Azure AI Foundry Evaluations Can Be an Enterprise Readiness Audit

A monetization play for Azure-heavy orgs: package Foundry evaluations (AI quality + NLP metrics + risk/safety metrics) into an enterprise readiness audit, then sell remediation sprints and ongoing scorecard retainers.

Source https://yetyield.com/blog/azure-ai-foundry-evaluations-as-enterprise-readiness-audit/

If a buyer is already committed to Azure, they don’t want “another evaluation stack.”

They want evidence their GenAI app is safe enough to deploy, and a repeatable way to keep it that way.

Azure AI Foundry evaluations are a useful substrate for packaging that evidence into a paid audit.

The monetization angle

Treat evaluation as an enterprise readiness audit:
• can we measure quality on real data?
• can we measure risk and safety harms?
• can we produce an evidence pack procurement and risk teams can accept?

Offer ladder:
1.Readiness audit (dataset + evaluation design + initial run)
2.Remediation sprint (fix top failure modes)
3.Ongoing evaluation scorecard retainer (monthly yield)

This complements:
• Bedrock Guardrails Grounding Checks Can Be a Compliance Monetization Layer (/blog/bedrock-guardrails-grounding-checks-as-compliance-monetization/)
• Phoenix Server-Side Evals Can Be a No-Code Evaluation Ops Offer (/blog/phoenix-server-side-evals-as-no-code-evaluation-ops-offer/)

The official Foundry surfaces you can cite

Azure AI Foundry describes running evaluation runs on substantial datasets, producing quantitative metrics and AI-assisted metrics, and managing evaluators in the portal.
Official entry point: https://learn.microsoft.com/en-us/azure/ai-foundry/how-to/evaluate-generative-ai-app

It also describes:

• supported dataset formats (CSV, JSONL)
• evaluation metric categories:
• AI quality (AI assisted)
• AI quality (NLP)
• risk and safety metrics
• data mapping for metrics (query, response, context, ground truth depending on metric)

These are all stable, audit-friendly anchors.

Why “risk & safety metrics” is a pricing wedge

Reliability gets funded.

Risk also gets funded, but from a different budget owner.

When your audit includes safety metrics (and a clear remediation plan), it becomes easier to sell:
• stakeholder alignment sessions
• policy decisions
• monitoring and incident playbooks

That turns evaluation into a compliance-priced trust layer.

The readiness audit deliverables

Deliverable 1: Dataset + mapping spec
• define required columns (query/response/context/ground truth)
• document mappings per metric category
• validate data quality and coverage

Deliverable 2: Evaluation plan (v1)

Pick a small set of metrics that map to business risk:
• groundedness (for RAG-like workflows)
• coherence / relevance (quality)
• risk and safety metrics that match your domain

Azure AI Foundry lists groundedness among its AI-assisted quality metrics and also separates risk and safety metrics as a category.
Official anchor: https://learn.microsoft.com/en-us/azure/ai-foundry/how-to/evaluate-generative-ai-app

Deliverable 3: Evidence pack

This is what makes the audit billable:
• the evaluation run configuration
• the metric outputs
• an executive summary translating metrics into decisions
• a prioritized remediation backlog

Deliverable 4: Remediation sprint

Fix the top 3 failure modes:
• prompt boundary repairs
• retrieval and grounding improvements
• policy adjustments and refusal tuning
• schema/format compliance

Deliverable 5: Monthly scorecard retainer

Monthly:
• rerun evaluation on updated datasets
• trend lines (quality + safety)
• new cases added from incidents
• updated evidence pack excerpt for governance

What to avoid
• “compliance theater” without tests
• shipping safety policies with no evidence pack
• scoring without thresholds and escalation rules

Next research direction: a unified “QA + compliance retainer” bundle that combines:
• quality gates (task success, tool-use correctness)
• safety and grounding checks
• monthly scorecards + incident response