AI Audit Readiness Framework

Author: Dr. Shaheen Gauher

AI powered products and solutions are increasingly being deployed and embedded in business workflows to influence day to day business operations and decision making. It is critical that these AI systems and solutions operate optimally and as expected. Periodic audit of these systems is conducted to ensure that they are functioning as intended, are fair, transparent, and trustworthy, and are in compliance with relevant laws, regulations, and ethics. It aims to identify and mitigate any risks including biases, inaccuracies, misuse of data, security vulnerabilities or business discontinuity. This serves in protecting both the company and the consumers from potential harm that may be caused by faulty or unethical AI operation. The harm can be financial, reputational, operational, or legal/regulatory among others.

An AI Audit Readiness Framework was designed to structure the process of auditing the AI systems. The Framework presented below identifies thirteen facets across four categories, against which an AI system/product/program/solution can be assessed.

#	Facet Category	Facet
1	Scope, governance & oversight processes	Solution Charter score
2	Scope, governance & oversight processes	RACI score
3	Scope, governance & oversight processes	Governance/Sign Offs score
4	Explainability and transparency of the AI system	Data Quality score
5	Explainability and transparency of the AI system	Data Explainability score
6	Explainability and transparency of the AI system	Model Transparency score
7	Explainability and transparency of the AI system	Model Fairness score
8	Explainability and transparency of the AI system	Model Explainability score
9	Explainability and transparency of the AI system	Solution Overview score
10	Monitoring & Measurement	Measurement Strategy score
11	Monitoring & Measurement	Monitoring Processes score
12	Quality assurance processes	Testing and QA strategy score

The twelve facets comprising the Audit Readiness Framework are described in detail below. An assessment of a facet can result in a score (5 scale score) for the facet as shown below.

No evidence found	Does not meet expectation	Minimally meets expectation	Meets expectation	Exceeds expectation

Solution Charter score

To calculate the Solution Charter score we will look for the existence of a document that contains the objective and scope of the solution clearly defined with measurable KPIs for success and failure. It outlines the sponsor/stakeholder's expectations. It should contain a vision and a plan to address the business problem with a detailed roadmap, milestones, and timelines. It should have sufficient details and mimic a statement of work to the extent possible.
A high score indicates the Solution has a clear project charter that outlines the objective and scope of the program and clearly defined and measurable KPIs for success and failure.

RACI score

To calculate the RACI score we will look for the existence of clearly documented roles for responsibility, and accountability, and appropriate stakeholder engagement. Additionally, we will look for clear communication evidence about the roles and responsibilities with the intended parties.
A high score indicates the Solution has a well-defined RACI. It is clear who is responsible and accountable for what components of the product/system.

Governance score

To calculate the Governance score, we will look for clearly documented governance requirements for the system/product/solution. We will look for data Governance (use, access, privacy) in place. We will look for evidence for sign offs from the appropriate stakeholders, evidence of approval on key decisions including minimum viable product, evidence of Legal, Compliance, and AUMSI review and approval of the requirements. Additionally, we will also look for communication of bias testing results to key business stakeholders before deployment to production.
A high score indicates the solution/system has adequate governance in place with clearly documented and archived sign offs and approvals from appropriate stakeholders throughout the product lifecycle.

Data Quality score

To calculate the Data Quality score, we will look for the existence of a data dictionary with properly defined and documented attributes of variables used for each of the training data sets corresponding to the model(s) comprising the solution. Data provenance and data lineage should be clearly outlined. Profiles for each of the training data sets corresponding to the model(s) comprising the solution should be available. We will look for evidence for no data contamination and no data leakage in modeling data. In production, we will look for evidence of quality control on the scoring or inference data.
A high score indicates the model(s) comprising the solution are developed using high quality data. Conversely, a low score indicates the quality of the data is unknown.

Data Explainability score

To calculate the Data Explainability score we will look for adequate and reasonable explanation for inclusion of the data elements in the training data for building a predictive model.

Categorizing features into buckets such as demographic, sdoh, utilization, clinical history, payment history, history of past digital behavior etc. are some examples of categories that the features can be bucketed in, for purposes of explanation. We will also look for a sign off from a subject matter expert to vouch for the reasonableness of inclusion of these features as predictors in predictive models.
A high score indicates the model(s) comprising the solution are developed using data that can be reasonably explained as fit for the intended purpose. Conversely, a low score indicates adequate explanation for using the specific data elements is not available.

Model Transparency score

To calculate the Model Transparency score we will look for the existence of suitable model documentation. Specifically, we will look for Model Cards to be filled out for each model (NLP card when applicable). When using off the shelf models, results from evaluation of the model to be fit for intended purpose, safe and reliable, should be documented.
A high score indicates the model(s) comprising the solution are documented and have Model Cards filled out. Conversely, a low score indicates the models lack documentation.

Model Fairness score

To calculate the Model Fairness score we will look for the existence of fairness analysis completed according to criteria outlined in Enterprise Bias Standard. SME communication and acceptance of how to segment population and identify at risk groups for the analysis should be available. The design for conducting bias analysis should typically be informed by the Bias Data and Equity Analysis template completion.
A high score indicates the model(s) comprising the solution have both training data tested for bias and model performance tested for bias for sub populations identified as at-risk within the population that will be served by the model(s).

Model Explainability score

To calculate the Model Explainability score we will look for clearly documented strategy for evaluating the outputs from the model for correctness. We will look for the existence of explanation of predictions (outputs) from the AI/model(s). The predictions should be reasonable from the use case standpoint and should be common sense explainable to the extent possible. SME communication and acceptance of outputs from AI/model runs should be available.
A high score indicates the AI/model(s) comprising the solution make predictions that can be explained and are reasonable. Conversely, a low score indicates the predictions from the model(s) cannot be explained reasonably.

Solution Overview score

To calculate the Solution overview score we will look for clear schematics to describe the solution, system architecture and design of the product/solution/system. Documents should tell a connected story across components, touch points and systems. The end-to-end data flow through the AI system should be clearly documented with diagrams clearly describing the

various data transformations and checkpoints. The diagram(s) should also outline: - system(s) references, master data store, system Integration references, supporting system platforms, show existing and new system/platform investment, high level capabilities and/or key features, internal/external users/system references, repositories (content, database, etc.), end user communities.
A high score indicates the solution has clear schematics to describe the solution, system, architecture, and design of the product/solution and can tell a connected story. Conversely, a low score indicates the solution does not have documentation to tell a connected story across components, touch points and systems.

Measurement Strategy score

To calculate the Measurement Strategy score we will look for the existence of a well-documented approach to tie the outputs from the model(s) comprising the AI solution/product to achieve the intended business objective (ROI). The data used for measurement should be clearly described and available. We will look for financial metrics that can be tied to the solution benefit trend. We will also look for the existence of a systematic and reproducible process for performing the measurement for the effectiveness of model(s) comprising the AI system/solution and of the AI system as a whole, in meeting stakeholders' expectations.
A high score indicates that appropriate data, metrics, and process exist to tie the predictions/insights from the solution to the business outcomes. Conversely, a low score indicates a systematic and reproducible process to measure effectiveness and ROI does not exist.

Monitoring Processes score

To calculate the Monitoring Processes score we will look for the existence of processes to monitor for a) model/AI degradation, b) model/AI fairness, c) data drift, d) data quality etc. in *production*. The baseline metrics and acceptable operating thresholds for all monitoring exercises should be clearly documented. We will also look for processes in place for System Health monitoring – e.g., meeting SLAs, system resource utilization metrics monitoring viz. CPU, memory, storage, transaction volume/traffic, logging, error notification/alerting etc. In the event of violation of acceptable thresholds during monitoring, we will look for processes in place for mitigation and for business continuity.
A high score indicates the AI system and the model(s) comprising the solution have monitoring processes in place with clearly defined operating thresholds in place.

Testing and Quality Assurance (QA) strategy score

To calculate the quality assurance (QA) process score we will look for the existence of documented testing and QA strategy for the AI/model outputs and overall solution/system. The testing scenarios, testing scripts and testing results should be available. Furthermore, the testing results should be reproducible.
A high score indicates that the AI system and the model(s) comprising the solution have a programmatic and systematic quality assurance and testing process in place.