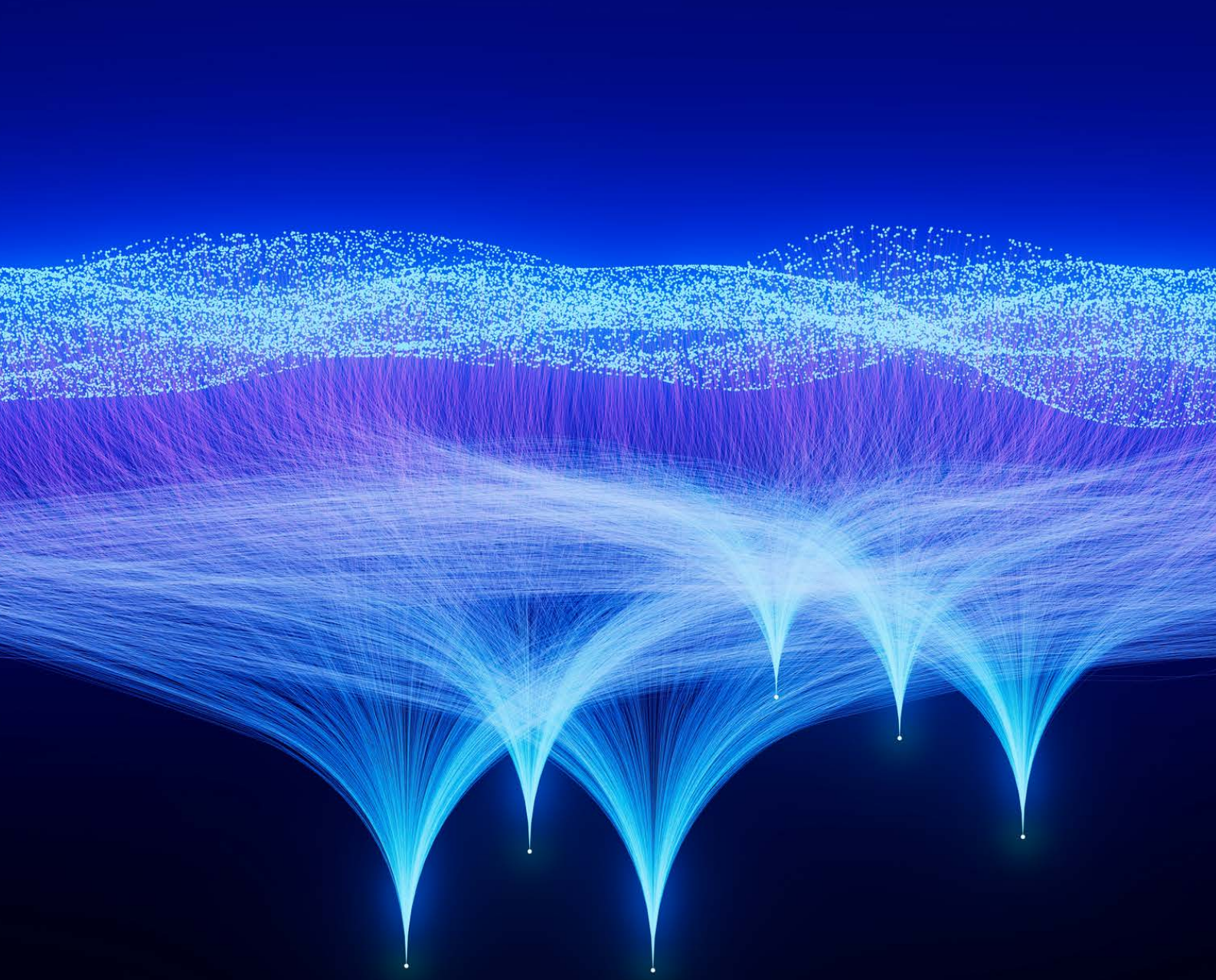
Pillar 4:
Non-Functional Assurance

AI systems that are accurate but unreliable under load, expensive beyond sustainable cost-per-inference thresholds, or vulnerable to adversarial API abuse are not production-ready regardless of their BLEU scores.
Non-Functional Assurance addresses the engineering dimensions of AI quality that determine whether a system can actually be operated at enterprise scale.

Performance benchmarking follows a sweep-analysis methodology: concurrent load testing with simulated user traffic, workload profiling across different prompt-to-completion ratios, and incremental request rate escalation to identify the knee of the latency-throughput curve. Key metrics include time to first token (TTFT), inter-token latency (ITL), tokens per second (TPS), and P95/P99 percentile latency — the tail behaviors that most affect user experience but are invisible in average-based reporting.

Resilience testing uses chaos engineering principles: injecting API failures, GPU out-of-memory conditions, and network partitions to validate that circuit breakers, graceful degradation patterns, and multi-provider routing configurations behave as designed. Cost-per-inference benchmarking compares managed API providers against open weights and self-hosted deployment options, quantifies the accuracy-versus-speed trade-off of quantization approaches, and measures semantic cache hit rates that reduce duplicate inference costs for common query patterns.

Security testing for AI systems extends traditional penetration testing with AI-specific threat vectors: prompt injection attacks using tools such as Garak and Giskard, PII exfiltration attempts with synthetic credential injection, model extraction attacks via API query patterns, data poisoning detection in training pipelines, and supply chain security review for third-party model components and ML libraries.



4

Benchmark Overview: 30 Tests across the Full AI Life Cycle

The four pillars are operationalized through a catalog of 30 structured assurance tests, each documented with a defined purpose, mitigation strategies for failure, available tooling (open-source and commercial), known shortcomings, and practical implementation guidance. They are organized into seven categories.

Category	Tests	Count
Data quality and integrity	e.g., Data Quality Assessment, Data Drift Detection, etc.	5
Model performance and reliability	Accuracy and Performance Benchmarking, Stress and Load Testing, Model Regression, etc.	5
Fairness and bias	e.g., Stereotype and Representation Testing	3
Safety and security	e.g., Prompt Injection, Toxicity and Harmful Content, Red-Teaming, etc	6
Gen AI-specific assurance	e.g., Hallucination Detection, Faithfulness and Groundedness, etc.	6
Explainability and transparency	Explainability and Interpretability Testing, Confidence Calibration Testing	2
Governance and compliance	Regulatory Compliance Testing, Model Documentation and Reproducibility Audit, Operational Monitoring and Observability	3
	Total	**30**

Table 3: 30 AI assurance tests across seven categories

Test design philosophy

Each test in the catalog was designed with three properties in mind. First, it must be executable — not an aspirational principle but a concrete procedure with defined inputs, methods, and measurable outputs. Second, it must be framework-mapped — every test is explicitly linked to one or more of the governance frameworks the client is expected to demonstrate compliance with. Third, it must be risk-calibrated — the depth of testing, the tooling employed, and the acceptance threshold applied are all scaled to the criticality of the use case and the potential harm of failure.

The catalog deliberately covers both pre-deployment validation and post-deployment monitoring. A common failure mode in enterprise AI programs is treating quality assurance as a gate rather than a practice — running a set of tests before launch and then assuming the system will continue to perform as validated. These 30 tests are designed to be run continuously, with automated monitoring for drift, performance degradation, and safety violations feeding back into a living evidence pack that supports ongoing governance reviews. 12

5

Key Findings from Research

The research underpinning Zensar's ZenseAI.AssureAI framework synthesizes findings from peer-reviewed literature in healthcare AI governance, automotive safety assurance, cross-domain standards development, software engineering quality, and large language model evaluation. Several findings are particularly consequential for practitioners designing enterprise AI quality programs.

Quality frameworks are converging on life cycle governance

Across domains, research consistently identifies that point-in-time testing is insufficient for AI systems. Healthcare AI guidance emphasizes six life cycle phases — data preparation, model development, validation, software development, impact assessment, and clinical implementation — with quality obligations at each stage.13

Automotive safety assurance frameworks similarly identify runtime monitoring and continuous observability as essential complements to pre-deployment scenario-based testing.14

Cross-domain frameworks are converging on ISO/IEC alignment, MLOps traceability integration, and risk-based governance as the common architecture for responsible AI quality practice. 15

Hallucination can be substantially reduced

Recent research has demonstrated that hallucination rates in LLMs can be reduced by 90 to 96% through targeted synthetic example training — a finding that has significant implications for production RAG deployments where factual accuracy is a requirement rather than a preference.16

However, hallucination detection itself remains an imperfect art: detection systems that rely on LLMs-as-judges are themselves susceptible to hallucination, and current automated metrics show limited correlation with human judgment on subtle factual errors. This reinforces the need for tiered evaluation methodologies that incorporate human review for high-stakes applications.

Differential privacy trades off against fairness

Empirical research has documented a troubling interaction between privacy-preserving training techniques and fairness outcomes. Differential privacy applied during fine-tuning has been shown to worsen AUC-based bias metrics for underrepresented demographic groups in LLMs.17

This means that privacy and fairness cannot be optimized independently — organizations that apply differential privacy to satisfy regulatory privacy requirements must conduct coordinated fairness audits to ensure that privacy protection is not achieved at the expense of equitable performance.

Prompt injection remains an unsolved problem

OWASP ranks prompt injection as the number-one security risk for LLM applications, and research confirms that no complete defense currently exists. New attack vectors continue to emerge as models and deployment patterns evolve, and indirect injection — where malicious instructions are embedded in external content that the model processes — is particularly difficult to detect and defend against. This finding argues for defense-in-depth architectures that combine input validation, privilege separation, output filtering, and human-in-the-loop controls for high-stakes actions, rather than relying on any single mitigation.

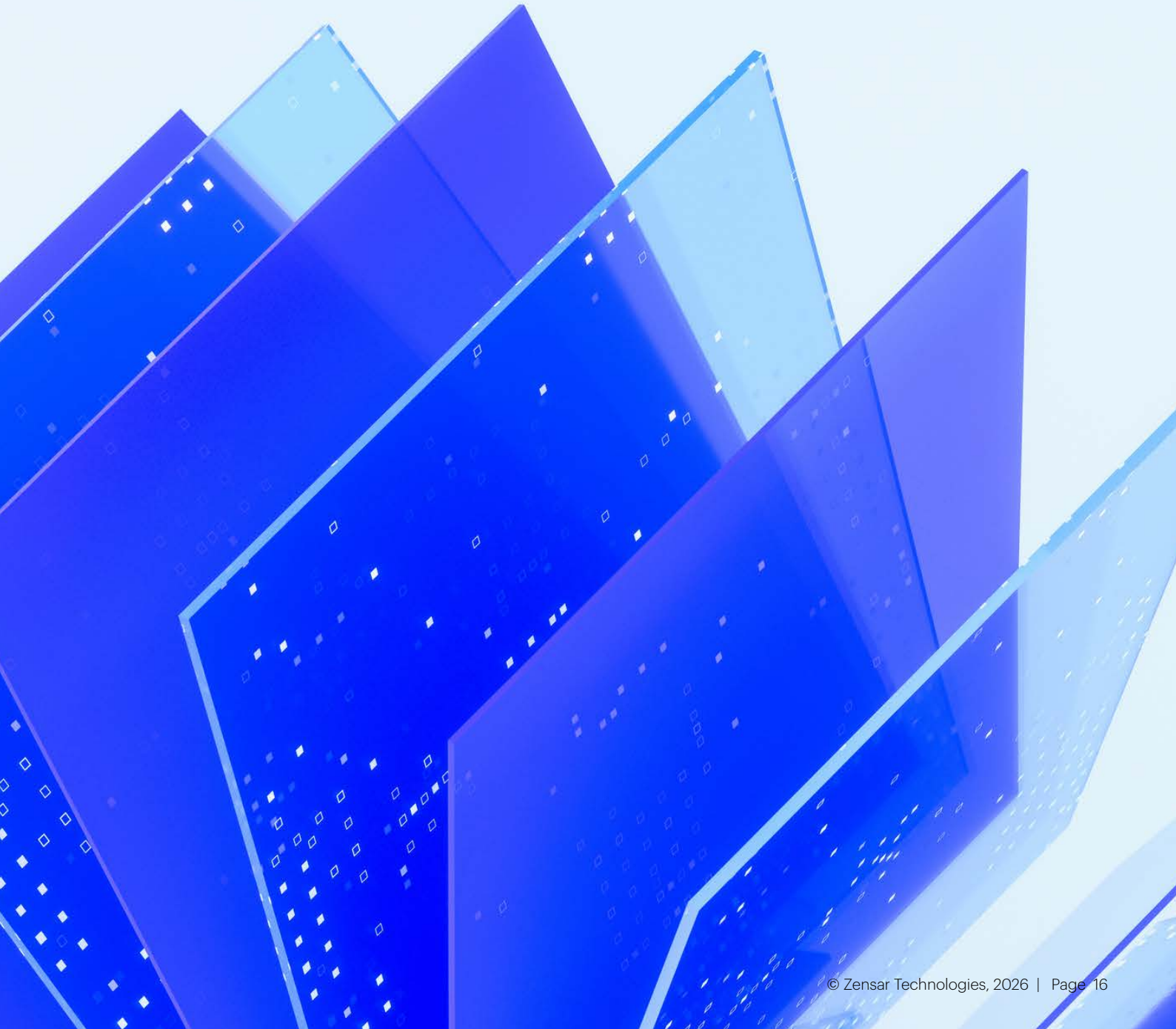
Documentation gaps are the leading cause of audit failure

Organizations often fail responsible AI audits not because they conducted no testing, but because their testing was fragmented and not traceable to governance decisions. Common weaknesses include missing use-case boundary definitions, absence of subgroup fairness analysis, unverifiable explanation claims, weak versioning of red team findings, and incomplete technical documentation for external review. The implication is that quality assurance must be documentation-first: every test must produce evidence artifacts that a regulator, auditor, or procurement reviewer can actually inspect.18

6

Technology Stack

ZenseAI.AssureAI framework is tool-agnostic by design — it specifies what must be tested and what evidence must be generated, leaving the tooling layer open for client preference and integration constraints. That said, Zensar has developed preferred tool stacks for each testing domain, combining established open-source frameworks with proprietary accelerators.



Domain	Open-source tooling	Zensar accelerator
RAG and LLM evaluation	Arize Phoenix , DeepEval, RAGAS, Promptfoo,	ZenseAI.QI
Agentic AI testing	LangSmith, Shap	ZenseAI.QI
Safety and red-teaming	Arize Phoenix, MAIA, DeepEval, LangGraph, CrewAI, LangSmith	ZenseAI.QI Red-Team Module
Data quality	Garak (NVIDIA), PyRIT (Microsoft), Giskard	ZenseAI.Data
Performance testing	Great Expectations, Evidently AI	ZenseAI.QI
Privacy and compliance	MLflow, Kubeflow, Grafana, Datadog	Non-Functional Assurance Suite
Observability and drift	Presidio (Microsoft), IBM Diffprivlib, OpenDP	ZenseAI.QI Compliance Module
	Evidently AI, OpenTelemetry, Prometheus + Grafana	ZenseAI.Data Monitoring

Table 4: ZenseAI.AssureAI technology stack

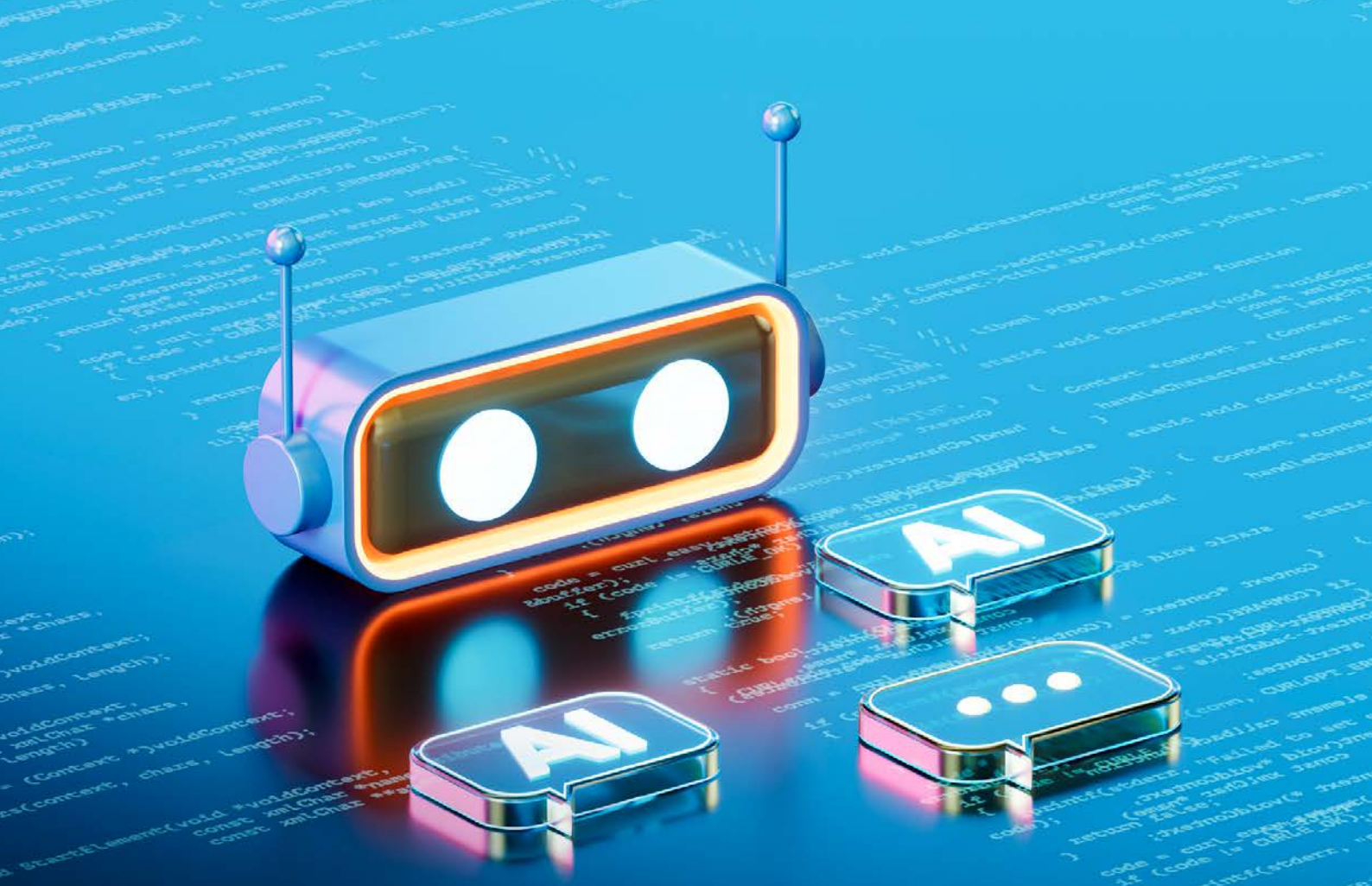
Evaluation methodology

Zensar's evaluation methodology is structured in three tiers, applied according to the risk profile of the use case:

Tier 1 — Deterministic checks: Rule-based validation for structured outputs, schema conformance, format compliance, and safety guardrail adherence. Fully automated, runs in CI/CD.

Tier 2 — LLM-as-judge: Model-evaluated assessments of open-ended generative quality, faithfulness, relevance, and instruction-following. Scalable and reference-free, with known evaluator bias as a managed risk.

Tier 3 — Human annotation: Expert and domain-specific human review for high-stakes, regulated, or ambiguous cases where automated evaluation is insufficient. Provides the ground truth that calibrates the other two tiers.



7

Use Cases

ZenseAI.AssureAI for AI framework is applicable across industries and deployment architectures, but two use-case patterns have emerged as particularly high-value initial applications: RAG-based enterprise knowledge assistants and agentic workflow automation.

RAG-based enterprise knowledge assistants

Retrieval-augmented generation assistants — where a user query triggers a search against an enterprise knowledge base, and the retrieved content is used to ground the model's response — are among the most widely deployed production LLM applications. They appear in insurance claims processing, financial services research, legal contract review, and healthcare clinical decision support.

The quality risks in RAG systems cluster around three failure modes: retrieval failure (the wrong documents are returned), generation failure (the model hallucinates or misrepresents the retrieved content), and contextual failure (the response is technically grounded but irrelevant to the user's actual intent). Zensar's RAG Retrieval Quality Testing (Test 23) and Faithfulness and Groundedness Testing (Test 21) address these failure modes directly, using RAGAS-based metrics for context precision, context recall, faithfulness, and answer relevance.19

In a regulated industry pilot applying the ZenseAI.AssureAI framework to a RAG-based insurance assistant, the evaluation surfaced a context precision gap where retrieved passages were topically related but not sufficiently specific to the query — a pattern that would cause the model to generate responses that sounded authoritative but missed the precise policy language the user needed.

Agentic workflow automation

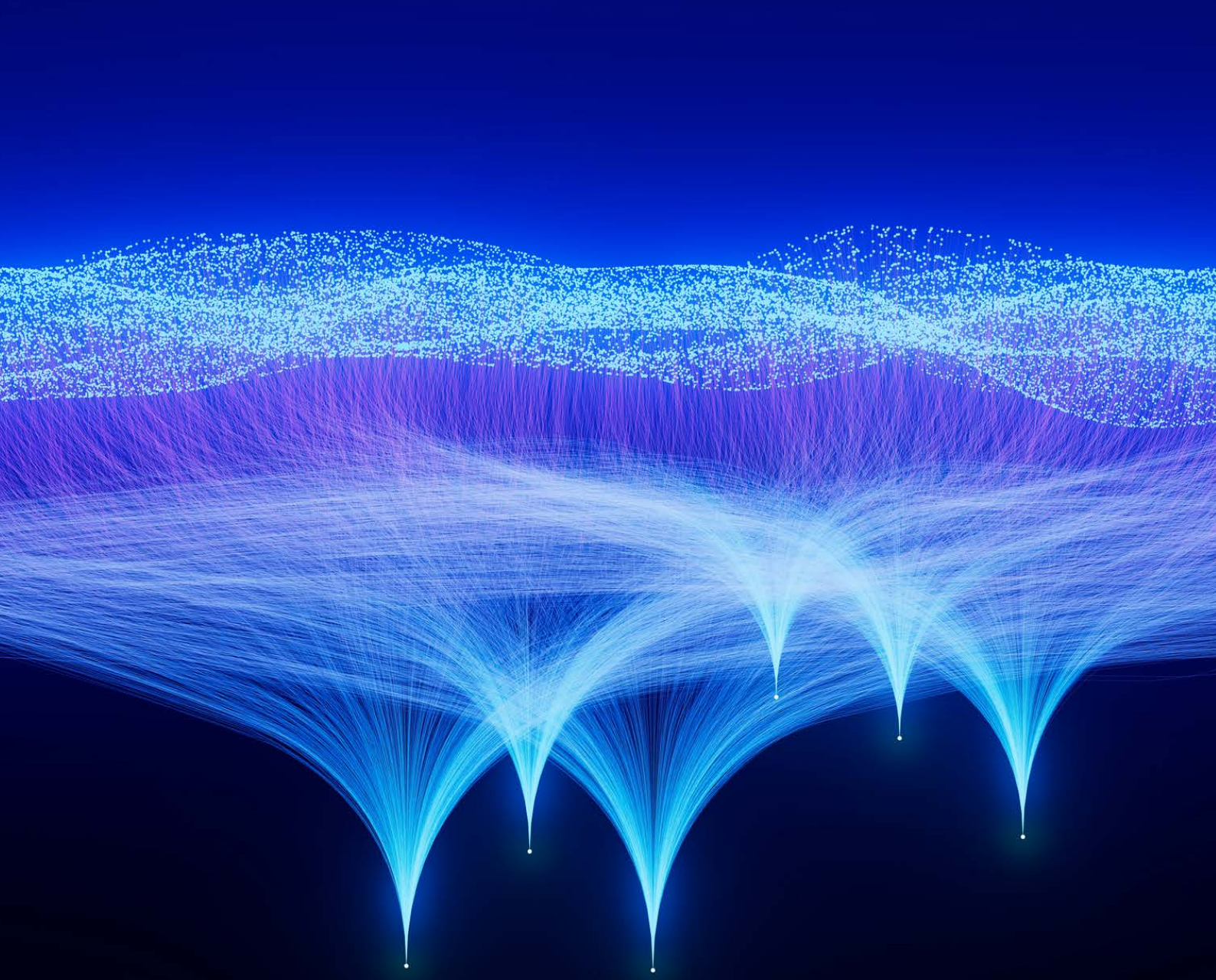
Agentic AI systems — where the model is given tools (code execution, database queries, email composition, API calls) and tasked with completing multi-step goals — are the frontier of enterprise AI deployment. They offer enormous productivity leverage, but their failure modes are qualitatively more serious than those of read-only LLM applications. A hallucination in a knowledge assistant is corrected when the user notices; a hallucination that causes an agentic system to execute an unintended transaction is not.

Zensar's AgentProbe addresses this through structured testing of tool-use authorization boundaries, multi-step reasoning coherence, graceful degradation when tool calls fail or return unexpected results, and instruction-following fidelity across complex, multi-constraint task specifications. The testing methodology explicitly validates that the system respects its intended scope even under adversarial inputs designed to expand its authority — a form of prompt injection that is specific to agentic contexts.

Regulated industry compliance

Organizations deploying AI in regulated industries — financial services, healthcare, insurance, and public sector — face a specific variant of the AI quality challenge: not only must the system work, it must be demonstrably compliant with sector-specific regulations alongside the horizontal obligations of the EU AI Act and NIST AI RMF.

Zensar's Trustworthy Assurance pillar is designed for exactly this context. It produces the evidence artifacts that auditors and regulators expect: a control matrix mapping system controls to regulatory requirements, a risk register with documented harms, mitigation owners, and residual risk sign-off, evaluation reports covering performance, safety, and fairness, a red-team report with adversarial findings and remediation, and a post-market monitoring plan demonstrating ongoing surveillance obligations. [20]



8

Benefits and Value Proposition

The ZenseAI.AssureAI framework delivers value across three dimensions that matter to different enterprise stakeholders: risk reduction for the enterprise risk function, regulatory readiness for legal and compliance teams, and delivery confidence for product and technology teams.

Risk reduction

The most direct benefit of a structured AI quality program is the reduction of AI-specific risk. Hallucinations caught in evaluation rather than in production, fairness gaps identified before a system is deployed to affect customers, security vulnerabilities discovered through red-teaming rather than through a breach — each of these represents a cost avoided that is orders of magnitude larger than the cost of the testing that prevented it.

The framework's emphasis on life cycle monitoring means that risk management does not stop at deployment. Drift detection catches performance degradation before it reaches users. Operational monitoring observability validates that the system's behavior can be traced and explained when something does go wrong. Incident response playbooks ensure that failures are addressed systematically rather than reactively.

Regulatory readiness

For organizations subject to the EU AI Act, ISO/IEC 42001, or sector-specific AI regulations, the ZenseAI.AssureAI framework provides a direct path to audit readiness. The evidence artifacts produced by the framework map to the technical documentation requirements of the EU AI Act's Annex IV, the management system controls required by ISO/IEC 42001, and the testing and evaluation evidence expected under the NIST AI RMF's Measure function.

Building compliance evidence retroactively — after the system is deployed and the regulator comes calling — is significantly more expensive and less reliable than building it as part of the development and deployment process. The framework is designed to generate evidence continuously, ensuring that the documentation is always current and that the evidence gap identified in so many responsible AI audits never has the opportunity to develop.

Delivery confidence

For product and technology teams, the framework delivers something more immediate: confidence that what is being shipped will actually work as expected. The tiered evaluation methodology provides fast feedback in CI/CD pipelines for deterministic checks, with richer evaluation available on demand for pre-release gates and periodic audits. Model regression testing ensures that fine-tuning, retraining, or infrastructure changes do not silently degrade performance that was previously validated.

The six-to-eight-week pilot engagement is designed to make the value tangible within a single project cycle. Starting with a specific, bounded use case — typically a RAG-based assistant with tool-use capabilities — the pilot applies the framework against real system behavior, generates the first set of evaluation evidence, identifies the most significant quality gaps, and delivers a remediation plan that the client team can begin executing immediately.

9

Risks and Considerations

No quality framework is a guarantee. AI quality assurance — like all forms of software quality assurance — operates in a world of incomplete information, evolving threats, and constrained resources. Practitioners engaging with the ZenseAI.AssureAI framework should understand its inherent limitations alongside its capabilities.

The impossibility of complete coverage

Adversarial robustness research confirms that defending against all possible adversarial attacks on an AI system is theoretically impossible. New attack vectors emerge continuously, and any fixed set of red-team prompts will be incomplete almost immediately after being created. The framework addresses this by treating red-teaming as a recurring control — not a one-time pre-deployment exercise — and by maintaining awareness of newly published attack families to keep test suites current.

Fairness impossibility theorems

Mathematical proofs have established that it is impossible to simultaneously satisfy all intuitive fairness criteria — demographic parity, equalized odds, and calibration cannot all hold at the same time except in trivial cases. The ZenseAI.AssureAI framework does not promise to eliminate unfairness; it promises to measure it, document it, and make explicit trade-offs transparent so that the humans responsible for deploying the system can make informed decisions about what they are willing to accept.

Evolving regulatory landscape

AI regulation is moving faster than any other domain of technology law. The EU AI Act is the most comprehensive existing framework, but its high-risk classification thresholds are subject to interpretation, and implementing acts and standards are still being developed. The NIST AI RMF is updated periodically, with the 2025 update adding new emphasis on model provenance and third-party assessment. Organizations must assume that the compliance landscape will continue to evolve and that their testing frameworks must be updated accordingly.

LLM-as-judge limitations

Using LLMs to evaluate LLM outputs — the LLM-as-judge paradigm employed in Tier 2 of the evaluation methodology — introduces evaluator bias. The judge model may share the same failure modes as the model being evaluated, may have positional bias (preferring longer or earlier responses), and may not be well-calibrated for domain-specific quality criteria. The framework mitigates this through careful judge model selection, multiple independent evaluations, and regular calibration against human judgments — but practitioners should not treat automated LLM-as-judge scores as ground truth.

9 Conclusion

The gap between AI's capability and the enterprise's ability to assure that capability is one of the defining technology risks of this decade. AI systems are being deployed at scale across customer relationships, regulatory environments, and operational workflows that demand reliability, fairness, safety, and accountability — properties that traditional software quality practices were never designed to verify.

Zensar's ZenseAI.AssureAI practice exists to close that gap. The four-pillar framework — Data Quality Assurance, Functional and Model Evaluation, Trustworthy Assurance, and Non-Functional Assurance — addresses the full spectrum of AI-specific quality risk, from the integrity of training data through the safety behavior of deployed agentic systems. The 30-test catalog operationalizes the framework into executable, evidence-generating procedures aligned to the governance frameworks that regulators and auditors are beginning to require.

The evidence is clear. Organizations that invest in structured AI quality assurance earlier in their AI adoption journey will spend less on remediation, face less regulatory exposure, and build more durable relationships with the customers and stakeholders who depend on their AI systems to behave as promised.

Zensar recommends a 6-to-8-week pilot engagement as the starting point. The pilot applies the ZenseAI.AssureAIframework to a bounded, production-representative AI system — typically a RAG-based assistant or agentic workflow — to generate the first set of evaluation evidence, identify quality gaps, and establish the measurement baselines that future assurance work will build upon.

To initiate a conversation about the ZenseAI.AssureAI pilot program, contact Zensar's AI Quality Engineering practice through your Zensar account team or through the Applied Research office at Zensar Technologies, Pune.

References

[1] A. de Hond et al., "Guidelines and quality criteria for artificial intelligence-based prediction models in healthcare: a scoping review," npj digital medicine, vol. 5, no. 1, Jan. 2022. doi: 10.1038/s41746-021-00549-7

[2] U. Mahmood et al., "Artificial intelligence in medicine: mitigating risks and maximizing benefits via quality assurance, quality control, and acceptance testing," vol. 1, Jan. 2024. doi: 10.1093/bjrai/ubae003

[3] C. Tao, J. Gao, and T. Wang, "Testing and Quality Validation for AI Software—Perspectives, Issues, and Practices," IEEE Access, vol. 7, pp. 120164–120175, Aug. 2019. doi: 10.1109/ACCESS.2019.2937107

[4] G. Marco, "Framework for the Integration and Governance of Complex Artificial Intelligence Systems - AIQ 1001:2025," Oct. 2025. doi: 10.5281/zenodo.17360353

[5] L. Ullrich, M. Buchholz, K. Dietmayer, and K. Graichen, "AI Safety Assurance for Automated Vehicles: A Survey on Research, Standardization, Regulation," IEEE Transactions on Intelligent Vehicles, Jan. 2024. doi: 10.1109/tiv.2024.3496797

[6] E. Shahul, J. James, L. E. Anke, and S. Schockaert, "RAGAS: Automated Evaluation of Retrieval Augmented Generation," arXiv, Sept. 2023. doi: 10.48550/arxiv.2309.15217

[7] NIST, Artificial Intelligence Risk Management Framework (AI RMF 1.0), January 2023. https://nvlpubs.nist.gov/nistpubs/ai/nist.ai.100-1.pdf

[8] NIST, Artificial Intelligence Risk Management Framework: Generative AI Profile (AI 600-1). https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.600-1.pdf

[9] European Commission, AI Act | Shaping Europe's digital future. https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai

[10] X. Sun, D. Ståhl, K. Sandahl, and C. Kessler, "Quality Assurance of LLM-generated Code: Addressing Non-Functional Quality Characteristics," Nov. 2025. doi: 10.1016/j.jss.2026.112885

[11] P. Dedu, W. Lopes, and N. Sirikonda, "An Observability Framework for Detecting Bias and Drift in AI-Based Student Early Alert Systems." https://ieeexplore.ieee.org/abstract/document/11355034/

[12] M. Kwiatkowska, "AI Driven DevOps and Machine Learning Systems for Privacy Preserving Healthcare and Digital Advertising." https://ijrai.org/index.php/ijrai/article/view/429

[13] K. Hamada et al., "Guidelines for Quality Assurance of Machine Learning-based Artificial Intelligence," pp. 335–341, Jan. 2020.

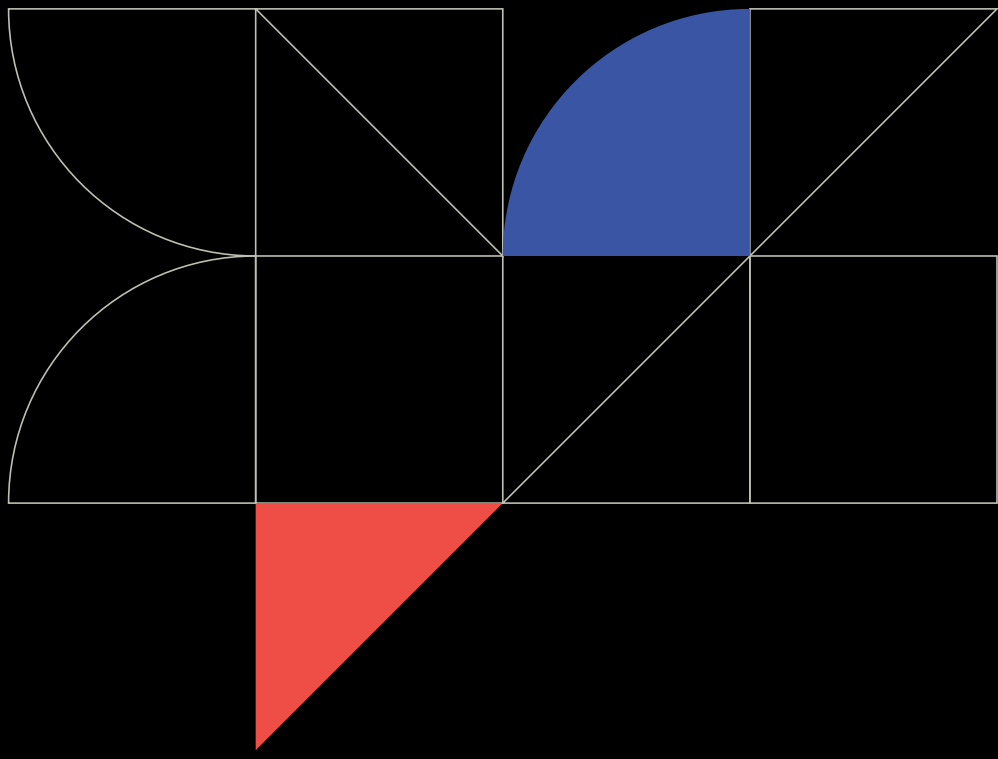
[14] ISO/IEC 42001:2023 Artificial intelligence — Management system. https://learn.microsoft.com/en-us/compliance/regulatory/offering-iso-42001

[15] Md. K. Islam et al., "Does Differential Privacy Impact Bias in Pretrained NLP Models?" Oct. 2024. doi: 10.48550/arxiv.2410.18749

[16] S. Roychowdhury et al., "Evaluation of RAG Metrics for Question Answering in the Telecom Domain," arXiv, abs/2407.12873, July 2024. doi: 10.48550/arxiv.2407.12873

[17] M. Sendak et al., "Editorial: Surfacing best practices for AI software development and integration in healthcare," Frontiers in Digital Health, vol. 5, Feb. 2023. doi: 10.3389/fdgth.2023.1150875

[18] Cloud Security Alliance, NIST AI Risk Management Framework: Agentic Profile. https://labs.cloudsecurityalliance.org/agentic/agentic-nist-ai-rmf-profile-v1/



zensar

An RPG Company

At Zensar, we're 'experience-led everything.' We are committed to conceptualizing, designing, engineering, marketing, and managing digital solutions and experiences for over 145 leading enterprises. Using our 3Es of experience, engineering, and engagement, we harness the power of technology, creativity, and insight to deliver impact.

Part of the $4.8 billion RPG Group, we are headquartered in Pune, India. Our 10,000+ employees work across 30+ locations worldwide, including Milpitas, Seattle, Princeton, Cape Town, London, Zurich, Singapore, and Mexico City.

For more information, please contact: info@zensar.com | www.zensar.com