Overview

There are different AI documentation Templates each unique and are used based on the requirements and models

Below are the Templates:

1. MODEL CARD TEMPLATE
2. DATA CARD TEMPLATE
3. AI DECISION LOG TEMPLATE
4. AI RISK ASSESSMENT FORM
5. AI USE CASE INTAKE FORM
6. AI MODEL APPROVAL FORM
7. AI AUDIT LOG TEMPLATE
8. MITRE ATLAS THREAT ASSESSMENT TEMPLATE

1. MODEL CARD TEMPLATE

1.1 Overview

- Model Name:
- Model Version:
- Model Owner:
- Business Unit / Product:
- Creation Date:
- Last Updated:

1.2 Purpose & Intended Use

- **Intended Purpose:** Describe what the model is designed to do.
- **Intended Users:** (Example: Business teams, analysts, end-users, systems.)
- **Supported Decisions / Actions:** (Mention What decisions does this model influence)
- **Not Intended For:** (Explicitly list forbidden / unsafe uses.)

1.3 Model Architecture

- **Model Type:** (LLM, CNN, XGBoost, Transformer, etc.)
- **Problem Type:** (Classification, Regression, NLP, CV, GenAI)
- **Input Features (summary):**
- **Output Format:** (probability, label, summary, etc.)

1.4 Training Data Summary

- **Data Sources:**
- **Collection Method:**
- **Time Period Covered:**
- **Populations / Demographics Included:**
- **Sensitive Attributes Present (and handling):**
- **Known Data Limitations:**

1.5 Evaluation Metrics

- **Performance Metrics:** (Accuracy, AUC, Precision, F1, BLEU, etc.)
- **Fairness Metrics**: (Group error rates, disparate impact, parity gap.)
- **Robustness Metrics:** (Performance under noise, perturbation, adversarial tests.)
- **Benchmark Datasets Used:** (Internal or public benchmarks.)

1.6 Ethical & Fairness Considerations

- **Potential ethical risks**
- **Impact on specific user groups**
- **Bias sources (data, label, historical, algorithmic)**
- **Fairness mitigation steps applied**
- **Ethical concerns that remain**

1.7 Safety & Security

- **Safety risks identified**
- **Security risks (poisoning, prompt injection, model theft)**
- **Tests performed (red-teaming, adversarial, stress tests)**
- **Guardrails implemented**
- **Abuse-prevention mechanisms**

1.8 MITRE ATLAS Threat Assessment

- **Threat IDs applicable: (AT1021, AT1005, AT1018, .)**
- **Threat descriptions**

- **Controls implemented**
- **Residual risk per threat (Low / Medium / High)**

1.9 Explainability & Transparency

- **Explainability methods used:(SHAP, LIME, Feature Importance)**
- **Who can access explanations: (end-users / auditors / regulators)**
- **Degree of interpretability: (glass-box / black-box)**
- **User-facing transparency disclosures**
- **Explainability limitations**

1.10 Deployment & Operational Controls

- **Deployment environment (cloud/on-prem/hybrid/API/batch)**
- **Access controls & permissions**
- **Pre-deployment governance checks**
- **Rollout strategy (pilot / shadow / phased / full)**
- **Fallback & failover mechanisms**
- **Manual override availability**

1.11 Known Limitations

- **Performance Limitations:** (Where model accuracy is weak.)
- **Data Limitations**:(Gaps in geography, demographics, time periods.)
- **Scope Limitations:**(Tasks the model must NOT be used for.)
- **Ethical Limitations:**(Groups that remain at risk.)
- **Communication of Limitations**:(How these are communicated to users.)

1.12 Maintenance & Monitoring Plan

1.12.1 Monitoring Areas:

- Accuracy
- Drift (data + concept)
- Fairness metrics by group
- Latency
- Security anomalies
- Hallucination or harmful output rate

1.12.2 Monitoring Frequency:

- Real-Time
- Daily
- Weekly
- Monthly
- Yearly

1.12.3 Retraining Plan:

- Retraining triggers
- Retraining frequency
- Owners responsible

1.12.4 Incident Response:

- Logging method
- Escalation path
- SLA for mitigation

1.12.5 Governance Review:

- Model Owner
- Review cycle (quarterly / annual)
- Next scheduled review date

2. DATA CARD TEMPLATE

2.1 Dataset Overview

- **Dataset Name**
- **Dataset Owner**
- **Purpose of Dataset**
- **Data Type (text, images, structured, logs, etc.)**

2.2 Data Collection

- **Source systems**
- **Collection methodology**

- **Frequency**
- **Consent requirements**
- **Third-party/vendor involvement**

2.3 Data Structure

- **Number of rows**
- **Key features / columns**
- **Target labels**
- **Labelling method**

2.4 Data Quality Assessment

- **Completeness**
- **Accuracy**
- **Consistency**
- **Timeliness**
- **Duplicate records**
- **Missing values**

2.5 Sensitive Attributes

List sensitive attributes and handling:

- **Removed**
- **Masked**
- **Aggregated**
- **Only used for fairness testing**

2.6 Bias Assessment

- **Representational bias**
- **Sampling bias**
- **Label bias**
- **Coverage imbalance**

2.7 Preprocessing Steps

- Data cleaning
- Feature engineering
- Normalization
- Encoding
- Filtering
- Anonymisation

2.8 Privacy & Compliance

- GDPR / PDPA / CCPA applicability
- Data minimization
- Retention periods
- Data access restrictions
- Encryption & storage policies

2.9 Risks & Limitations

- Drift risk
- Data freshness issues
- Demographic underrepresentation
- Structural bias

2.10 Controls Implemented

- Technical (encryption, hashing, tokenization)
- Process (restricted roles, access approvals)
- Validation checks
- Data lineage tracking

3. AI DECISION LOG TEMPLATE

Fields:

- **Decision ID**
- **Date**
- **Decision Maker(s)**
- **Decision Category (Risk / Security / Fairness / Model Change / Deployment / …)**
- **Context / Problem Description**
- **Options Considered**
- **Final Decision**
- **Rationale for Decision**
- **Impact (High / Medium / Low)**
- **Evidence / Attachments**
- **Status (Open / Closed / In Review)**

4. AI RISK ASSESSMENT FORM

4.1 Use Case Details

- **Use Case Name**
- **Business Owner**
- **Model Owner**
- **Description**

4.2 Risk Scoring (RAG)

- **Impact (1–5)**
- **Likelihood (1–5)**
- **Detectability (1–5)**
- **Risk Score = Impact × Likelihood × Detectability**
- **RAG Rating (Red / Amber / Green)**

4.3 Risk Areas

- **Fairness**
- **Privacy**
- **Explainability**
- **Security**
- **Ethical**
- **Reputational**
- **Operational**
- **Compliance**

4.4 MITRE ATLAS Threat Review

- **Threat IDs mapped**
- **Threat scenarios**
- **Controls in place**
- **Residual threat level**

4.5 Controls Required

- **Data controls**
- **Model controls**
- **Deployment controls**
- **Monitoring controls**

4.6 Final Recommendation

- **Approve**
- **Approve with Mitigation**
- **Reject**

5. AI USE CASE INTAKE FORM

5.1 Basic Info

- **Use Case Name**
- **Requestor**
- **Business Unit**
- **Problem Statement**

- **Expected Business Value**

5.2 AI Fit Assessment

- **Why AI is needed**
- **Alternatives considered**
- **Proposed model type**

5.3 Data Requirement

- **Data needed**
- **Data sources**
- **Sensitive attributes involved**

5.4 Quick Risk Indicators

- **Impacts individuals?**
- **Sensitive data?**
- **Safety-critical?**
- **Legal/regulatory exposure?**

5.5 Approval Route

Low / Medium / High → Corresponding approval workflow

6. AI MODEL APPROVAL FORM

6.1 Model Information

- **Model Name**
- **Version**
- **Owner**
- **Use Case**

6.2 Documentation Checklist

Must provide:

- Model Card
- Data Card
- Risk Assessment

- Explainability Report
- ATLAS Threat Assessment
- Monitoring Plan

6.3 Control Validation

- **Fairness controls**
- **Privacy controls**
- **Security controls**
- **Deployment readiness**
- **Monitoring readiness**

6.4 Governance Sign-Off

- **AI Program Manager**
- **AI Risk Manager**
- **Data Protection Officer**
- **Security Lead**
- **AI Governance Board**

7. AI AUDIT LOG TEMPLATE

Fields:

- **Audit ID**
- **Date**
- **Model Name**
- **Issue Type (Bias / Drift / Security / Privacy / Hallucination…)**
- **Severity**
- **Description**
- **Root Cause**
- **Evidence**
- **Action Taken**
- **Owner**
- **Status**

8. MITRE ATLAS THREAT ASSESSMENT TEMPLATE

8.1 Threat Surface

- **Input attacks**
- **Data poisoning**
- **Model inversion**
- **Membership inference**
- **Model extraction**
- **Jailbreak / prompt injection**
- **Unauthorized API access**

8.2 Threat Mapping Table

- **Threat ID**
- **Threat Name**
- **Threat Type**
- **Description**
- **Affected Stages**
- **Likelihood**
- **Impact**
- **Risk Score**
- **Controls Applied**
- **Residual Risk**
- **Status**

8.3 Controls Summary

- **Input filtering**
- **Adversarial training**
- **Differential privacy**
- **Monitoring & logs**
- **API rate limiting**
- **LLM guardrails**

8.4 Residual Risk

LOW / MEDIUM / HIGH (with Justification)