AQI Automotive Quality Institute

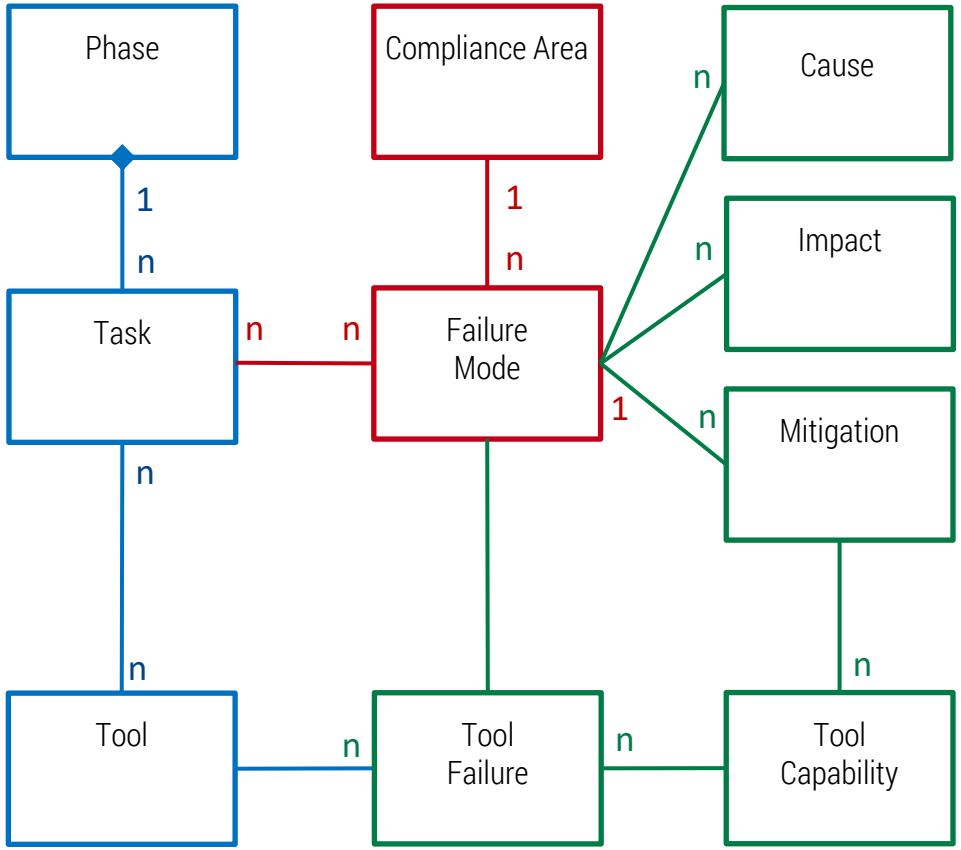
Q in AI: Requirements for tools and processes for assurance

Instructions for qualifying tools AssessML

Dr. Björn Schünemann (bjoern.schuenemann@aqigmbh.de)
Dr. Jürgen Großmann (juergen.grossmann@fokus.fraunhofer.de)

Relationship between concepts

Information model: Project procedure and terminology used



- Systematic listing of **life cycle phases (phase)** and **AI tasks (task)** as well as exemplary **tools (tool).**
- Identification of the **failure mode (failure mode)** along the **AI tasks** for each **compliance area (compliance area).**
- Identification of **causes, effects (impact)** and possible countermeasures (mitigation) for each **failure mode.**
- Identification of specific tool failures **(tool failure)** and **required tool capabilities (tool capability)** for a **failure mode.**

Phase: Phase from the AI life cycle

Task: Task during the AI life cycle that is supported by tools .

Tool: Tool that supports/executes a task.

Compliance Area: Area of compliance (functional safety, data protection, AI regulation).

Failure Mode: Hazard in the event of non-compliance.

Tool Failure: Tool error as the cause of an error condition.

Cause: Causes of the risk.

Impact: Effects of an error condition.

Mitigation: Measures to reduce the risk.

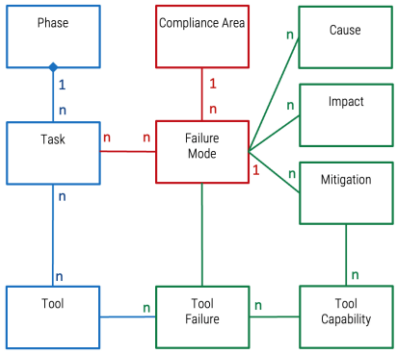
Tool Capability: Capabilities required of a tool to address the risk.

Systematic presentation of the developed content

Result is a Python-based web application

AQI | Automotive Quality Institute

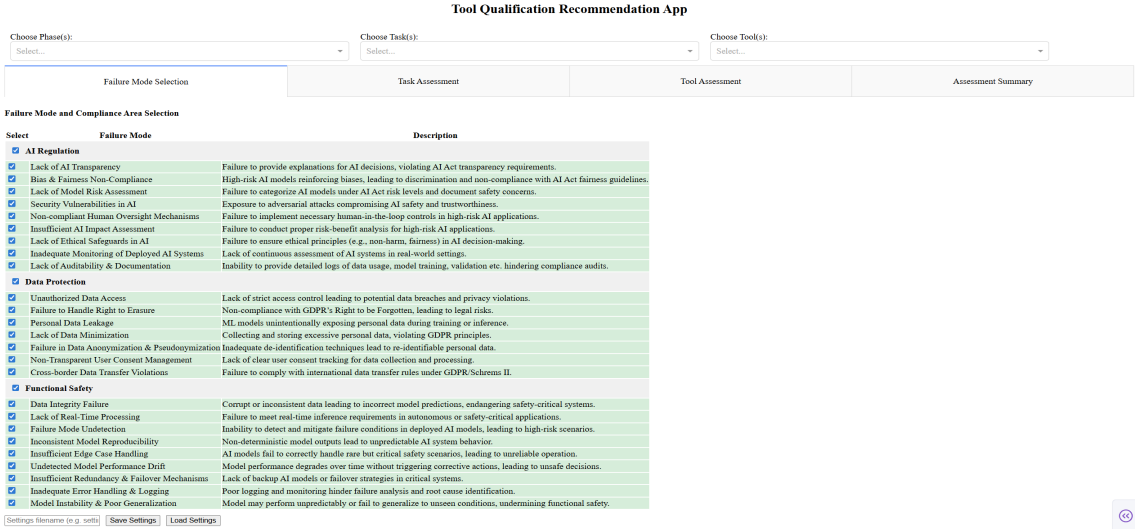
Terminology



The contents of the Excel spreadsheet are compiled in accordance with the terminology used

Compliance Area	Phase	Task	Failure Mode	Impact	Severity	Causes	Mitigation	Occurrence	Detectability	RPN	Potential Tool Failure	Recommended Tool Capability
Functional Safety	Training & Validation	Deep Learning & ML Frameworks	Lack of Real-Time Processing	Trained models and supporting frameworks are unable to deliver low-latency, real-time inference performance required for critical applications such as autonomous systems.		Model architectures are too large or complex for target real-time environments. Frameworks do not leverage hardware acceleration efficiently (e.g., GPU, TPU, Edge devices). Pipeline overhead (e.g., preprocessing, postprocessing) adds excessive latency during inference.	Apply model optimization techniques such as pruning, quantization, or knowledge distillation. Ensure deployment-ready models are benchmarked and optimized for specific hardware accelerators. Optimize full pipeline latency by batching operations and reducing unnecessary transformations.				May fail to guarantee deterministic training execution and runtime validation, resulting in trained models that do not meet real-time safety-critical performance requirements.	Framework support for model optimization (e.g., TensorRT, ONNX Runtime optimizations). Support for hardware-specific runtime optimizations and deployment targeting (e.g., EdgeTPU, CUDA kernels). Framework support for pipeline fusion, minimal preprocessing APIs, and asynchronous inference.
Functional Safety	Training & Validation	Deep Learning & ML Frameworks	Inadequate Error Handling & Logging	Frameworks and models lack robust error detection, recovery, and structured logging capabilities, making root cause analysis and system recovery difficult.		Training and inference exceptions are not properly caught or logged. Losses, gradients, or other training dynamics are not systematically monitored for anomalies. Training and deployment pipelines lack comprehensive logging and metadata tracking.	Integrate structured exception handling and mandatory logging for all critical training and inference operations. Implement runtime monitors for numerical instabilities (e.g., NaNs, divergence) and alert on thresholds. Embed pipeline-wide metadata capture and systematic experiment logging into the training framework.				May fail to provide comprehensive error logging and runtime monitoring, resulting in trained models with limited traceability for failure analysis in safety-critical systems.	Framework support for structured error reporting, custom callbacks, and logging integration. Built-in hooks for anomaly detection during training and validation (gradient checkers, divergence detectors). ML lifecycle management tools (e.g., MLflow, Weights & Biases) integration for metadata and error tracking.
Functional Safety	Training & Validation	Deep Learning & ML Frameworks	Model Instability & Poor Generalization	Inconsistent training leading to unreliable AI behaviour in critical applications.		Lack of deterministic/reproducible training settings due to parallel computation and floating-point arithmetic. Limited built-in validation or monitoring. No built-in numerical stability mechanisms.	Implement deterministic training pipelines with seed control and precision configuration to ensure repeatability. Integrate continuous performance monitoring during training with automatic validation checkpoints. Use frameworks that incorporate gradient clipping, normalization, and regularization for improved numerical robustness.				May fail to enforce deterministic training procedures and runtime validation, resulting in trained models with unpredictable behavior and poor generalization in safety-critical applications.	Explicit deterministic training support with reproducibility settings and controlled parallelism. Built-in validation dashboards and metric tracking APIs. Numerical stability modules with configurable regularization options.

Dashboard created based on Excel spreadsheet



Risk-based tool qualification (details)

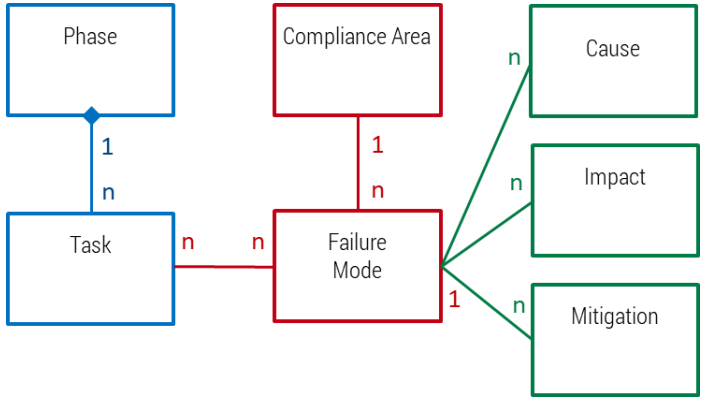
Two-stage process for identifying task-related risks and qualifying tools based on the risks

AQI Automotive Quality Institute

1. Identification of risks in development tasks (according to FMEA)

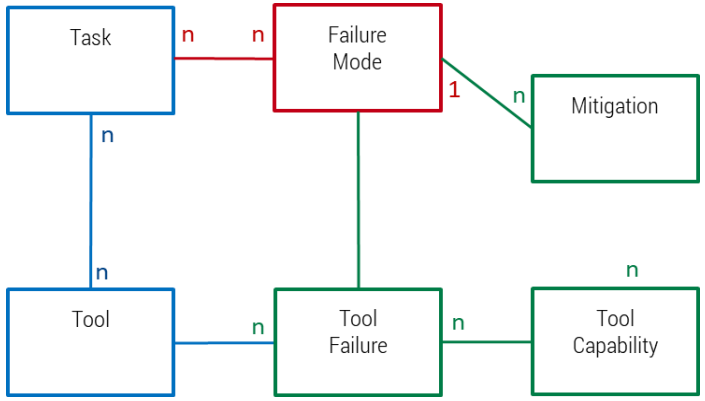
RPN = Severity * Occurrence * Detection

1.1 Selection of relevant tasks phases	1.2 Selection of area of compliance and applicable failure modes	1.3 Calculation of the RPN for selected failure modes in a task



2. Identification of qualification requirements for AI tools

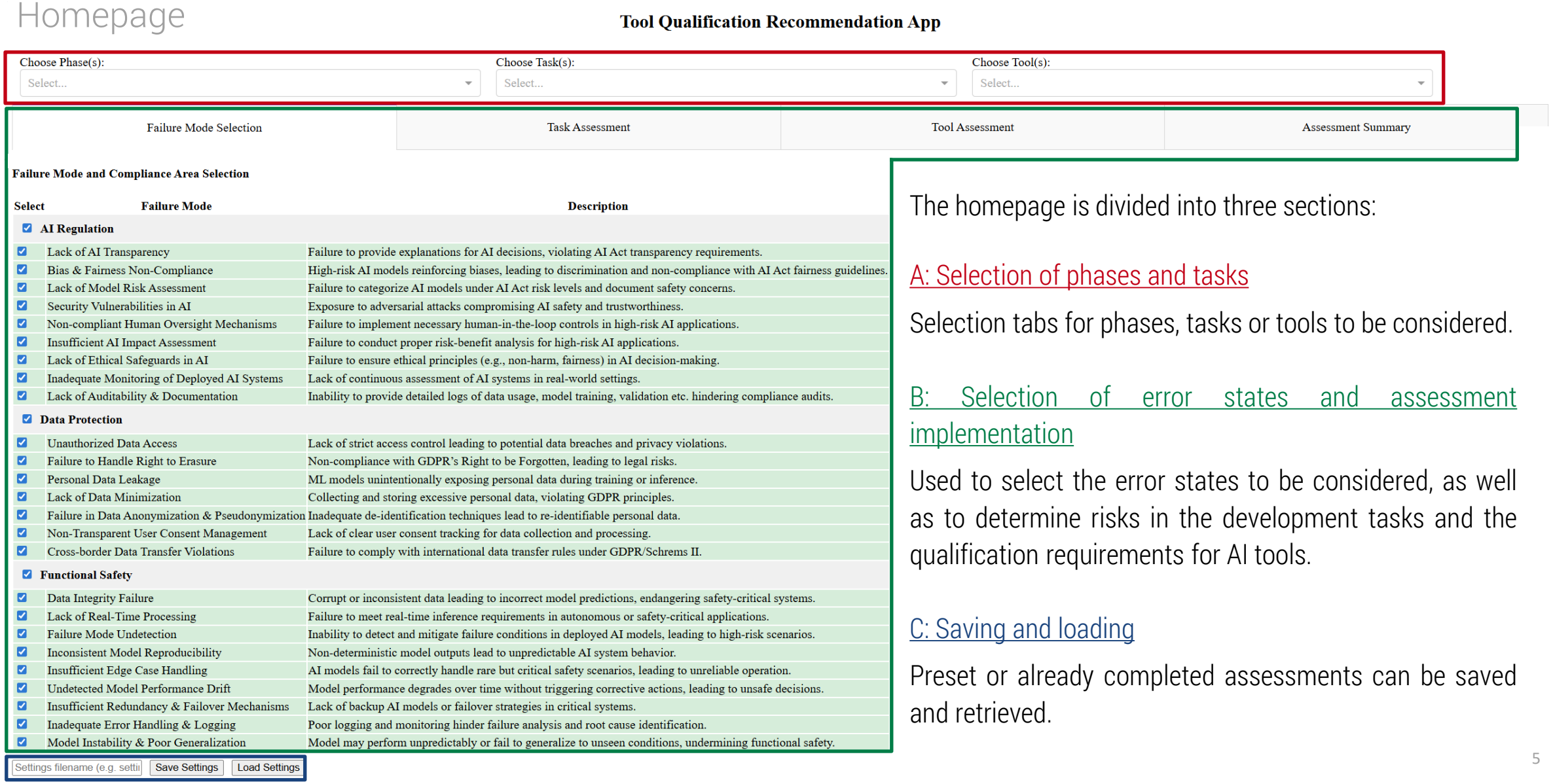
2.1 Assessment of the significance of the tool for an error (Tool Impact)	2.2 Evaluation of tool error detection and determination of the tool confidence level (TCL) for each conformity area	2.3 Qualification recommendation based on RPN and TCL



AssessML: structure of the dashboard

Automotive Quality Institute

Homepage



The homepage is divided into three sections:

A: Selection of phases and tasks

Selection tabs for phases, tasks or tools to be considered.

B: Selection of error states and assessment implementation

Used to select the error states to be considered, as well as to determine risks in the development tasks and the qualification requirements for AI tools.

C: Saving and loading

Preset or already completed assessments can be saved and retrieved.

AQI Automotive Quality Institute

1. Identification of risks in development tasks

1.1. Selection of relevant tasks phases

AssessML: selection area

What is the selection area used for?

The dashboard contains 25 error states that are selected from the start.

This automatically transfers all statuses to the task and tool assessment as well as to the assessment summary, which can quickly make the board confusing.

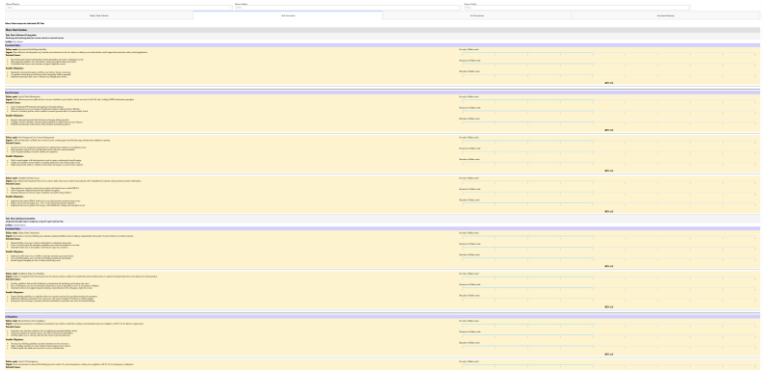
Failure Mode Selection | Task Assessment

Failure Mode and Compliance Area Selection

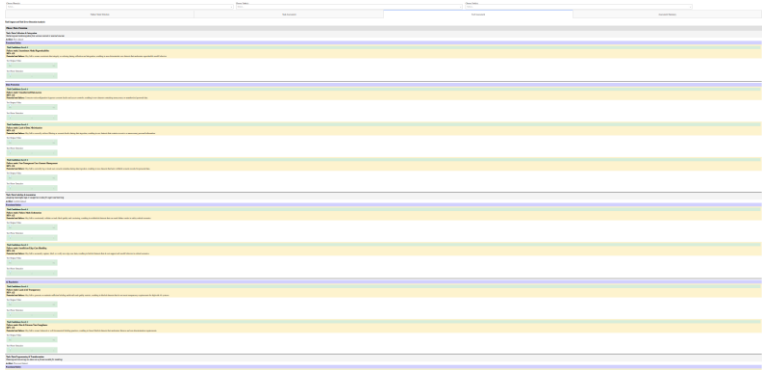
Select	Failure Mode	Description
☑	**AI Regulation**	
☑	Lack of AI Transparency	Failure to provide explanations for AI decisions, violating AI Act transparency requirements.
☑	Bias & Fairness Non-Compliance	High-risk AI models reinforcing biases, leading to discrimination and non-compliance with AI Act fairness guidelines.
☑	Lack of Model Risk Assessment	Failure to categorize AI models under AI Act risk levels and document safety concerns.
☑	Security Vulnerabilities in AI	Exposure to adversarial attacks compromising AI safety and trustworthiness.
☑	Non-compliant Human Oversight Mechanisms	Failure to implement necessary human-in-the-loop controls in high-risk AI applications.
☑	Insufficient AI Impact Assessment	Failure to conduct proper risk-benefit analysis for high-risk AI applications.
☑	Lack of Ethical Safeguards in AI	Failure to ensure ethical principles (e.g., non-harm, fairness) in AI decision-making.
☑	Inadequate Monitoring of Deployed AI Systems	Lack of continuous assessment of AI systems in real-world settings.
☑	Lack of Auditability & Documentation	Inability to provide detailed logs of data usage, model training, validation etc. hindering compliance audits.
☑	**Data Protection**	
☑	Unauthorized Data Access	Lack of strict access control leading to potential data breaches and privacy violations.
☑	Failure to Handle Right to Erasure	Non-compliance with GDPR's Right to be Forgotten, leading to legal risks.
☑	Personal Data Leakage	ML models unintentionally exposing personal data during training or inference.
☑	Lack of Data Minimization	Collecting and storing excessive personal data, violating GDPR principles.
☑	Failure in Data Anonymization & Pseudonymization	Inadequate de-identification techniques lead to re-identifiable personal data.
☑	Non-Transparent User Consent Management	Lack of clear user consent tracking for data collection and processing.
☑	Cross-border Data Transfer Violations	Failure to comply with international data transfer rules under GDPR/Schrems II.
☑	**Functional Safety**	
☑	Data Integrity Failure	Corrupt or inconsistent data leading to incorrect model predictions, endangering safety-critical systems.
☑	Lack of Real-Time Processing	Failure to meet real-time inference requirements in autonomous or safety-critical applications.
☑	Failure Mode Undetection	Inability to detect and mitigate failure conditions in deployed AI models, leading to high-risk scenarios.
☑	Inconsistent Model Reproducibility	Non-deterministic model outputs lead to unpredictable AI system behavior.
☑	Insufficient Edge Case Handling	AI models fail to correctly handle rare but critical safety scenarios, leading to unreliable operation.
☑	Undetected Model Performance Drift	Model performance degrades over time without triggering corrective actions, leading to unsafe decisions.
☑	Insufficient Redundancy & Failover Mechanisms	Lack of backup AI models or failover strategies in critical systems.
☑	Inadequate Error Handling & Logging	Poor logging and monitoring hinder failure analysis and root cause identification.
☑	Model Instability & Poor Generalization	Model may perform unpredictably or fail to generalize to unseen conditions, undermining functional safety.

Settings filename (e.g. setti | Save Settings | Load Settings

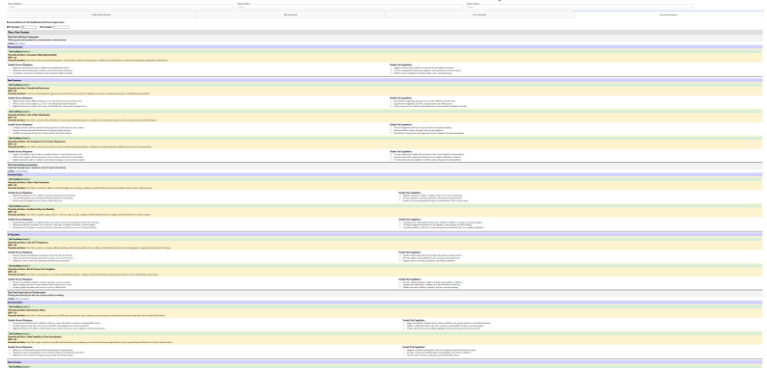
Task Assessment



Tool Assessment



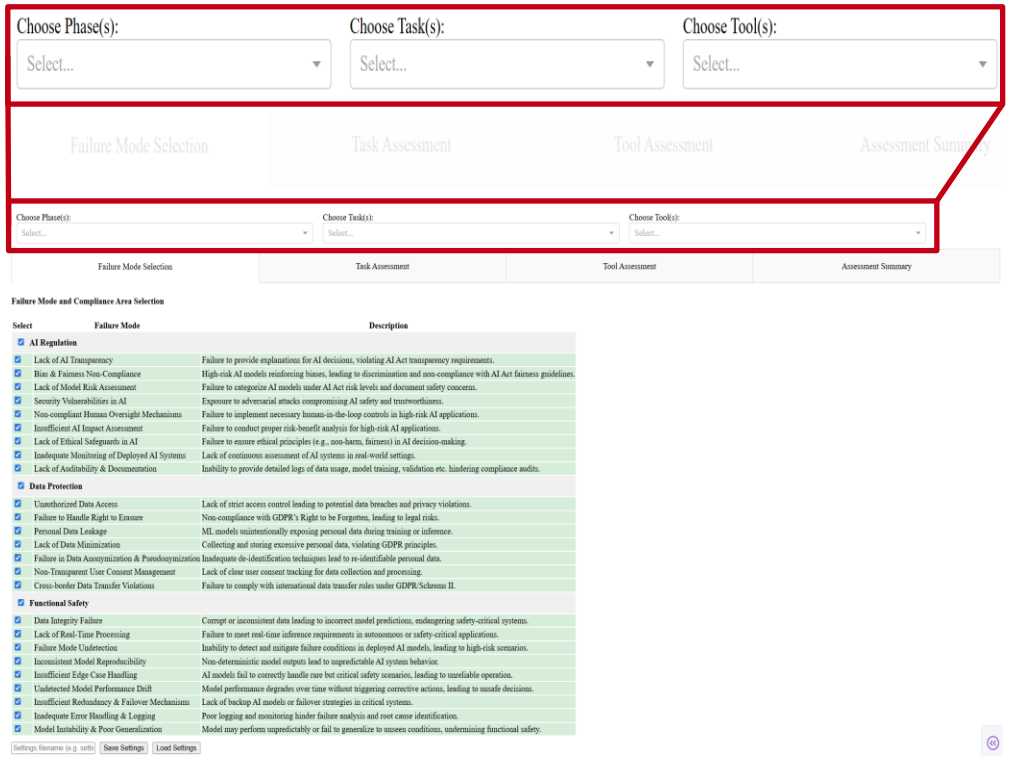
Assessment Summery



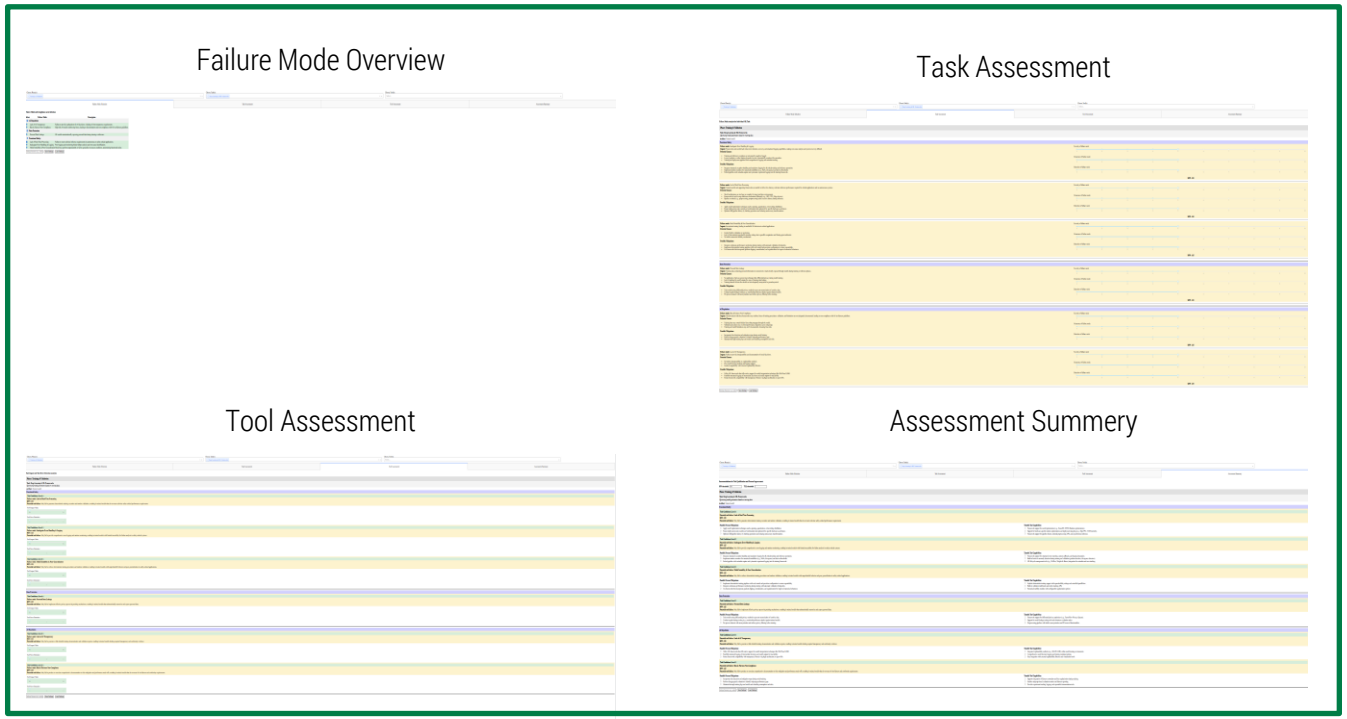
AssessML: selection area

What is the selection area used for?

To enable more targeted and efficient processing, the dashboard offers the option of filtering via the selection area.



Each selection made reduces the number of error states displayed, thereby simplifying the processing of the following steps in the assessments and in the Assessment Summary.



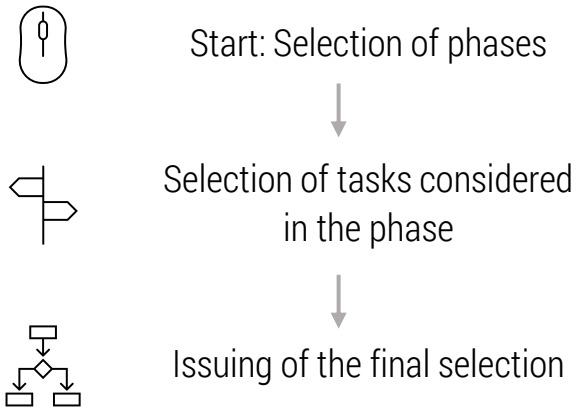
Instructions for assessments follow from slide 18 onwards

AssessML: selection area

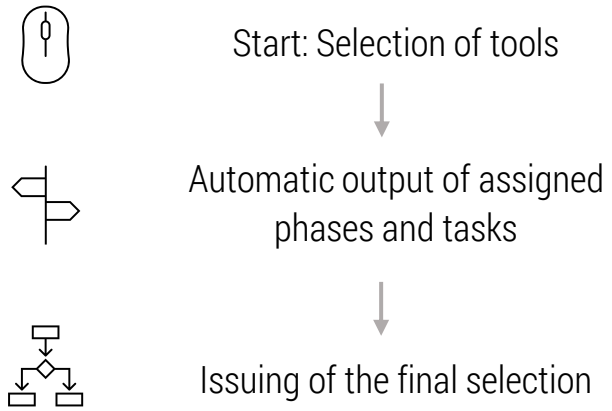
1. Selection of phases and tasks

The selection area is divided into three **search tabs**: *Phase, Task and Tools*, allowing users to limit their search to their preferred phases and tasks. There are **two** ways to specify a phase and task:

<u>1st approach: Phases and tasks</u>



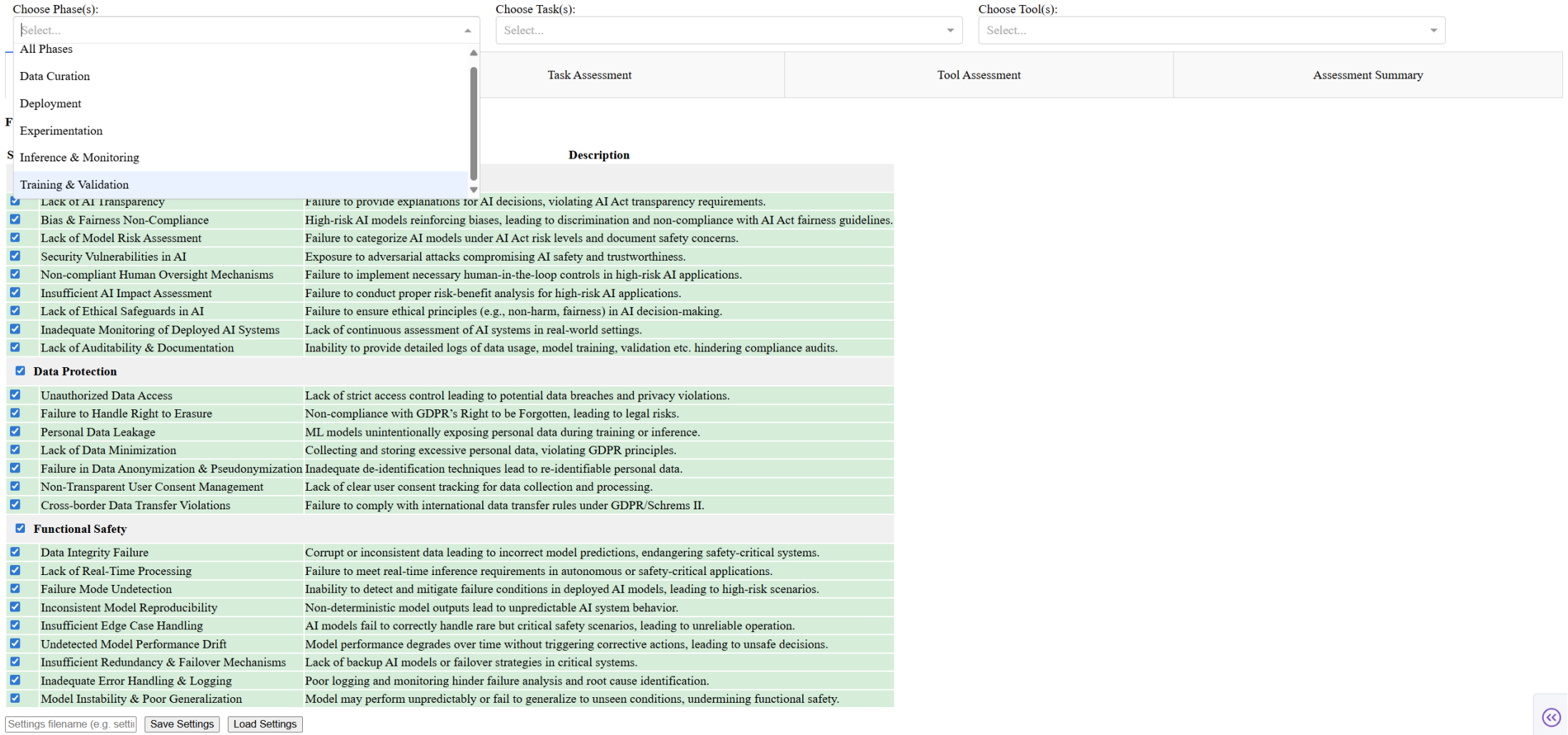
<u>2nd approach: Tools</u>



AssessML: selection area

1st approach: Phases and tasks

Start: Selection of phases

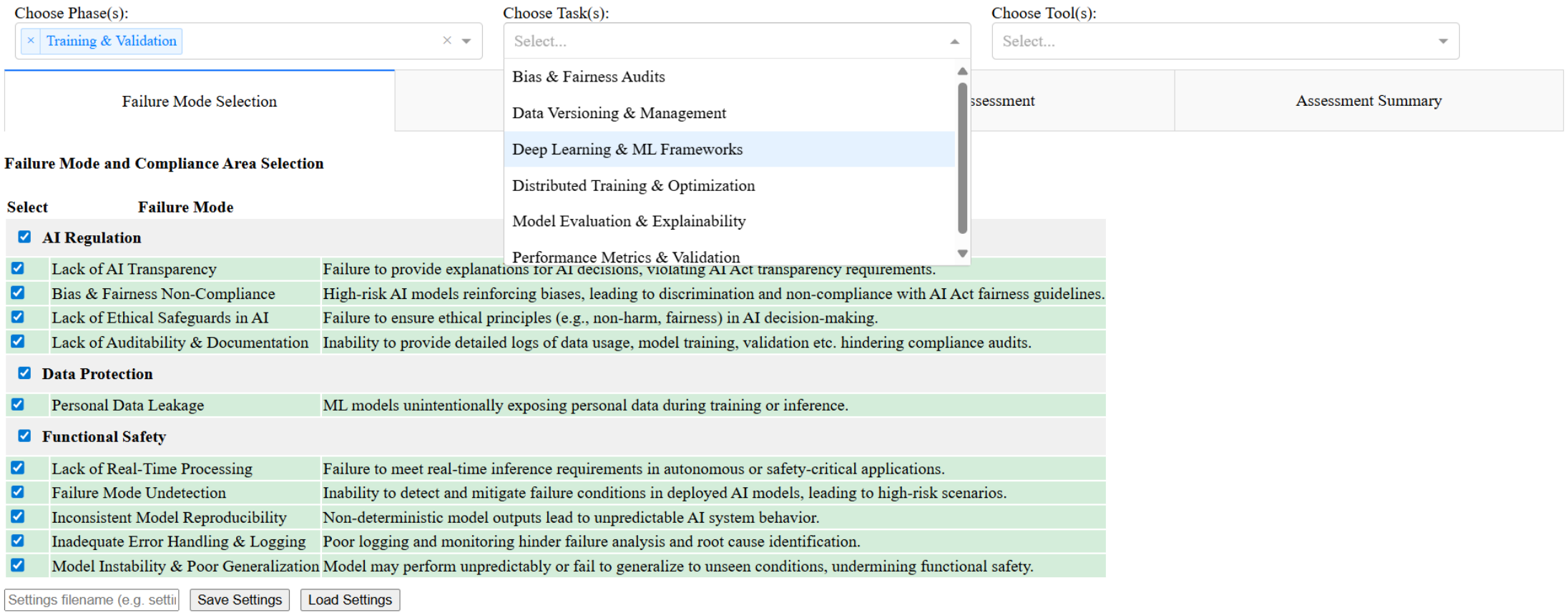


AssessML: selection area

1st approach: Phases and tasks

Automotive Quality Institute

Selection of tasks considered in the phase



Note:

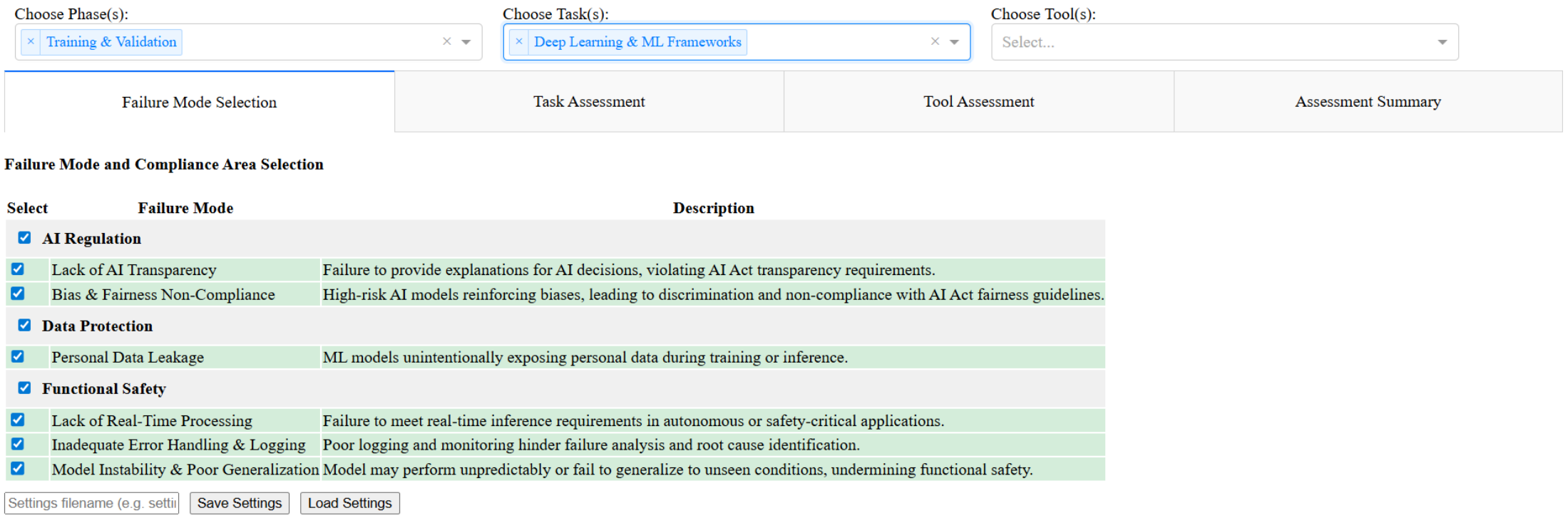
Selecting the phases automatically filters the fault conditions.

AssessML: selection area

1st approach: Phases and tasks

Automotive Quality Institute

Issuing of the final selection



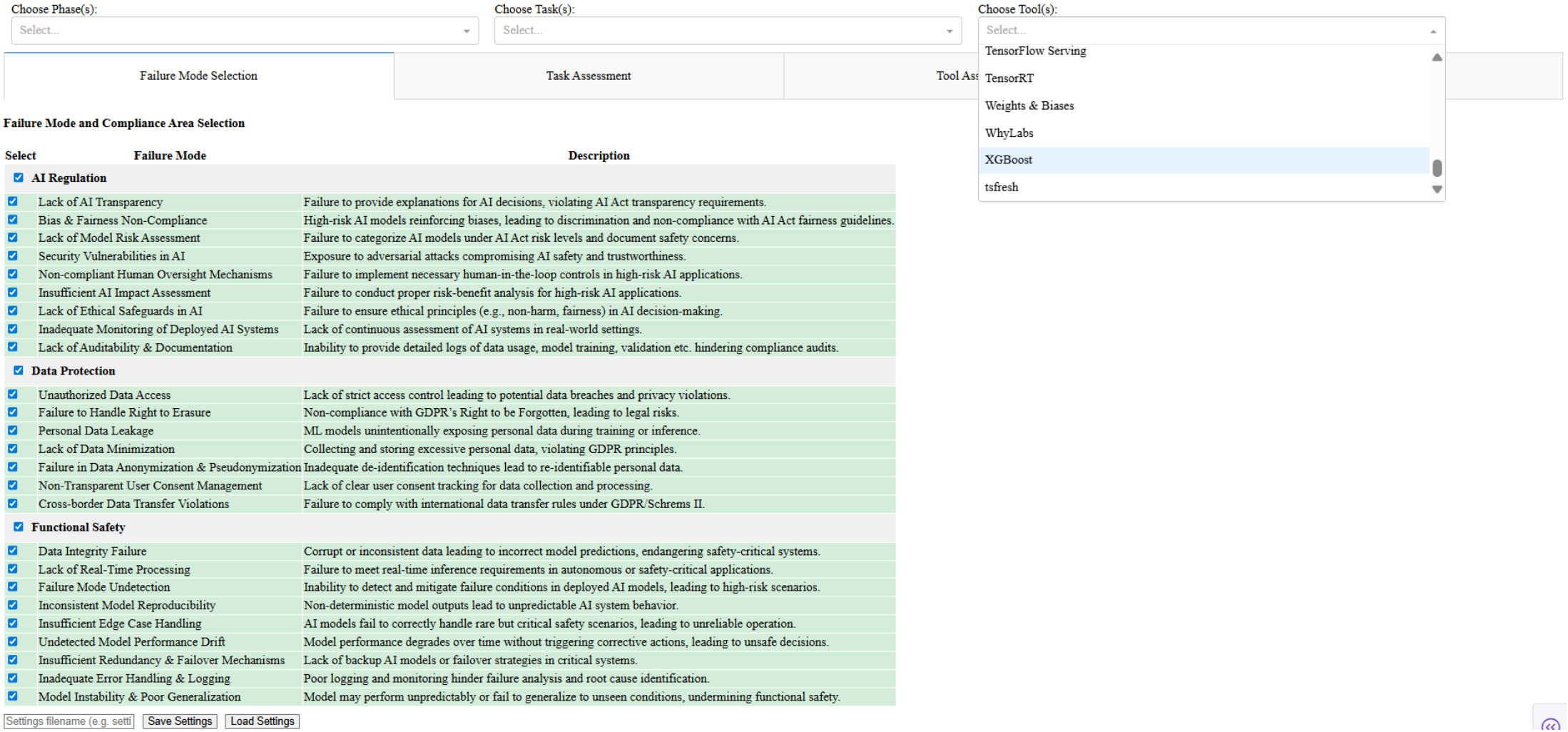
After going through the ‘Phases’ and ‘Tasks’ tabs from left to right, the dashboard displays all potential error states that remain after filtering. → The possible error states are restricted!

AssessML: selection area

2nd approach: Tools

Automotive Quality Institute

Start: Selection of tools used

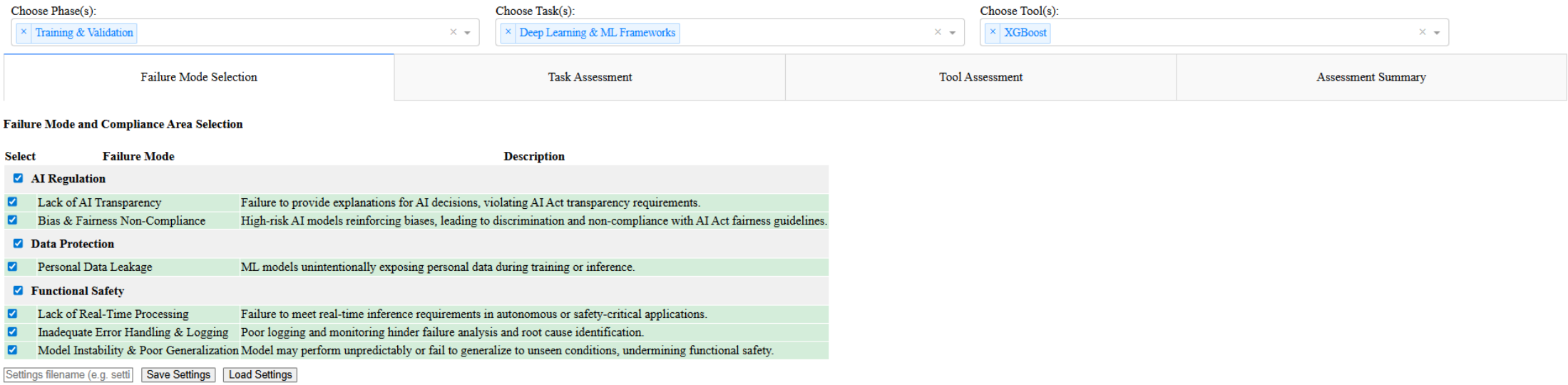


AssessML: selection area

2nd approach: Tools

AQI Automotive Quality Institute

Automatic output of assigned phases and tasks



After going through the ‘Tools’ tab, the dashboard displays all error states associated with the selected tool.

→ There is a specific restriction of error states!

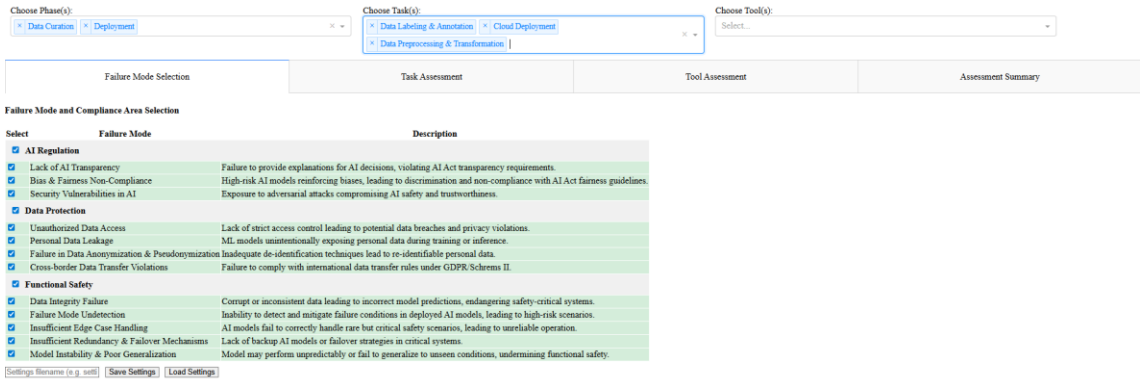
AssessML: selection area

Note

Multiple selections are also possible; this applies to **both** approaches!

1st approach: Phases and tasks

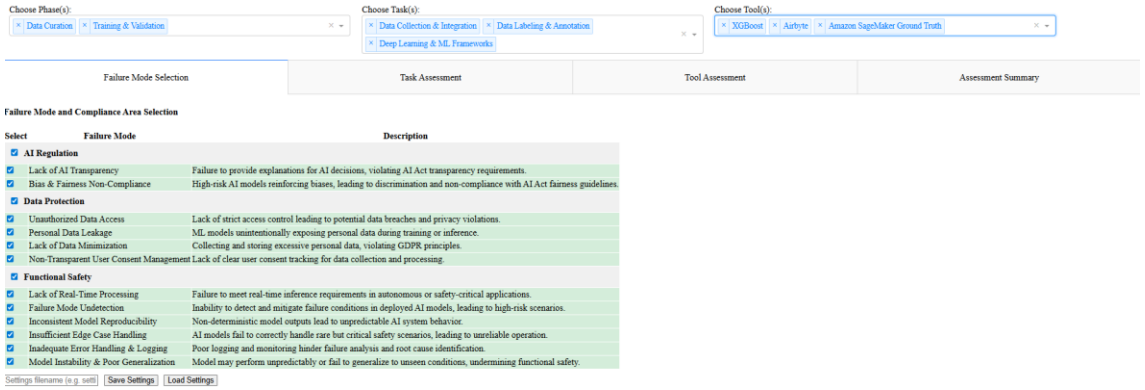
If several phases and tasks are selected, the error states that are relevant to all selected phases and tasks are displayed.



There is a restriction on the possible error states!

2. Approach: Tools

If several tools are selected, the error states that are relevant to all selected tools are displayed.



There is a specific restriction of error states!

AQI Automotive Quality Institute

1. Identification of risks in development tasks

1.2 Selection of area of compliance and applicable failure modes

AssessML: selection of failure modes

AQI Automotive Quality Institute

Structure

Failure Mode Selection | Task Assessment | Tool Assessment | Assessment Summary

Failure Mode and Compliance Area Selection

Select	Failure Mode	Description
☑	**AI Regulation**	
☑	Lack of AI Transparency	Failure to provide explanations for AI decisions, violating AI Act transparency requirements.
☑	Bias & Fairness Non-Compliance	High-risk AI models reinforcing biases, leading to discrimination and non-compliance with AI Act fairness guidelines.
☑	Lack of Model Risk Assessment	Failure to categorize AI models under AI Act risk levels and document safety concerns.
☑	Security Vulnerabilities in AI	Exposure to adversarial attacks compromising AI safety and trustworthiness.
☑	Non-compliant Human Oversight Mechanisms	Failure to implement necessary human-in-the-loop controls in high-risk AI applications.
☑	Insufficient AI Impact Assessment	Failure to conduct proper risk-benefit analysis for high-risk AI applications.
☑	Lack of Ethical Safeguards in AI	Failure to ensure ethical principles (e.g., non-harm, fairness) in AI decision-making.
☑	Inadequate Monitoring of Deployed AI Systems	Lack of continuous assessment of AI systems in real-world settings.
☑	Lack of Auditability & Documentation	Inability to provide detailed logs of data usage, model training, validation etc. hindering compliance audits.
☑	**Data Protection**	
☑	Unauthorized Data Access	Lack of strict access control leading to potential data breaches and privacy violations.
☑	Failure to Handle Right to Erasure	Non-compliance with GDPR's Right to be Forgotten, leading to legal risks.
☑	Personal Data Leakage	ML models unintentionally exposing personal data during training or inference.
☑	Lack of Data Minimization	Collecting and storing excessive personal data, violating GDPR principles.
☑	Failure in Data Anonymization & Pseudonymization	Inadequate de-identification techniques lead to re-identifiable personal data.
☑	Non-Transparent User Consent Management	Lack of clear user consent tracking for data collection and processing.
☑	Cross-border Data Transfer Violations	Failure to comply with international data transfer rules under GDPR/Schrems II.
☑	**Functional Safety**	
☑	Data Integrity Failure	Corrupt or inconsistent data leading to incorrect model predictions, endangering safety-critical systems.
☑	Lack of Real-Time Processing	Failure to meet real-time inference requirements in autonomous or safety-critical applications.
☑	Failure Mode Undetection	Inability to detect and mitigate failure conditions in deployed AI models, leading to high-risk scenarios.
☑	Inconsistent Model Reproducibility	Non-deterministic model outputs lead to unpredictable AI system behavior.
☑	Insufficient Edge Case Handling	AI models fail to correctly handle rare but critical safety scenarios, leading to unreliable operation.
☑	Undetected Model Performance Drift	Model performance degrades over time without triggering corrective actions, leading to unsafe decisions.
☑	Insufficient Redundancy & Failover Mechanisms	Lack of backup AI models or failover strategies in critical systems.
☑	Inadequate Error Handling & Logging	Poor logging and monitoring hinder failure analysis and root cause identification.
☑	Model Instability & Poor Generalization	Model may perform unpredictably or fail to generalize to unseen conditions, undermining functional safety.

The assessment section is divided into four slides:

- Failure Mode Selection (List of error conditions),
- Task Assessment (Identification of risks in development tasks),
- Tool Assessment (Training requirements for AI tools) and
- Assessment Summery (Qualification summary based on the RPN and the TCL)

AssessML: selection of failure modes

AQI Automotive Quality Institute

Fine selection - Failure Mode Overview

Failure Mode Selection | Task Assessment | Tool Assessment | Assessment Summary

Failure Mode and Compliance Area Selection

Select	Failure Mode	Description
☑	**AI Regulation**	
☑	Lack of AI Transparency	Failure to provide explanations for AI decisions, violating AI Act transparency requirements.
☑	Bias & Fairness Non-Compliance	High-risk AI models reinforcing biases, leading to discrimination and non-compliance with AI Act fairness guidelines.
☑	Lack of Model Risk Assessment	Failure to categorize AI models under AI Act risk levels and document safety concerns.
☑	Security Vulnerabilities in AI	Exposure to adversarial attacks compromising AI safety and trustworthiness.
☑	Non-compliant Human Oversight Mechanisms	Failure to implement necessary human-in-the-loop controls in high-risk AI applications.
☑	Insufficient AI Impact Assessment	Failure to conduct proper risk-benefit analysis for high-risk AI applications.
☑	Lack of Ethical Safeguards in AI	Failure to ensure ethical principles (e.g., non-harm, fairness) in AI decision-making.
☑	Inadequate Monitoring of Deployed AI Systems	Lack of continuous assessment of AI systems in real-world settings.
☑	Lack of Auditability & Documentation	Inability to provide detailed logs of data usage, model training, validation etc. hindering compliance audits.
☑	**Data Protection**	
☑	Unauthorized Data Access	Lack of strict access control leading to potential data breaches and privacy violations.
☑	Failure to Handle Right to Erasure	Non-compliance with GDPR's Right to be Forgotten, leading to legal risks.
☑	Personal Data Leakage	ML models unintentionally exposing personal data during training or inference.
☑	Lack of Data Minimization	Collecting and storing excessive personal data, violating GDPR principles.
☑	Failure in Data Anonymization & Pseudonymization	Inadequate de-identification techniques lead to re-identifiable personal data.
☑	Non-Transparent User Consent Management	Lack of clear user consent tracking for data collection and processing.
☑	Cross-border Data Transfer Violations	Failure to comply with international data transfer rules under GDPR/Schrems II.
☑	**Functional Safety**	
☑	Data Integrity Failure	Corrupt or inconsistent data leading to incorrect model predictions, endangering safety-critical systems.
☑	Lack of Real-Time Processing	Failure to meet real-time inference requirements in autonomous or safety-critical applications.
☑	Failure Mode Undetection	Inability to detect and mitigate failure conditions in deployed AI models, leading to high-risk scenarios.
☑	Inconsistent Model Reproducibility	Non-deterministic model outputs lead to unpredictable AI system behavior.
☑	Insufficient Edge Case Handling	AI models fail to correctly handle rare but critical safety scenarios, leading to unreliable operation.
☑	Undetected Model Performance Drift	Model performance degrades over time without triggering corrective actions, leading to unsafe decisions.
☑	Insufficient Redundancy & Failover Mechanisms	Lack of backup AI models or failover strategies in critical systems.
☑	Inadequate Error Handling & Logging	Poor logging and monitoring hinder failure analysis and root cause identification.
☑	Model Instability & Poor Generalization	Model may perform unpredictably or fail to generalize to unseen conditions, undermining functional safety.

The Failure Mode Selection section shows:

- List of all error states,
- Explanation for each error condition

and is used to select the desired/required error states of the three compatibility areas.

AssessML: selection of failure modes

Automotive Quality Institute

Fine selection - Failure Mode Overview

For a detailed search, the user can apply further filters using the 'Select boxes':

- If you make a selection for the individual failure modes, the modes that are not taken into account are highlighted in **white**.
- If a selection is made for an entire compatibility area, the area that is not taken into account is highlighted in **grey**.

Select the failure modes by clicking on the 'Select' boxes **for each failure mode** and/or **for each compliance area**.

Select	Failure Mode
☑	**AI Regulation**
☑	Lack of AI Transparency
☑	Bias & Fairness Non-Compliance
☑	Lack of Model Risk Assessment
☑	Security Vulnerabilities in AI
☐	Non-compliant Human Oversight Mechanisn
☐	Insufficient AI Impact Assessment
☑	Lack of Ethical Safeguards in AI
☑	Inadequate Monitoring of Deployed AI Syste
☐	Lack of Auditability & Documentation
☑	**Data Protection**
☑	Unauthorized Data Access
☐	Failure to Handle Right to Erasure
☑	Personal Data Leakage
☑	Lack of Data Minimization
☐	Failure in Data Anonymization & Pseudonyn
☐	Non-Transparent User Consent Management
☑	Cross-border Data Transfer Violations
☑	**Functional Safety**
☑	Data Integrity Failure
☑	Lack of Real-Time Processing
☐	Failure Mode Undetection
☑	Inconsistent Model Reproducibility

Select	Failure Mode	
☐	**AI Regulation**	
☑	Lack of AI Transparency	Fail
☑	Bias & Fairness Non-Compliance	Hig
☑	Lack of Model Risk Assessment	Fail
☑	Security Vulnerabilities in AI	Exp
☑	Non-compliant Human Oversight Mechanisms	Fail
☑	Insufficient AI Impact Assessment	Fail
☑	Lack of Ethical Safeguards in AI	Fail
☑	Inadequate Monitoring of Deployed AI Systems	Lac
☑	Lack of Auditability & Documentation	Inat
☑	**Data Protection**	
☑	Unauthorized Data Access	Lac
☑	Failure to Handle Right to Erasure	Nor
☑	Personal Data Leakage	ML
☑	Lack of Data Minimization	Coll
☑	Failure in Data Anonymization & Pseudonymization	Inac
☑	Non-Transparent User Consent Management	Lac
☑	Cross-border Data Transfer Violations	Fail
☐	**Functional Safety**	
☑	Data Integrity Failure	Cor
☑	Lack of Real-Time Processing	Fail
☑	Failure Mode Undetection	Inat

AQI Automotive Quality Institute

1. Identification of risks in development tasks

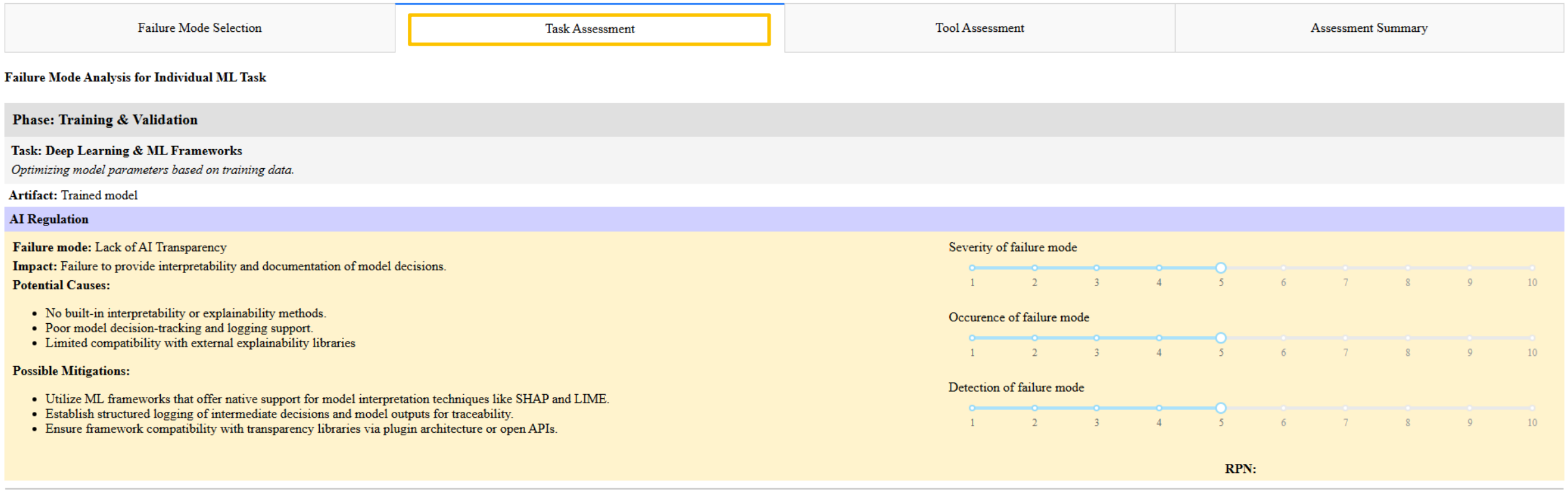
1.3 Calculation of the RPN for selected failure modes in a task

AssessML: evaluation of failure modes

AQI Automotive Quality Institute

Task Assessment - Calculation of the RPN for selected failure modes

The Task Assessment area is used to calculate the risk priority number **(RPN)** for each selected fault condition and/or compatibility area.



AssessML: evaluation of failure modes

AQI Automotive Quality Institute

Task Assessment - Calculation of the RPN for selected failure modes

The RPN is the sum of the three indicators:

- Severity (Significance of the error sequence),
- Occurence (Probability of occurrence of the cause of the fault) and
- Detection (Probability of discovering the error or its cause)

RPN = Severity * Occurrence * Detection

Phase: Training & Validation

Task: Deep Learning & ML Frameworks

Optimizing model parameters based on training data.

Artifact: Trained model

Risk of Deep Learning & ML Frameworks in AI Regulation

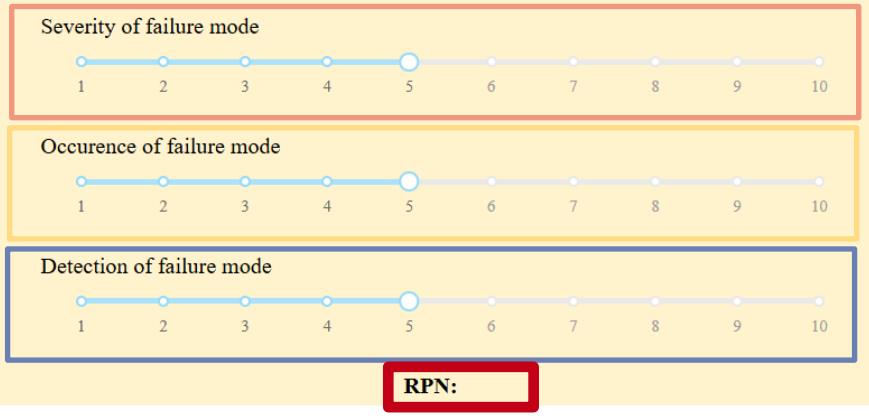
Failure mode: Lack of AI Transparency: Failure to provide interpretability and documentation of model decisions.

Potential Causes:

- No built-in interpretability or explainability methods.
- Poor model decision-tracking and logging support.
- Limited compatibility with external explainability libraries

Possible Mitigations:

- Utilize ML frameworks that offer native support for model interpretation techniques like SHAP and LIME.
- Establish structured logging of intermediate decisions and model outputs for traceability.
- Ensure framework compatibility with transparency libraries via plugin architecture or open APIs.



AssessML: evaluation of failure modes

Automotive Quality Institute

Significance of the failure mode (Severity)

Action:

Assess the severity for the failure modes assigned to the selected ML tasks in the tool.

Phase: Training & Validation

Task: Deep Learning & ML Frameworks

Optimizing model parameters based on training data.

Artifact: Trained model

Risk of Deep Learning & ML Frameworks in AI Regulation

Failure mode: Lack of AI Transparency: Failure to provide interpretability and documentation of model decisions.

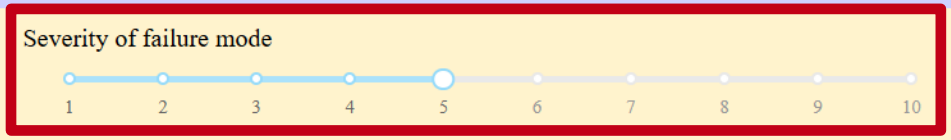
Potential Causes:

- No built-in interpretability or explainability methods.
- Poor model decision-tracking and logging support.
- Limited compatibility with external explainability libraries

Possible Mitigations:

- Utilize ML frameworks that offer native support for model interpretation techniques like SHAP and LIME.
- Establish structured logging of intermediate decisions and model outputs for traceability.
- Ensure framework compatibility with transparency libraries via plugin architecture or open APIs.

By moving the slider, the values 1 – 3: 4 – 6; 7 – 10 can be set.



RPN:

AssessML: evaluation of failure modes

AQI | Automotive Quality Institute

Probability of occurrence of the cause of the failure mode (Occurence)

Action:

Assess the **probability of occurrence** for the failure modes assigned to the selected ML tasks in the tool.

Phase: Training & Validation

Task: Deep Learning & ML Frameworks

Optimizing model parameters based on training data.

Artifact: Trained model

Risk of Deep Learning & ML Frameworks in AI Regulation

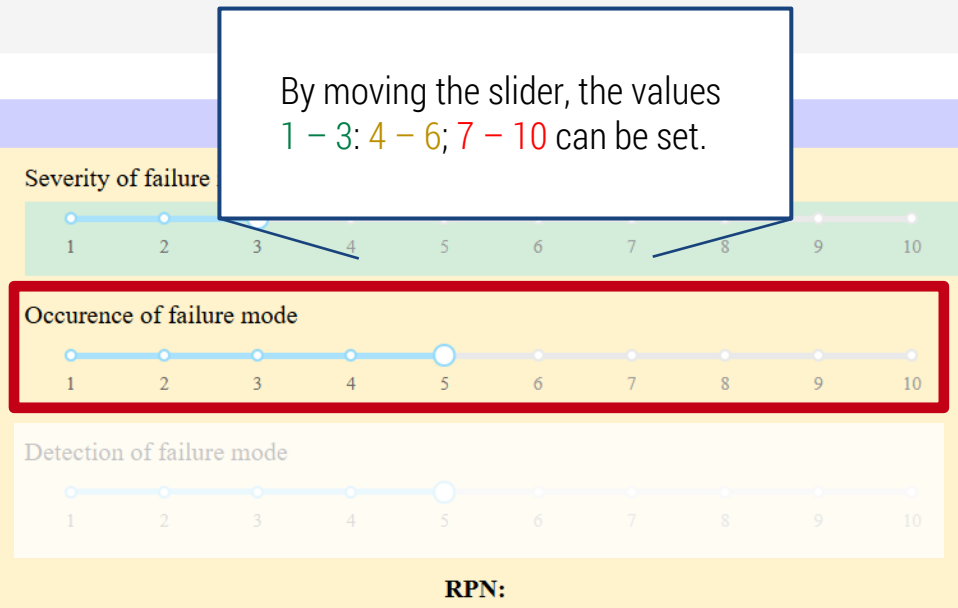
Failure mode: Lack of AI Transparency: Failure to provide interpretability and documentation of model decisions.

Potential Causes:

- No built-in interpretability or explainability methods.
- Poor model decision-tracking and logging support.
- Limited compatibility with external explainability libraries

Possible Mitigations:

- Utilize ML frameworks that offer native support for model interpretation techniques like SHAP and LIME.
- Establish structured logging of intermediate decisions and model outputs for traceability.
- Ensure framework compatibility with transparency libraries via plugin architecture or open APIs.



AssessML: evaluation of failure modes

AQI Automotive Quality Institute

Probability of discovering the failure mode or its cause (Detectability)

Action:

Assess the **detectability** for the failure modes assigned to the selected ML tasks in the tool.

Phase: Training & Validation

Task: Deep Learning & ML Frameworks

Optimizing model parameters based on training data.

Artifact: Trained model

Risk of Deep Learning & ML Frameworks in AI Regulation

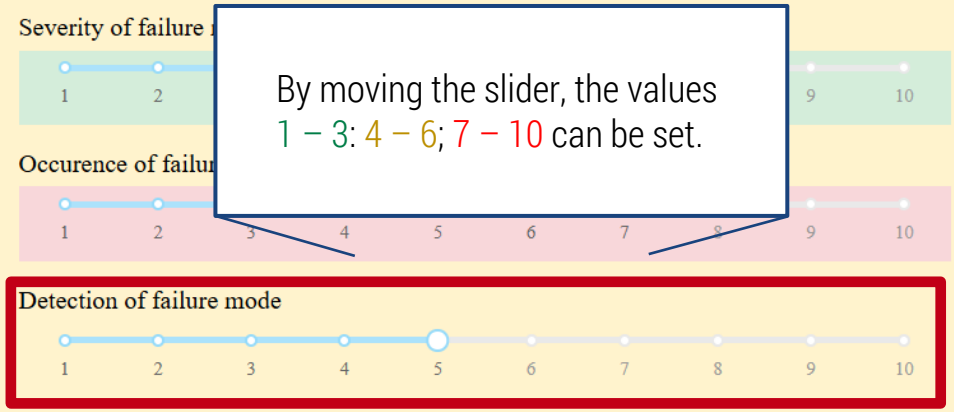
Failure mode: Lack of AI Transparency: Failure to provide interpretability and documentation of model decisions.

Potential Causes:

- No built-in interpretability or explainability methods.
- Poor model decision-tracking and logging support.
- Limited compatibility with external explainability libraries

Possible Mitigations:

- Utilize ML frameworks that offer native support for model interpretation techniques like SHAP and LIME.
- Establish structured logging of intermediate decisions and model outputs for traceability.
- Ensure framework compatibility with transparency libraries via plugin architecture or open APIs.



RPN:

AssessML: evaluation of failure modes

AQI | Automotive Quality Institute

Calculation of the RPN

Action:

Calculation of the RPN for the failure modes assigned to the selected ML tasks in the tool.

Phase: Training & Validation

Task: Deep Learning & ML Frameworks

Optimizing model parameters based on training data.

Artifact: Trained model

Risk of Deep Learning & ML Frameworks in AI Regulation

Failure mode: Lack of AI Transparency: Failure to provide interpretability and documentation of model decisions.

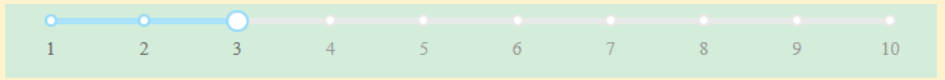
Potential Causes:

- No built-in interpretability or explainability methods.
- Poor model decision-tracking and logging support.
- Limited compatibility with external explainability libraries

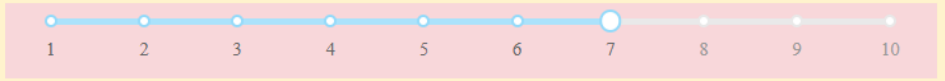
Possible Mitigations:

- Utilize ML frameworks that offer native support for model interpretation techniques like SHAP and LIME.
- Establish structured logging of intermediate decisions and model outputs for traceability.
- Ensure framework compatibility with transparency libraries via plugin architecture or open APIs.

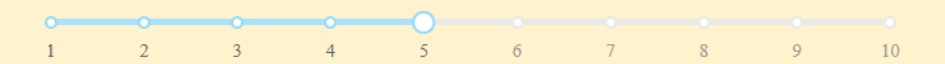
Severity of failure mode



Occurence of failure mode



Detection of failure mode



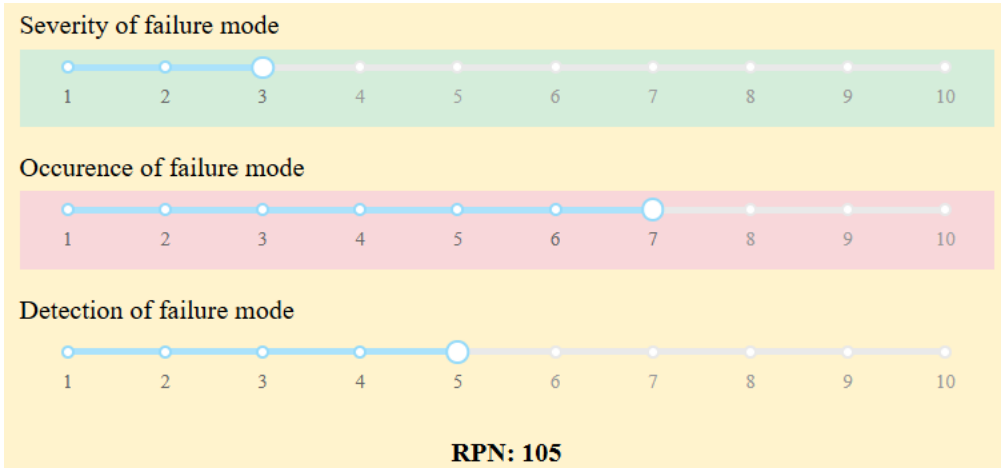
RPN: 105

Calculation of the RPN (1 – 29; 30 – 299; 300 -1000).

AssessML: evaluation of failure modes

Task Assessment - Calculation of the RPN for selected failure modes in a task

RPN = Severity * Occurrence * Detection



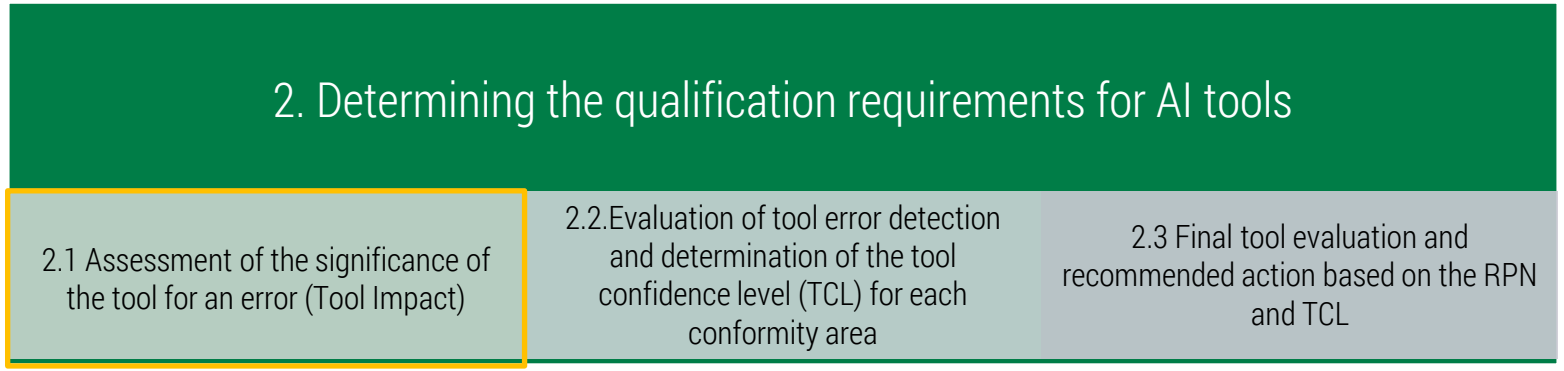
Evaluation of the **risk priority number** (RPN):

- Low RPN values (RPN 1–30) indicate low criticality and no need for action
- Medium RPN values (RPN 31–299) indicate medium criticality and thus a need for action
- High values (RPN 300–1000) indicate high criticality and a greater need for action.

→ The determination of risks (RPN) in the development tasks (according to FMEA) focuses on the determination <u>per use case</u>

2. Determining the qualification requirements for AI tools

2.1 Assessment of the significance of the tool for an error (Tool Impact)

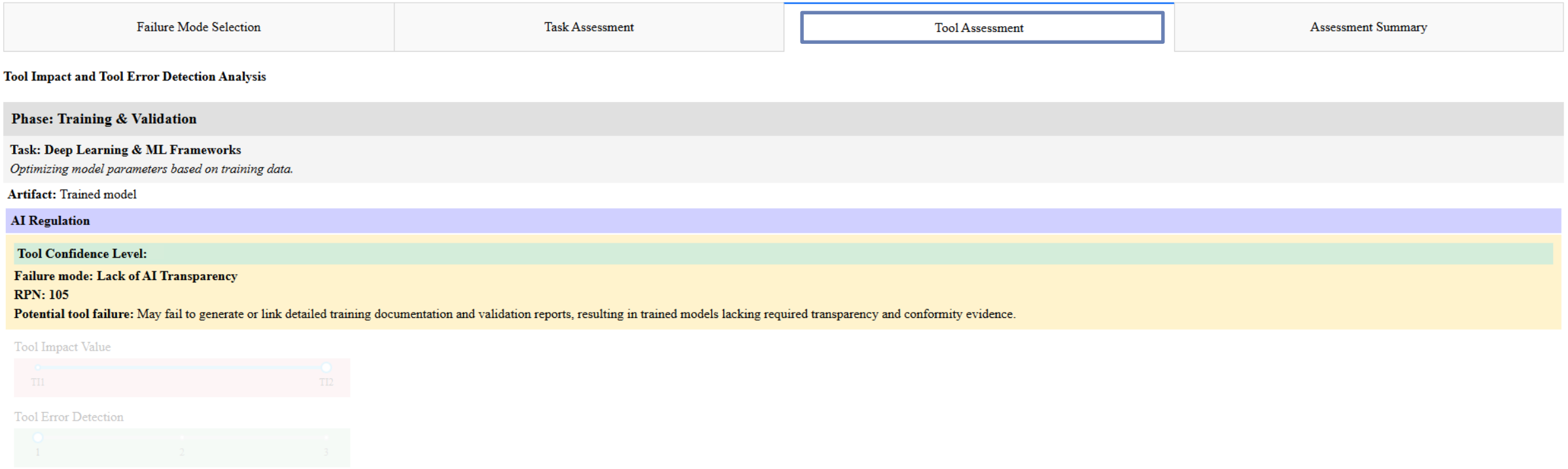


AssessML: impact of a tool failure

Tool Assessment - Identifying training needs for AI tools

AQI | Automotive Quality Institute

The Tool Assessment area is used to determine the qualification requirements for AI tools for each selected error state and/or compatibility area.



AssessML: impact of a tool failure

Effectiveness of the tool

Automotive Quality Institute

Failure Mode Selection | Task Assessment | **Tool Assessment** | Assessment Summary

Tool Impact and Tool Error Detection Analysis

Phase: Training & Validation

Task: Deep Learning & ML Frameworks

Optimizing model parameters based on training data.

Artifact: Trained model

AI Regulation

Tool Confidence Level:

Failure mode: Lack of AI Transparency

RPN: 105

Potential tool failure: May fail to generate or link detailed training documentation and validation reports, resulting in trained models lacking required transparency and conformity evidence.

Tool Impact Value

TI1 TI2

Tool Error Detection

1 2 3

Based on section 11.4.5.2 of ISO 26262:2018, the **Tool Confidence Level (TCL)** is determined based on the following three parts:

- **Tool Impact (TI)**, assesses whether a tool error could be safety-relevant,
- **Tool Error Detection (TD)** , assesses the probability that an error will be detected and results in the
- **Tool Confidence Level (TCL)**, the required level of confidence in the tool (TCL1 to TCL3).

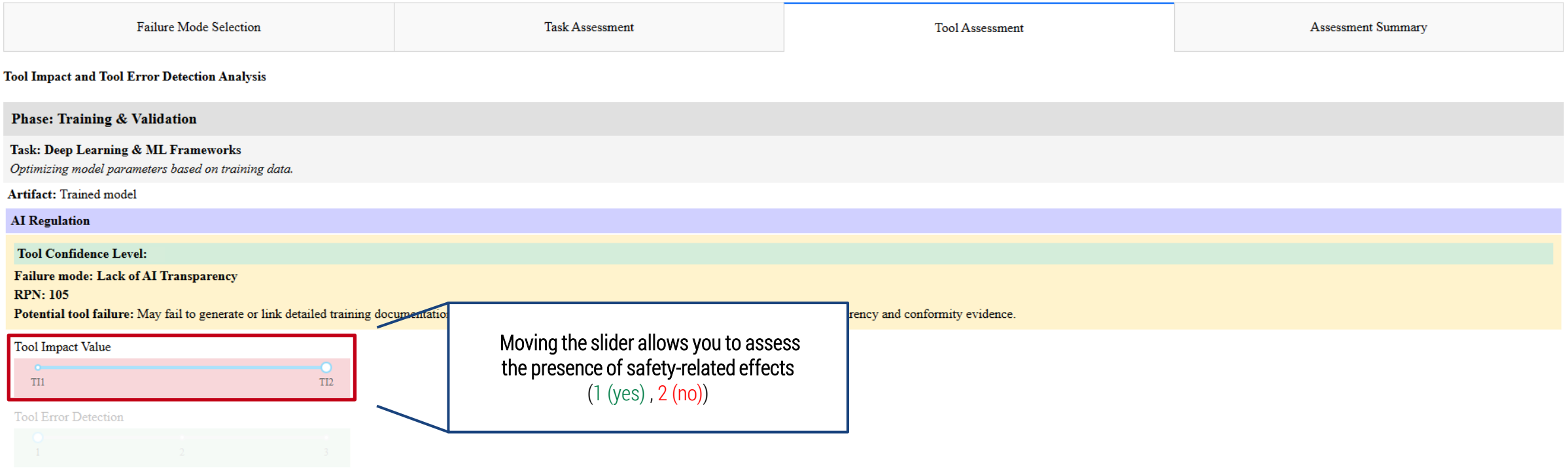
Source: Own representation based on ISO 26262:2018 Section 11.4.5.2

AssessML: impact of a tool failure

Assessment of the significance of the tool for an error (Tool Impact)

Automotive Quality Institute

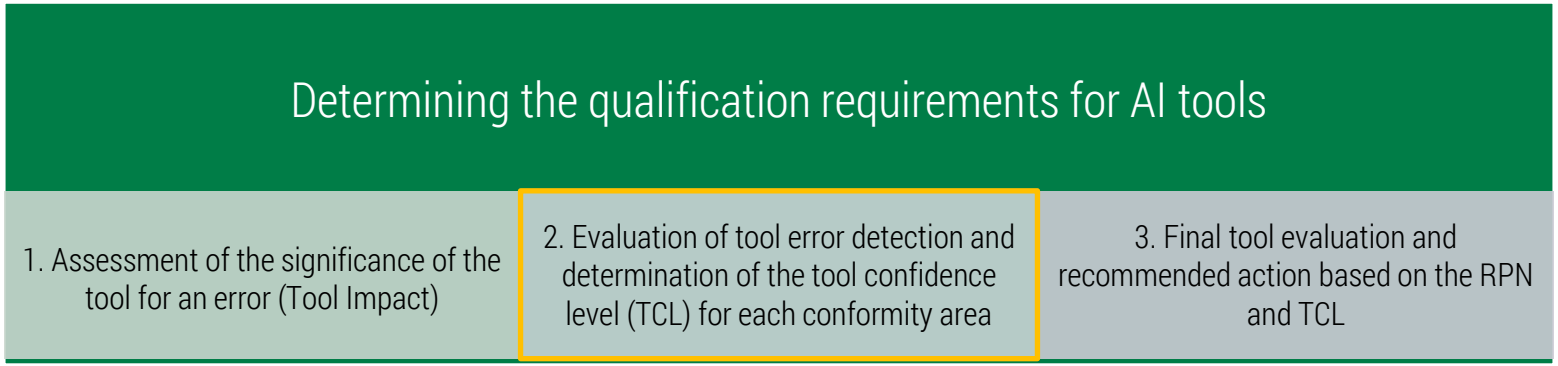
Action: Does a potential error in the tool affect the functionality of safety-related systems?



AQI Automotive Quality Institute

2. Determining the qualification requirements for AI tools

2.2. Evaluation of tool error detection and TCL calculation

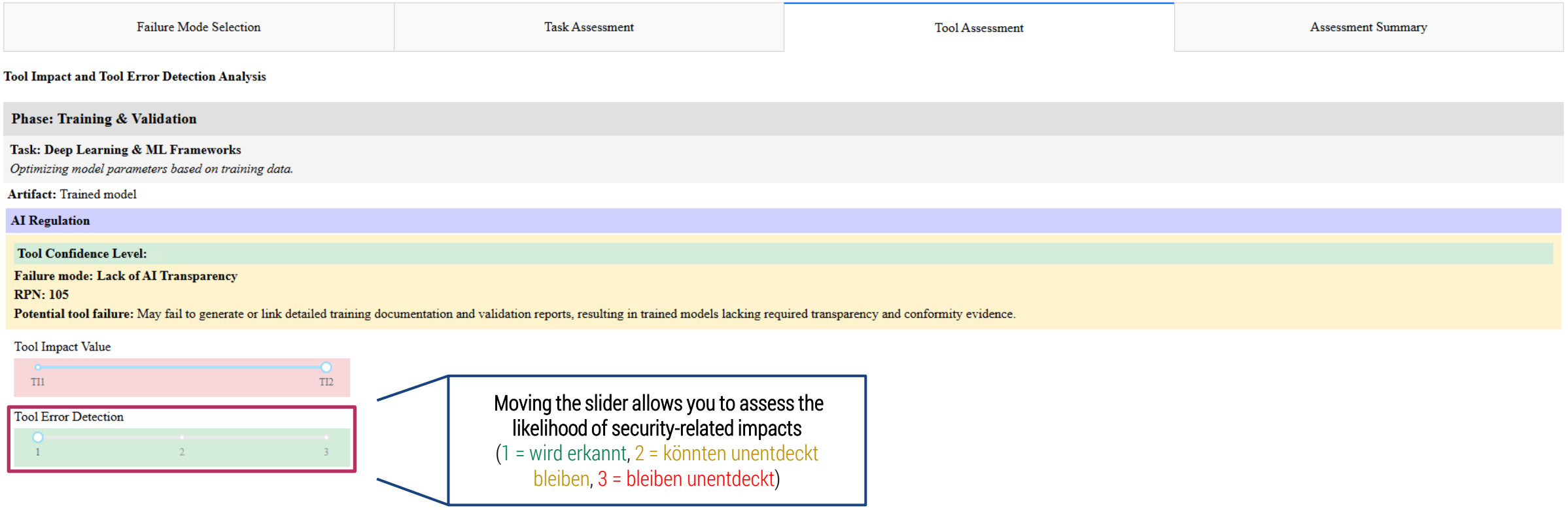


*AssessML: d*etectability of tool error

AQI | Automotive Quality Institute

Assessment of tool error detectability (Tool Error Detection)

Action: How well can errors caused by the tool be detected in the further development process?



Background: calculation of TCL

AQI Automotive Quality Institute

Need for trust in tool usage

Starting point: How high must the level of trust in the tool be (or is tool qualification necessary)?

Table 3 — Determination of the Tool Confidence Level (TCL)

		Tool error detection		
		TD1	TD2	TD3
Tool impact	TI1	TCL1	TCL1	TCL1
	TI2	TCL1	TCL2	TCL3

Source: ISO 26262:2018, Section. 11.4.5.4

Aspect	Picture	Description
Target	***Assignment of the appropriate Tool Confidence Level (TCL)*** *based on Tool Impact (TI) and Tool Error Detection (TD)*	*Determine* ***how high the level of trust*** *in the tool must be (or whether tool qualification is necessary)*
TCL1 (low)	▪ Tool has no safety-related influence (TI1) or ▪ Errors are reliably detected (TI2 + TD1) → **No tool qualification necessary**	▪ **Low significance** for product quality ▪ No tool qualification necessary → Confidence in tool behaviour from an ISO 26262 perspective **not critical**
TCL2 (medium)	▪ Tool has safety-related influence (TI2) ▪ Error **not reliably detectable (TD2)**	▪ Tool **important for product quality** ▪ Tool qualification required ▪ Requirements depend on ASIL level
TCL3 (high)	▪ Requirements depend on ASIL level (TI2) ▪ Errors **remain undetected (TD3)**	▪ **High importance** for product quality ▪ **Tool qualification absolutely necessary**, high requirements ▪ Differences between TCL2 and TCL3 are methodologically relevant but not dramatic (depending on ASIL)

Background: calculation of TCL

AQI Automotive Quality Institute

Need for trust in tool usage

Starting point: How high must the level of trust in the tool be (or is tool qualification necessary)?

Table 3 — Determination of the Tool Confidence Level (TCL)

		Tool error detection		
		TD1	TD2	TD3
Tool impact	TI1	TCL1	TCL1	TCL1
	TI2	TCL1	TCL2	TCL3

Source: ISO 26262:2018, Section. 11.4.5.4

Aspect	Picture	Description
Target	***Assignment of the appropriate Tool Confidence Level (TCL)*** *based on Tool Impact (TI) and Tool Error Detection (TD)*	*Determine* ***how high the level of trust*** *in the tool must be (or whether tool qualification is necessary)*
TCL1 (low)	▪ Tool has no safety-related influence (TI1) or ▪ Errors are reliably detected (TI2 + TD1) → No tool qualification necessary	▪ **Low significance** for product quality ▪ No tool qualification necessary → Confidence in tool behaviour from an ISO 26262 perspective **not critical**
TCL2 (medium)	▪ Tool has safety-related influence (TI2) ▪ Error **not reliably detectable (TD2)**	▪ Tool **important for product quality** ▪ Tool qualification required ▪ Requirements depend on ASIL level
TCL3 (high)	▪ Requirements depend on ASIL level (TI2) ▪ Errors **remain undetected (TD3)**	▪ **High importance** for product quality ▪ **Tool qualification absolutely necessary**, high requirements ▪ Differences between TCL2 and TCL3 are methodologically relevant but not dramatic (depending on ASIL)

Background: calculation of TCL

AQI Automotive Quality Institute

Need for trust in tool usage

Starting point: How high must the level of trust in the tool be (or is tool qualification necessary)?

Table 3 — Determination of the Tool Confidence Level (TCL)

		Tool error detection		
		TD1	TD2	TD3
Tool impact	TI1	TCL1	TCL1	TCL1
	TI2	TCL1	TCL2	TCL3

Source: ISO 26262:2018, Section. 11.4.5.4

Aspect	Picture	Description
Target	***Assignment of the appropriate Tool Confidence Level (TCL)*** *based on Tool Impact (TI) and Tool Error Detection (TD)*	*Determine* ***how high the level of trust*** *in the tool must be (or whether tool qualification is necessary)*
TCL1 (low)	▪ Tool has no safety-related influence (TI1) or ▪ Errors are reliably detected (TI2 + TD1) → No tool qualification necessary	▪ Low significance for product quality ▪ No tool qualification necessary → Confidence in tool behaviour from an ISO 26262 perspective not critical
TCL2 (medium)	▪ Tool has safety-related influence (TI2) ▪ Error **not reliably detectable (TD2)**	▪ Tool **important for product quality** ▪ Tool qualification required ▪ Requirements depend on ASIL level
TCL3 (high)	▪ Requirements depend on ASIL level (TI2) ▪ Errors **remain undetected (TD3)**	▪ High importance for product quality ▪ **Tool qualification absolutely necessary**, high requirements ▪ Differences between TCL2 and TCL3 are methodologically relevant but not dramatic (depending on ASIL)

Background: calculation of TCL

AQI Automotive Quality Institute

Need for trust in tool usage

Starting point: How high must the level of trust in the tool be (or is tool qualification necessary)?

Table 3 — Determination of the Tool Confidence Level (TCL)

		Tool error detection		
		TD1	TD2	TD3
Tool impact	TI1	TCL1	TCL1	TCL1
	TI2	TCL1	TCL2	TCL3

Source: ISO 26262:2018, Section. 11.4.5.4

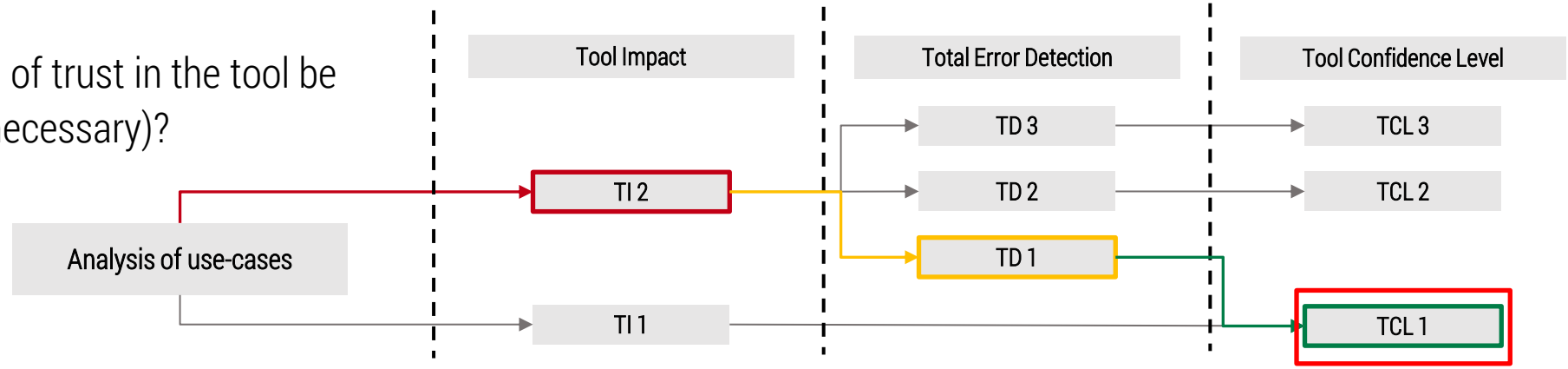
Aspect	Picture	Description
Target	***Assignment of the appropriate Tool Confidence Level (TCL)*** *based on Tool Impact (TI) and Tool Error Detection (TD)*	*Determine* ***how high the level of trust*** *in the tool must be (or whether tool qualification is necessary)*
TCL1 (low)	▪ Tool has no safety-related influence (TI1) or ▪ Errors are reliably detected (TI2 + TD1) → No tool qualification necessary	▪ Low significance for product quality ▪ No tool qualification necessary → Confidence in tool behaviour from an ISO 26262 perspective not critical
TCL2 (medium)	▪ Tool has safety-related influence (TI2) ▪ Error not reliably detectable (TD2)	▪ Tool important for product quality ▪ Tool qualification required ▪ Requirements depend on ASIL level
TCL3 (high)	▪ Requirements depend on ASIL level (TI2) ▪ Errors **remain undetected (TD3)**	▪ **High importance** for product quality ▪ **Tool qualification absolutely necessary**, high requirements ▪ Differences between TCL2 and TCL3 are methodologically relevant but not dramatic (depending on ASIL)

Background: calculation of TCL

AQI | Automotive Quality Institute

Determination of the Tool Confidence Level (TCL)

Action: How high must the level of trust in the tool be (or is tool qualification necessary)?



Tool Impact and Tool Error Detection Analysis

Phase: Training & Validation

Task: Deep Learning & ML Frameworks
Optimizing model parameters based on training data.

Artifact: Trained model

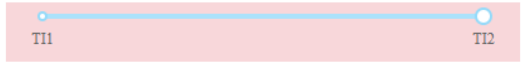
AI Regulation

Tool Confidence Level: 1

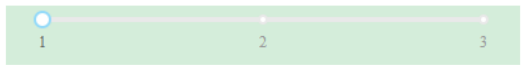
Calculation of confidence level
(1 = low, 2 = medium, 3 = high)

Failure mode: Lack of AI Transparency
RPN: 105
Potential tool failure: May fail to generate or link detailed training documentation and validation reports, resulting in trained models lacking required transparency and conformity evidence.

Tool Impact Value



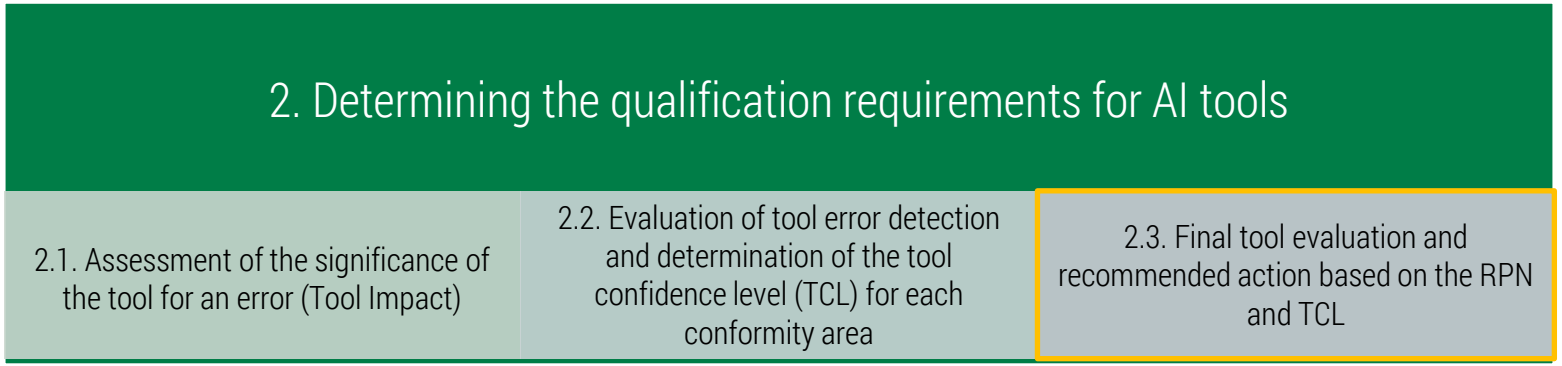
Tool Error Detection



AQI Automotive Quality Institute

2. Determining the qualification requirements for AI tools

2.3. Tool assessment and recommended actions based on RPN and TCL

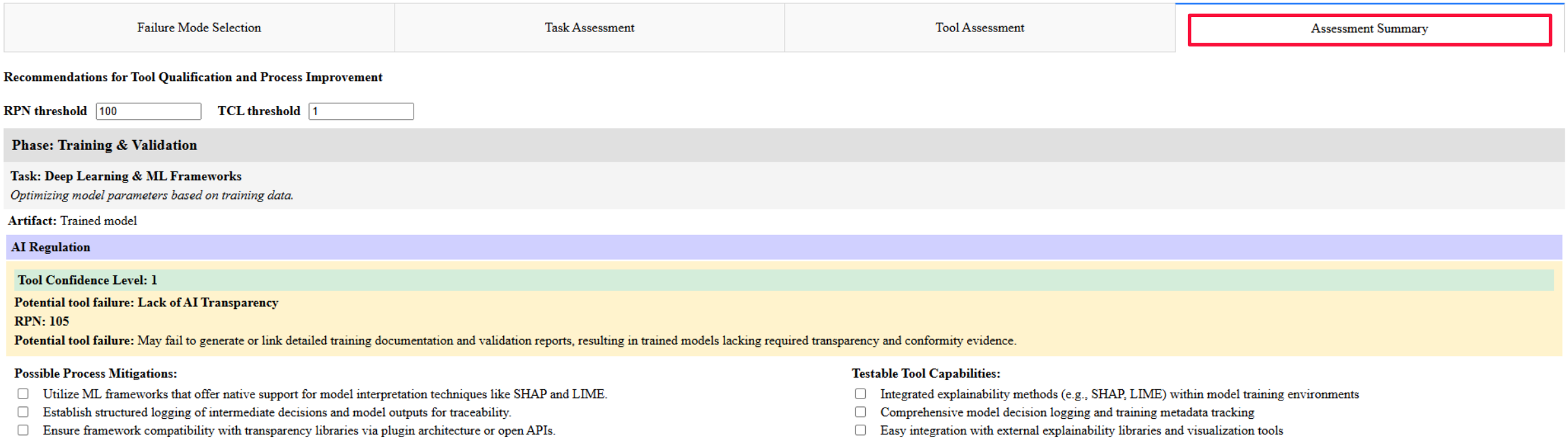


AssessML: evaluation and recommended action

AQI Automotive Quality Institute

Assessment Summery - Qualification summary

The Assessment Summery section is used to evaluate and prioritise the qualification summary based on the RPN and TCL for selected fault conditions.

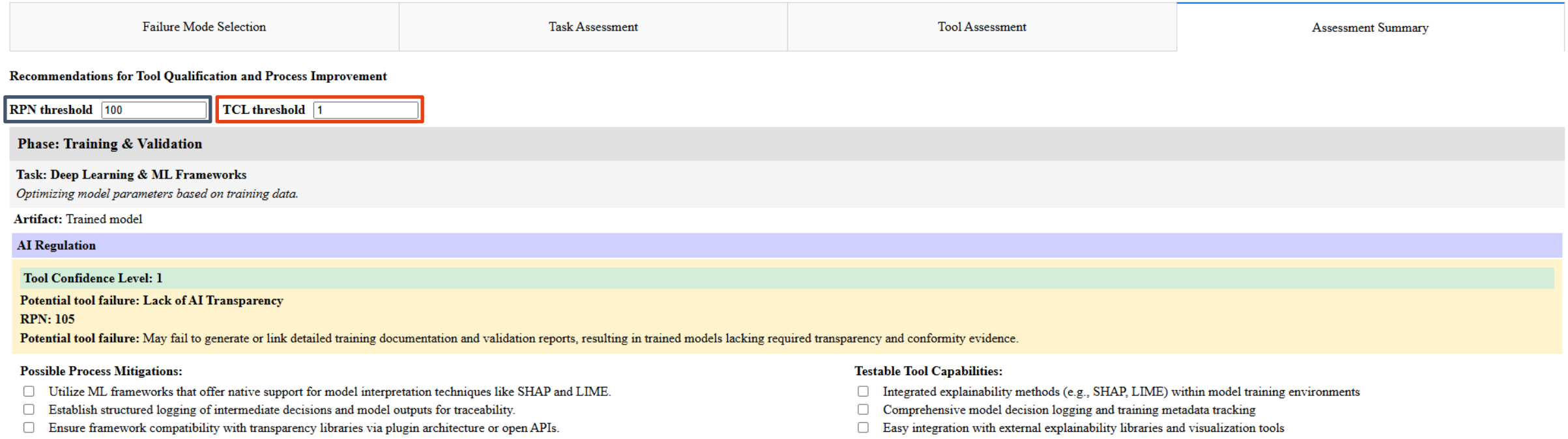


AssessML: evaluation and recommended action

Overview of the specific risk priority score and tool confidence level

The error states can be filtered using the inputs for **thresholds**. Two different specifications can be weighted:

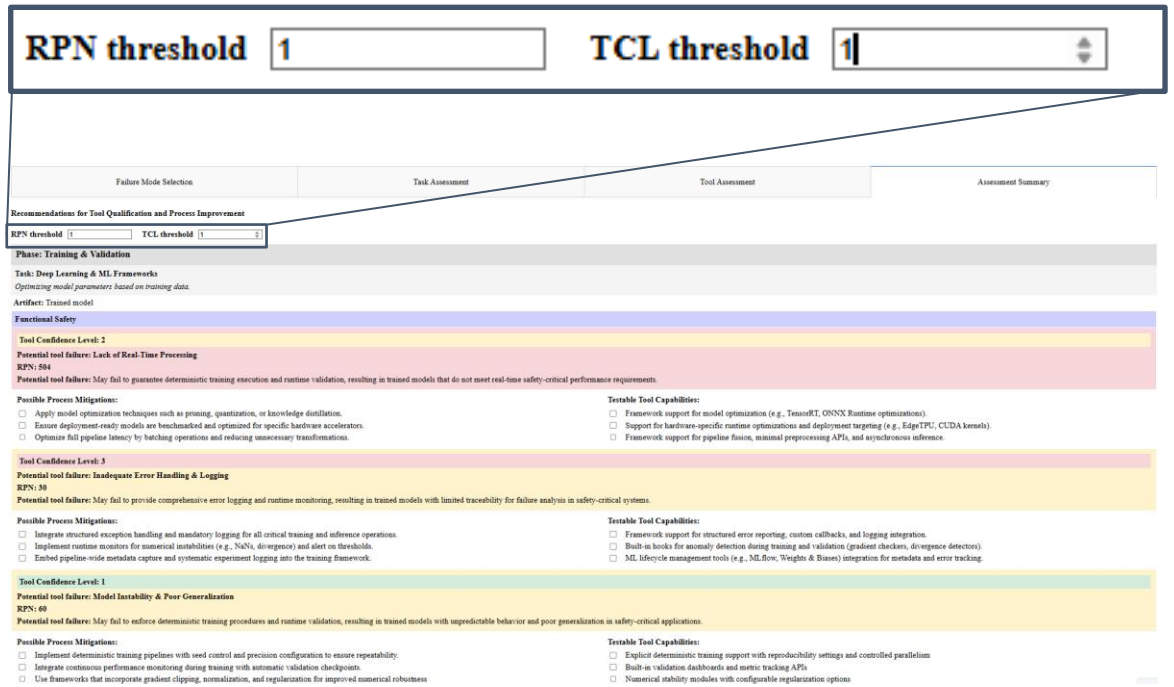
- RPN threshold (limit value for the risk priority number)
- TCL threshold (limit value for the tool confidence level)



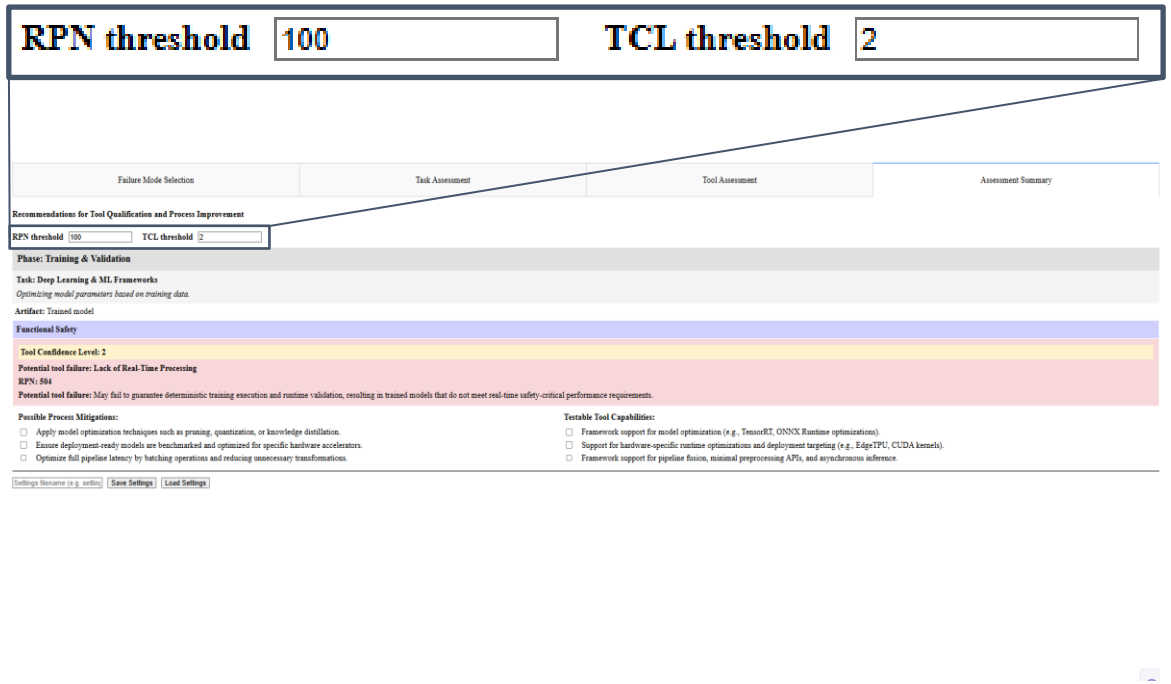
AssessML: evaluation and recommended action

What is the purpose of thresholds?

All statuses are automatically transferred to the assessment summary, which can quickly make the board confusing.



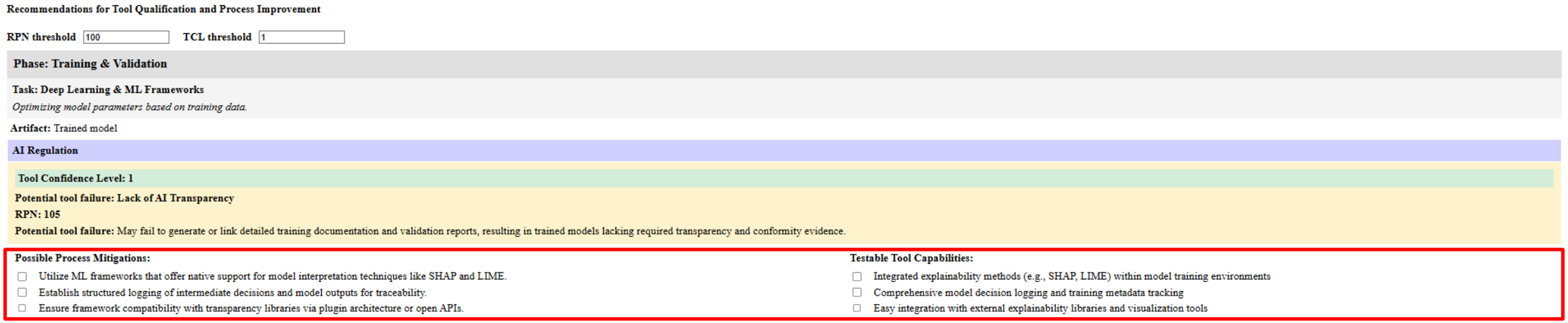
Threshold values can be used to filter out **critical error states** or **unsafe tools** by setting them appropriately. Setting the filter makes the analysis clearer.



AssessML: evaluation and recommended action

Overview of the specific risk priority score and tool confidence level

The select boxes (Checkboxes) listed below are used to select specific tool capabilities or possible process measures:



Recommendations for Tool Qualification and Process Improvement

RPN threshold 100 **TCL threshold** 1

Phase: Training & Validation

Task: Deep Learning & ML Frameworks

Optimizing model parameters based on training data.

Artifact: Trained model

AI Regulation

Tool Confidence Level: 1

Potential tool failure: Lack of AI Transparency

RPN: 105

Potential tool failure: May fail to generate or link detailed training documentation and validation reports, resulting in trained models lacking required transparency and conformity evidence.

Possible Process Mitigations:

- ☐ Utilize ML frameworks that offer native support for model interpretation techniques like SHAP and LIME.
- ☐ Establish structured logging of intermediate decisions and model outputs for traceability.
- ☐ Ensure framework compatibility with transparency libraries via plugin architecture or open APIs.

Testable Tool Capabilities:

- ☐ Integrated explainability methods (e.g., SHAP, LIME) within model training environments
- ☐ Comprehensive model decision logging and training metadata tracking
- ☐ Easy integration with external explainability libraries and visualization tools

Possible Process Mitigations:

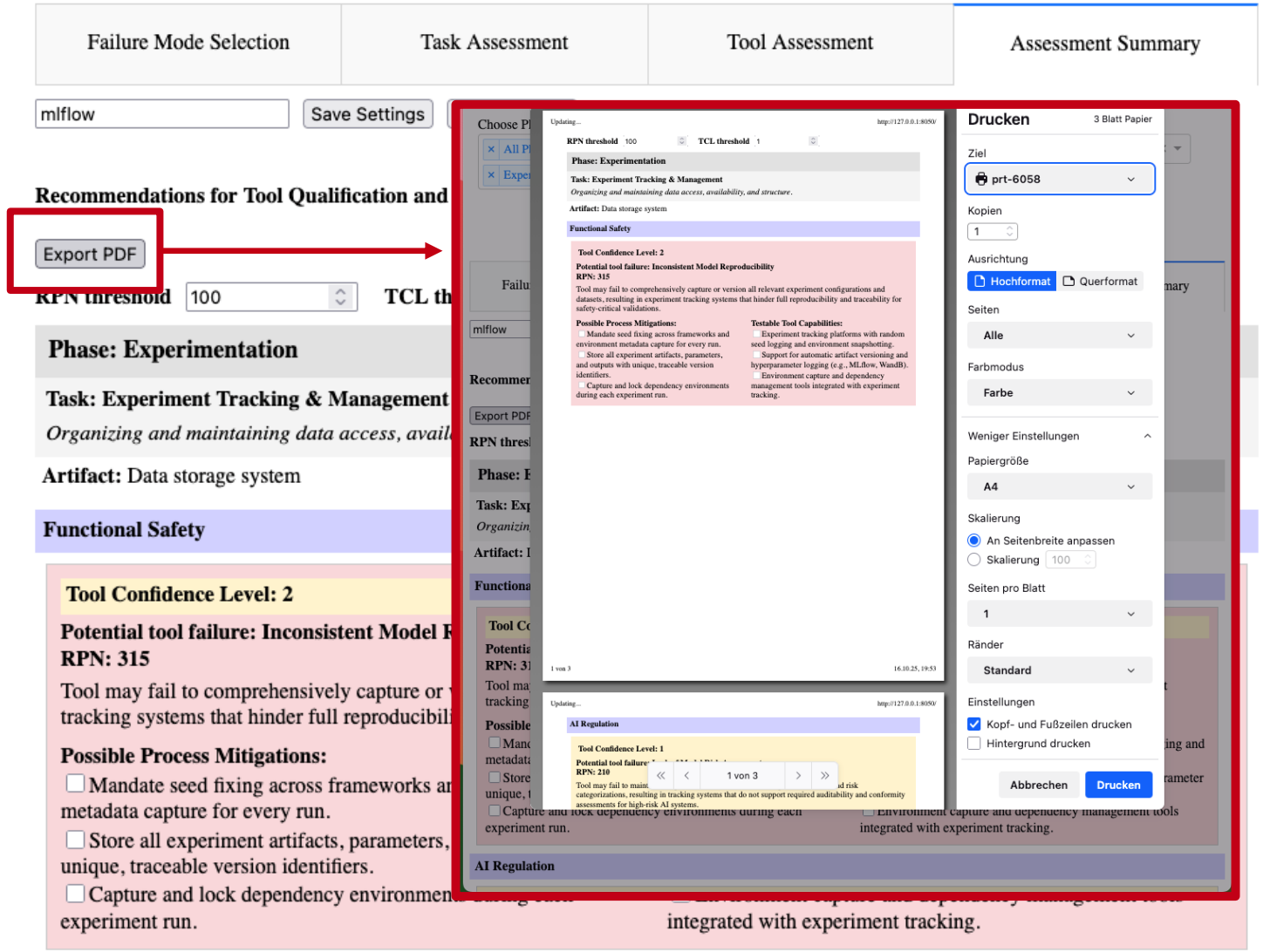
Here, methods or procedures can be selected that are intended to reduce the risk of a specific error condition.

Testable Tool Capabilities:

Here, specific features or functions of tools that are to be tested or reviewed in the assessment can be marked.

AssessML: export and print

The results in the Assessment Summary can be exported as a PDF and printed by pressing the ‘Export PDF’ button.



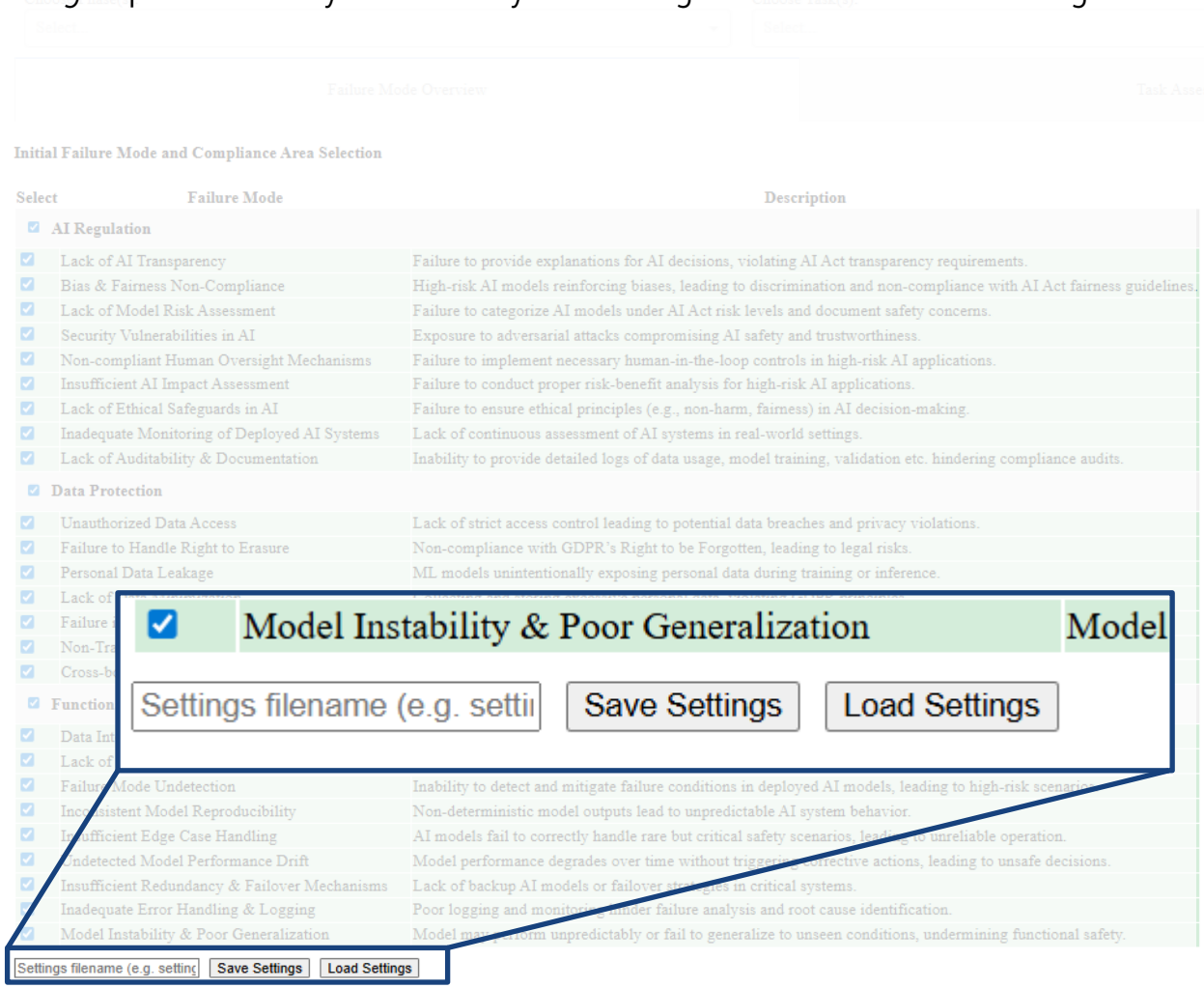
AQI Automotive Quality Institute

Additional features of the AssessML

AssessML: saving and loading

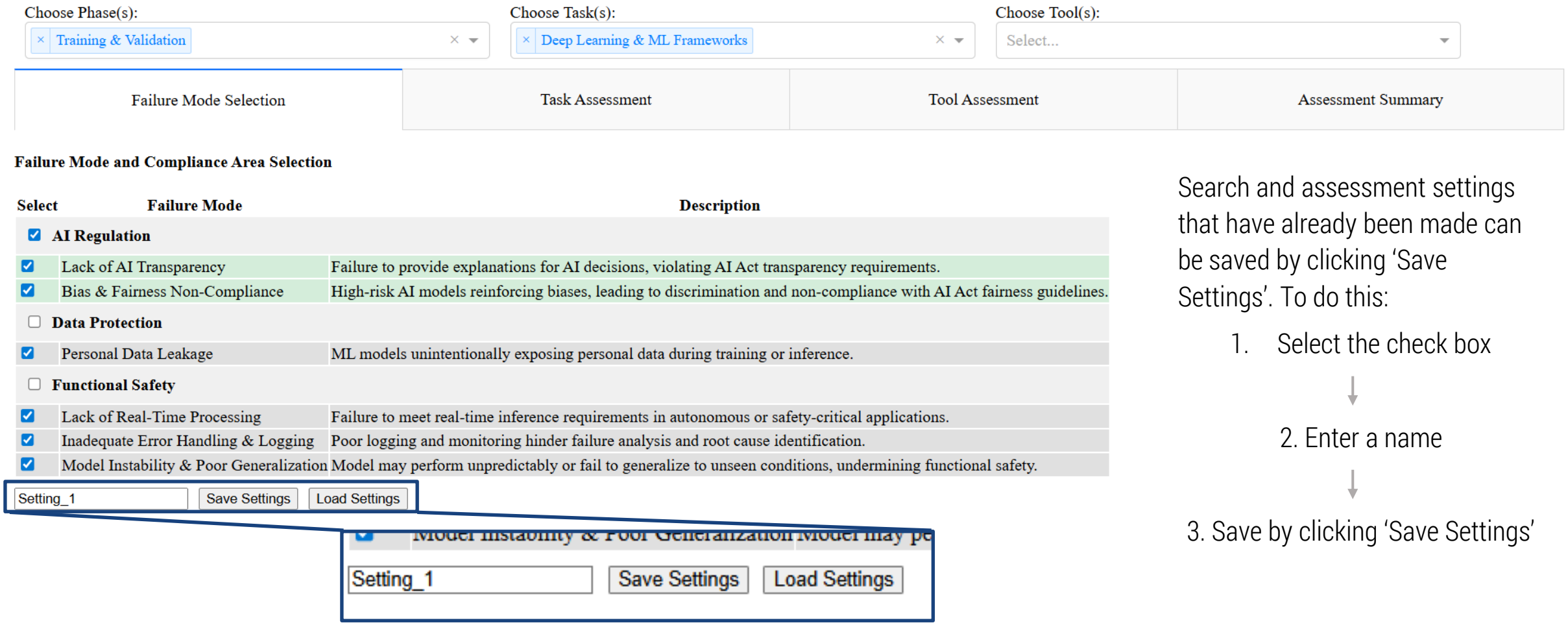
Why should one save and load?

The *Save* and *Load Settings* option allows you to save your settings and assessment settings and reload them as needed.



AssessML: saving and loading

Explanation of the storage process



Search and assessment settings that have already been made can be saved by clicking 'Save Settings'. To do this:

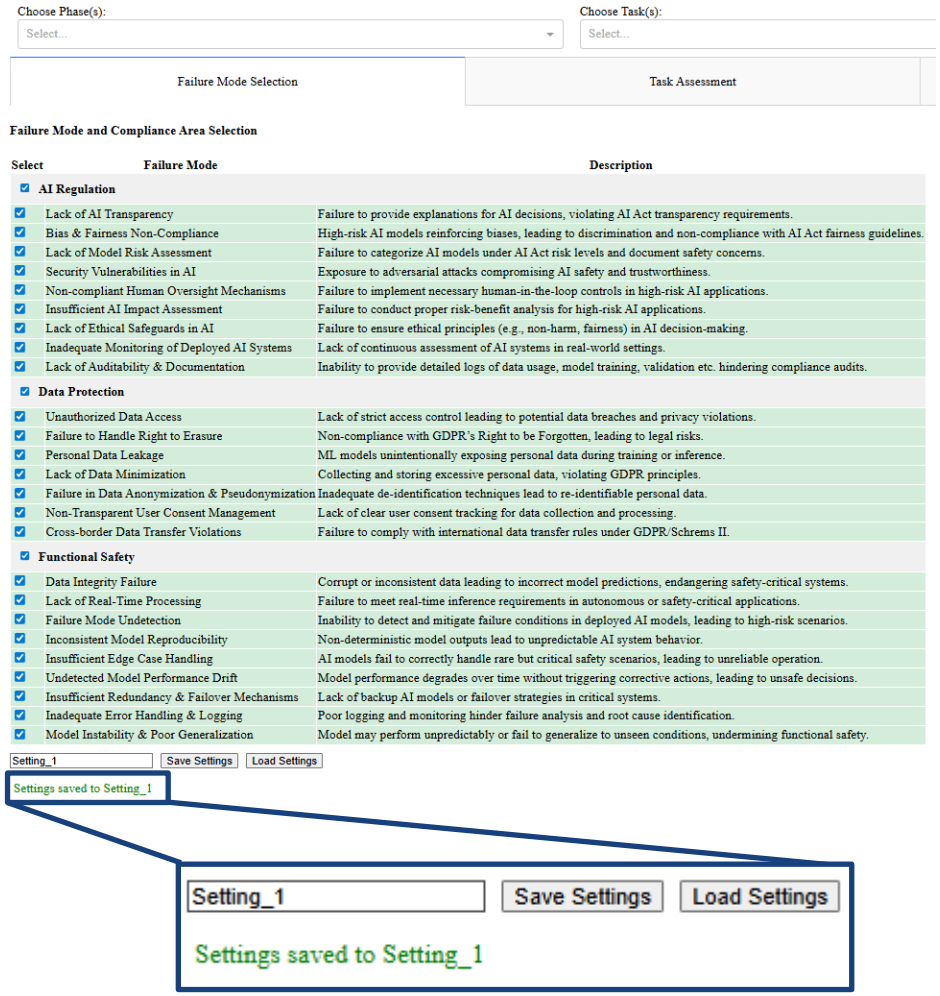
1. Select the check box
2. Enter a name
3. Save by clicking 'Save Settings'

AssessML: saving and loading

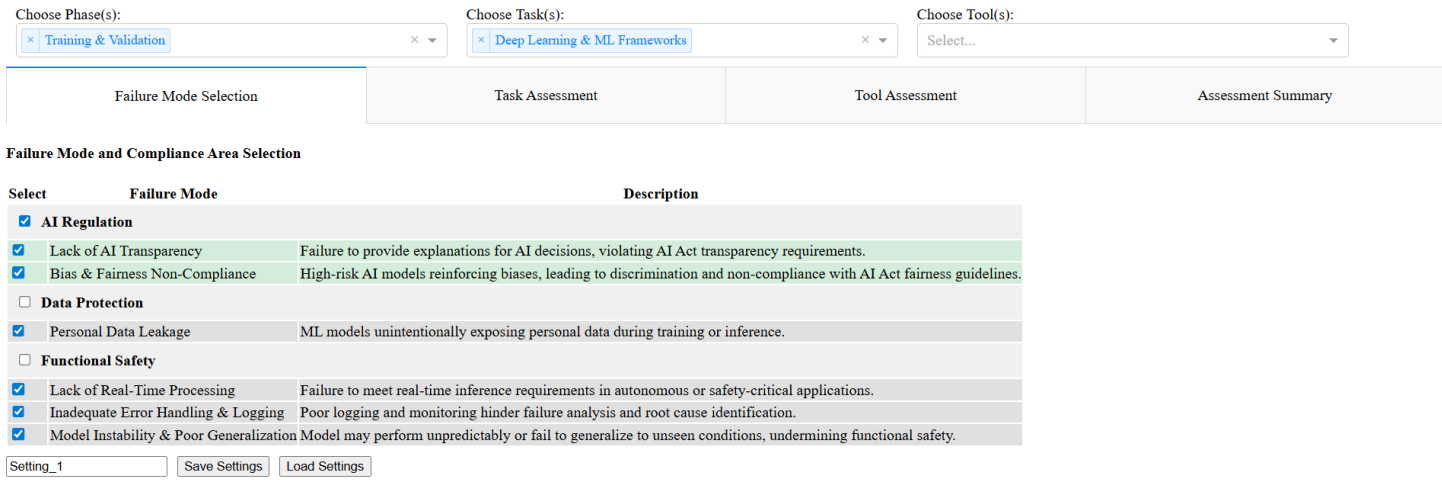
Explanation of the loading process

Automotive Quality Institute

Saved settings are listed below. To retrieve saved settings:



1. Enter setting names
2. Click Load Settings
3. Dashboard loads configuration



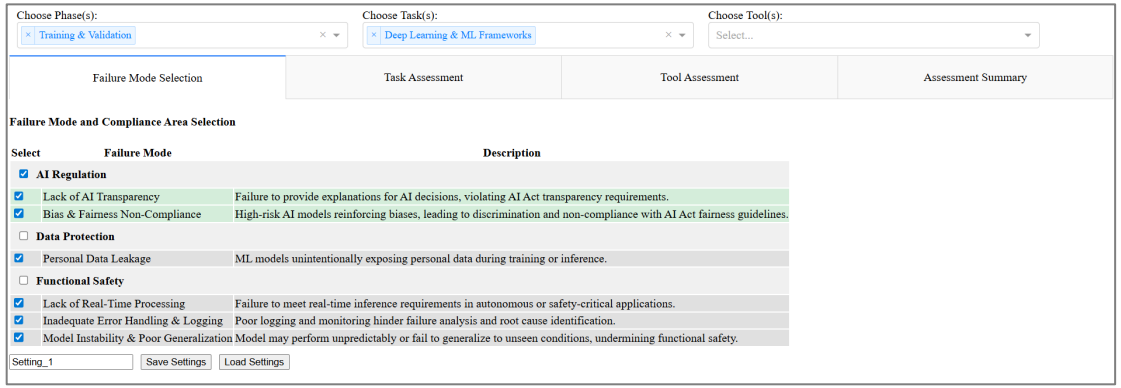
AssessML: saving and loading

Automotive Quality Institute

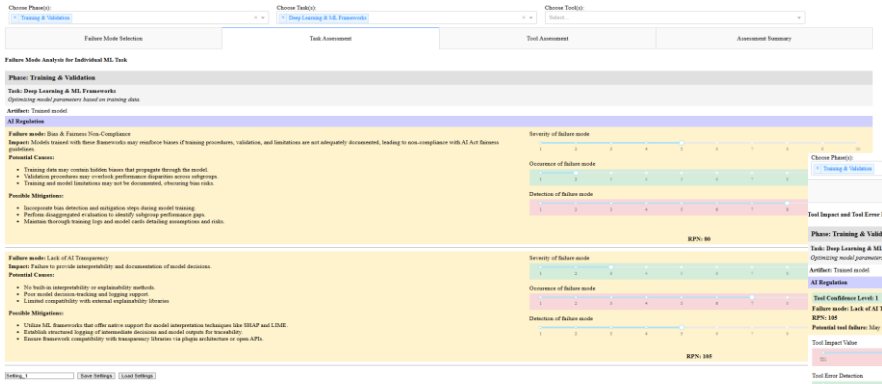
Notes

Saved settings include both the selected error states and assessments already performed, as well as the summary.

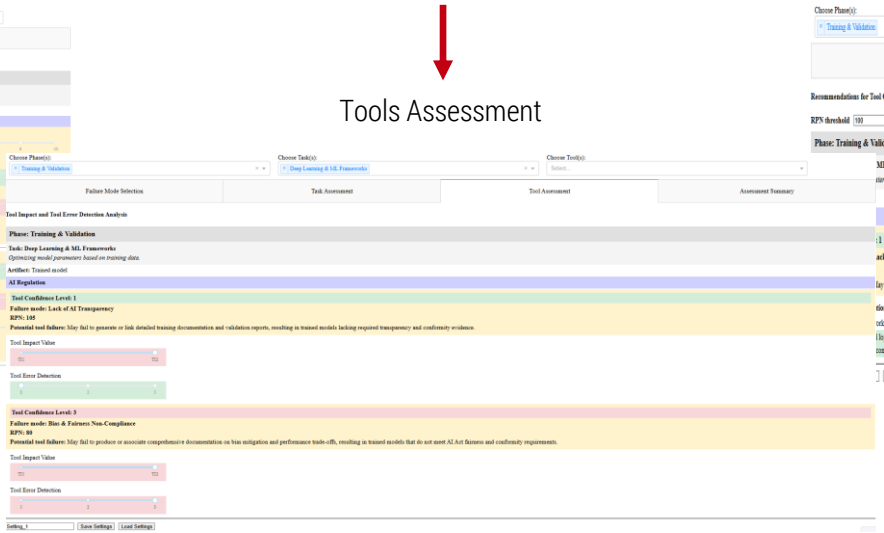
Failure Mode Overview



Task Assessment



Tools Assessment



Assessment Summery