TÜV AUSTRIA Best Practice Guideline:

Technical Documentation

Version number: 1.3

Release date: 11.03.2025

This template document provides a structure for the technical documentation of the developed AI system in accordance with EU AI Act, Art. 11, ISO/IEC 42001:2023 and the TÜV AUSTRIA TRUSTED AI standard. It ensures that all relevant technical details are captured, enabling users and stakeholders to understand the AI system design, functionality, and compliance with regulatory standards.

The applicable sections of the template depend on the applicable regulations, company size, risk-level, and type of the developed AI system.

RTR

Note: This document template was jointly coordinated with the AI Service Desk established at RTR-GmbH and is, in accordance with § 194a para 1 no 2 TKG, also available on its website at https://ai.rtr.at
This document template is intended as a guideline for fulfilling the requirements on technical documentation of AI systems imposed in the AI Act. It is not a precedent for expected harmonised standards, nor of guidelines of the AI Office.

For reasons of clarity, the TÜV AUSTRIA Best Practice Guideline: Technical Documentation *primarily uses the term "AI system", but also refers in some places to "AI models", "GPAI models" or "models". The terms mentioned and their derivatives do not anticipate any legal qualification of the respective technical component as an "AI system" within the meaning of Art. 3 No. 1 EU AI Act, "GPAI model" within the meaning of Art. 3 No. 66 EU AI Act, as an "AI model" (term not explicitly defined by the EU AI Act), etc., and the associated documentation requirements. For example, the guidelines refer to the "responsibilities (including contact details) of the person(s) or organization(s) developing the AI system". Corresponding information must also be provided in the case of a legal classification as a "GPAI model".* The guidelines published by the European Commission[1] may also be consulted in interpreting the term 'AI system' within the meaning of Art. 3 no. 1 EU AI Act.

[1] https://digital-strategy.ec.europa.eu/en/library/commission-publishes-guidelines-ai-system-definition-facilitate-first-ai-acts-rules-application

1. Key Facts

Description of:

- Responsibilities (including contact details of)
  - Person or organization developing the AI system, if applicable
  - Person or organization responsible for operating the AI system and those accountable for system use, especially for handling system failures or managing updates if applicable
- Release date of the AI system
- Version number reflecting its relation to previous versions
- Model type(s) and model architecture(s) underlying the AI system
- (High-level) information about training algorithms, parameters, fairness constraints, or other applied approaches, used foundation models and features
- Paper or other resources for more information
- Details on how the AI system should be cited
- License information for the AI system
- Contact for questions and comments about the AI system

2. Intended Purpose

Description of:

- Primary intended uses and (simplified version of) the application domain definition for the AI system
- Primary intended users
- Foreseeable misuse
- Out-of-scope use cases
- Risks when applying the AI system and mitigating measure including the sources of risks to
  - health and safety
  - fundamental rights, and
  - discrimination,

  if applicable,

and

- Instructions for use for the deployer and a basic description of the user interface, if applicable

3. Results of Risk Analysis

Description of:

- The known and the reasonably foreseeable risks that the high-risk AI system can pose to health, safety, to fundamental rights, or discrimination, when the high-risk AI system is used in accordance with its intended purpose
- The estimation and evaluation of the risks that may emerge when the high-risk AI system is used in accordance with its intended purpose, and under conditions of reasonably foreseeable misuse
- The evaluation of other risks possibly arising, based on the analysis of data gathered from the post-market monitoring system referred to in EU AI Act, Art. 72
- The adoption of appropriate and targeted risk management measures designed to address the risks identified
- Risk management measures, that ensure:
    - Elimination or reduction of risks identified and evaluated pursuant to the above paragraph, as far as technically feasible, through adequate design and development of the high-risk AI system
    - Where appropriate, implementation of adequate mitigation and control measures addressing risks that cannot be eliminated
    - Provision of information required pursuant to EU AI Act, Art. 13 and, where appropriate, training of deployers

4. Results of Impact Assessment

Description of the potential consequences for individuals or groups of individuals, or both, and societies that can result from the AI system throughout its life cycle, e.g., through an impact assessment according to ISO/IEC 42001:2023 Annex B.5.

Following EU AI Act, Art. 27, specific deployers need to carry out and document an impact assessment that addresses the following points:

- a description of the deployer's processes in which the high-risk AI system will be used in line with its intended purpose
- a description of the period of time within which, and the frequency with which, each high-risk AI system is intended to be used
- the categories of natural persons and groups likely to be affected by its use in the specific context;
- the specific risks of harm likely to have an impact on the categories of natural persons or groups of persons identified pursuant to the previous paragraph
- a description of the implementation of human oversight measures, according to the instructions for use;
- the measures to be taken in the case of the materialisation of the risks mentioned above, including the arrangements for internal governance and complaint mechanisms.

5. Compliance

Description of:

- The risk management system in accordance with EU AI Act, Art. 9
- The harmonized standards which apply in full or in part, the references of which have been published in the Official Journal of the European Union. Where no such harmonized standards have been applied, a detailed description of the solutions adopted to meet the requirements set out in EU AI Act, Chapter III, Section 2, including a list of other relevant standards and technical specifications applied
    - Compliance with the requirements (EU AI Act, Art. 8)
    - Risk management system (EU AI Act, Art. 9)
    - Data and data governance (EU AI Act, Art. 10)
    - Technical documentation (EU AI Act, Art. 11)
    - Record-keeping (EU AI Act, Art. 12)
    - Transparency and provision of information to deployers (EU AI Act, Art. 13)
    - Human oversight (EU AI Act, Art. 14)
    - Accuracy, robustness, and cybersecurity (EU AI Act, Art. 15)
- A list of other relevant standards and technical specifications applied

and

- A copy of the EU declaration of conformity referred to in EU AI Act, Art. 47

6. Development Process

Description of:

- The usage of pre-trained systems or tools used for the development of the AI system and how those were integrated or modified
- The general logic of the AI system
    - Including main classification choices, what the system is designed to optimize for, and the relevance of different parameters to achieving these goals
- The training data sets used, specifically:
    - A general description of the data set (data sources and data collection process, dataset composition)
    - The provenance of the data and information about the selection and labelling procedure
    - Cleaning methodologies
- Steps involved in data preprocessing (EU AI Act, Art. 10)
- The training methodologies and techniques, including
    - Information about fine-tuning of hyperparameters and the associated cost function if applicable
    - Decisions about possible trade-offs made regarding the technical solutions adopted to balance different performance metrics (e.g. in terms of fairness requirements for the application)

- The computational resources needed for training, validation and testing of the AI system

7. Technical Details

Description of:

- The interaction of the system with Hardware and Software:
    - Information about how the AI system interacts with other hardware/software, including other AI systems, if applicable
    - The interaction between the data-driven programmed parts and non-data-driven programmed parts
- The system architecture, including:
    - High-Level Architecture: A (UML) diagram and description of the overall system architecture
    - System Integration: How the AI system integrates with other systems or components (REST, etc.)
        - Key design choices and assumptions, including the rationale regarding the intended users or groups the system is designed to serve
- The minimum and recommended hardware specifications, including:
    - CPU: The minimum and recommended processor specifications needed to run the AI system. This could include the number of cores, clock speed, and specific processor models, if necessary
    - GPU: For systems that leverage GPUs for acceleration, specify the minimum and recommended GPU specifications, such as the type (e.g., NVIDIA), model, VRAM (Video RAM) size, and CUDA cores
    - RAM: The minimum and recommended amount of system memory required to operate the AI system efficiently
    - Storage: Minimum and recommended storage requirements, including the type of storage (HDD, SSD) and space required for data storage, model files, and logs
- The software requirements
    - Operating System: Supported operating systems (e.g., Windows, Linux, macOS) and specific versions
    - Dependencies: List of all software dependencies required to run the AI system, including:
        - Programming Languages: Specify the programming languages used
        - Libraries and Frameworks: Detailed list of libraries and frameworks, including version numbers
        - Databases: Details on the specific databases required (engines, formats, etc.)
        - Other Tools: Additional tools or software required
    - Installation Instructions: Step-by-step instructions for installing each dependency, including commands and configurations required, necessary authorizations and integration tests
- The Input/Output, including but not limited to:
    - Data Sources
        - Include information about sensory or other access points via which the system accesses inputs
    - Requirements for Input Data (input format and potential restrictions for the input data)

- Data Storage (how and where data is stored, including database schemas or file structures)
- Format of System Output (data types, etc.)
- Interpretation of the output (probabilities, logits, etc.)

- The Software/Firmware versions:
  - Information of the relevant software/firmware versions and update requirements
- The forms of availability:
  - Description of all the forms in which the AI system is available (e.g., packages, downloads, APIs)
- The product component details:
  - If the AI system is a product component, provide photographs or illustrations showing external features, markings, and internal layout
- The human oversight measures:
  - A description of the human-machine interface tools provided to ensure the AI system can be effectively overseen
  - A description how human operators are enabled to:
    - adequately understand the relevant capabilities and limitations of the high-risk AI system and properly monitor its operation
    - interpret the results of the AI system (detect anomalies, malfunctions and unexpected performance)
    - recognize and be aware of so-called "automation bias" (overconfidence in the output produced by the AI system)
    - disregard, override or reverse the output of a high-risk AI system in certain situations, or
    - to intervene if necessary (e.g. to interrupt procedures).

8. Test and Validation Data Set and Methodologies

Description of:

- The composition of the test data set used to assess the model performance, including a description of the preprocessing
- The validation and testing procedures used, including the characteristics of the data used in these processes, as well as the metrics used to measure accuracy, robustness, and compliance with requirements, specifically:
  - Validation and testing data: Description of validation and testing datasets, including their sources, composition, and characteristics
  - Testing methodology: Explanation of testing procedures, including any real-world testing and evaluation
  - Metrics and logs: Detailed description of performance metrics (e.g., accuracy, robustness, potential discriminatory impacts) and a record of test logs and reports. Ensure all test logs are dated and signed by responsible individuals

9. AI System Performance Metrics

Description of the model metrics for the specific AI system:

- Description of the appropriateness of each metric and the acceptable target range
- Evaluation of appropriate metrics that indicate potential biases and their implications
- Uncertainty of performance measures, i.e. confidence intervals or results of statistical test, if applicable
- System capabilities and limitations in performance metrics, with respect to robustness
- Interpretation of the performance metrics with respect to the reliability of the system
- Public Evaluation Protocols: If applicable, reference and describe any public protocols or tools utilized (e.g., GLUE, SuperGLUE, HELM)
- Alternative Evaluation Methodologies: If applicable, detail other methods used in evaluation, including performance in task-specific and general-use cases

10. Monitoring of the AI System

Description of:

- The plans for managing failures of the system
- The rollback plans for the AI system, procedures for:
    - Turning off features
    - Update processes
    - Plans for notifying customers and users of changes or system failures, and
    - Mitigation strategies
- The processes for monitoring the health of the AI system in the post-market phase. This includes:
    - Ensuring the AI system operates as intended and within normal operating margins (observability) and addressing AI system failures
    - Monitoring of relevant events, prioritization and review of event logs, investigation of failures, and failure prevention measures
- The cybersecurity measures put in place (EU AI Act, Annex IV, 2h), especially technical and organizational measures for securing the development and operating infrastructure, as well as means against adversarial attacks
- Relevant changes made by the provider to the system through its lifecycle

11. Lifecycle of the AI System

Description of:

- The pre-determined changes, including:
    - An overview of pre-determined changes (i.e., modifications, updates, or adjustments that are planned and incorporated into the design or operational lifecycle of the AI system) and which measures are in place to guarantee the continuous compliance of the system with relevant technical and regulatory standards, with respect to any anticipated updates or modifications

- The procedures the organization has in place to address operational changes, including communication with users and internal evaluations. The documentation should be:
    - Up-to-date
    - Accurate, and
    - Approved by relevant management

12. Additional Requirements for General-Purpose AI Models

This section follows EU AI Act, Art. 53 and Annexes XI to XIII.

- General Description of the General-Purpose AI Model
    - Intended Tasks and Integration: A comprehensive overview of the tasks the model is intended to perform and the types and nature of AI systems into which it can be integrated
    - Acceptable Use Policies, guidelines or restrictions set by the AI provider to specify how the AI model should and should not be used
    - Description of the methods used for its distribution
- Detailed Description and Resources used during the Development Process
    - A sufficiently detailed summary of the data used for training the general-purpose AI model (types and provenance of the data, pre-preprocessing methodology, number of available data points and their main features, data quality metrics). The summary should be made publicly available.
    - Technical Means for Integration: Information on the technical means required (e.g., instructions of use, infrastructure, tools) to integrate the general-purpose AI model into AI systems
    - Computational Resources Used:
        - Training Resources: Details such as the number of floating-point operations and training time. Mention other relevant information
    - Energy Consumption:
        - Known or Estimated Energy Use: The model's known or estimated energy consumption (e.g., Based on the computational resources used)
- A policy to comply with Union law on copyright and related rights according to EU AI Act, Art. 53, No. 1c)
- Criteria for Systemic Risk Assessment
    - Number of Parameters: Provide the parameter count for the model and explain its relevance to the model's capabilities
    - Supported Modalities: List all modalities the model supports (e.g., text-to-text, text-to-image)
    - Thresholds for High-Impact Capabilities: Describe the state-of-the-art benchmarks and thresholds relevant to high-impact capabilities in each modality
    - Benchmark Results: Detail any benchmark tests performed and the model's adaptability to new tasks
    - Reach and User Base: Detail the extent of market reach, including statistics on registered business users within the EU (10,000+ threshold as a criterion)
    - Number of Registered End-Users: Document the number of end-users interacting with or benefiting from the model

In case that the GPAI poses systemic risk:

- Adversarial Testing and Model Adaptations
    - Description of Adversarial Testing Processes: Explain whether adversarial testing was conducted and detail the procedures, e.g., internal testing, red teaming exercises
    - Measures for Red Teaming: Outline the specific setup and scenarios tested, such as edge-case inputs or situations to identify vulnerabilities
    - Alignment and Fine-Tuning Measures: Describe techniques used to improve the model's alignment with ethical, legal, and performance standards, such as reinforcement learning with human feedback (RLHF)
    - External Collaboration in Testing: If external testers or community feedback were used, provide details of the contributions and integrations

Glossary: Terms and Definitions

Confidence Interval

Confidence interval is an interval which is expected to typically contain the parameter being estimated with a certain confidence level.

Commonly used in statistical testing.

Edge Case

Sample that is close to the decision boundary for classification problems.

Explainability

Property of an ML system that allows humans to understand the factors influencing the ML system's output.

Explainability can be achieved by applying a toolstack of methods such as LIME, SHAP etc.

Hyperparameters

Characteristics of a machine learning algorithm that affect its learning process. They are selected prior to training.

Examples of hyperparameters include the number of network layers, width of each layer, type of activation function, optimization method, learning rate for neural networks; the choice of kernel function in a support vector machine; number of leaves or depth of a tree; the K for K-means clustering; the maximum number of iterations of the expectation maximization algorithm; the number of Gaussians in a Gaussian mixture.

Robustness

The capability of an AI module to cope with erroneous, noisy, unknown, or adversarially constructed input data

The two types of robustness include adversarial robustness (AR) and corruption robustness (CR). AR means a quality of the model to cope with adversarial attacks, perturbations, i.e., maleficent changes in the data by a third party. CR refers to the model being able to cope with unintentional corruption, i.e., difference between data used for training and in-deployment.

Test Set

Dataset used to assess the performance of a final ML model.

Training

Process to determine parameters of an ML model, based on an ML algorithm operating on the training set.

Training Set

Dataset used to train the ML model.

Validation Set

Dataset used before, during and after training to select model class and tune hyperparameters.