Design of a Quality Management System based on the EU Artificial Intelligence Act

Henryk Mustroph and Stefanie Rinderle-Ma

Technical University of Munich, TUM School of Computation, Information and Technology, Garching, Germany
{henryk.mustroph,stefanie.rinderle-ma}@tum.de

Abstract. The Artificial Intelligence Act of the European Union mandates that providers and deployers of high-risk AI systems establish a quality management system (QMS). Among other criteria, a QMS shall help to i) identify, analyze, evaluate, and mitigate risks, ii) ensure evidence of compliance with training, validation, and testing data, and iii) verify and document the AI system design and quality. Current research mainly addresses conceptual considerations and framework designs for AI risk assessment and auditing processes. However, it often overlooks practical tools that actively involve and support humans in checking and documenting high-risk or general-purpose AI systems. This paper addresses this gap by proposing requirements derived from legal regulations and a generic design and architecture of a QMS for AI systems verification and documentation. A first version of a prototype QMS is implemented, integrating LLMs as examples of AI systems and focusing on an integrated risk management sub-service. The prototype is evaluated on i) a user story-based qualitative requirements assessment using potential stakeholder scenarios and ii) a technical assessment of the required GPU storage and performance.

Keywords: EU Artificial Intelligence Act · Quality Management System · AI Systems · Software as a Service · Compliance Management

1 Introduction

Over recent years, the rise of Artificial Intelligence (AI) has introduced rapid advancements and heightened risks, particularly in critical sectors like medicine, finance, and law, where AI increasingly participates in or even controls decision-making processes. To have risks under control, the European Commission introduced its initial draft of the Artificial Intelligence Act (EU AIA) in 2021, which was updated and adopted by the European Parliament on 13 March 2024, accepted by the EU member states on 21 May 2024 and published on 12 July 2024 in the Official Journal of the European Union and will therefore come into legal force on 1 August 2024. It is based on a risk-based approach that categorizes AI systems into distinct risk classes. Specific evidence of safety and security measures must be provided for high-risk AI systems. However, a notable exception

applies to general-purpose AI (GPAI) models, as stated by the Future of Life Institute (FLI) [10]. The most prominent examples of GPAIs are large language models (LLMs), which have gained immense popularity since the emergence of OpenAI's ChatGPT in November 2022. These systems are trained on diverse datasets and applicable across various domains and tasks, making it challenging to neatly classify them into specific risk categories. Consequently, the FLI proposed that GPAIs should provide the same verification and documentation as high-risk AI systems. Even for GPAI models without systematic risk and with less stringent mandatory compliance regulations, implementing the same verification and documentation process as for high-risk AI systems would be beneficial.

Theoretical considerations related to planning and designing risk frameworks by integrating risk standards, human expertise, and audit systems to evaluate AI systems throughout their development lifecycle and during market use have been explored since 2021. The goal is to ensure unified processes and designs as detailed by works such as [7,24,27,28,33]. Secondly, attention has been directed towards evaluating various characteristics of AI systems, such as performance, explainability, and robustness, and developing techniques to comply with the technical documentation regulations in the EU AIA, as seen in works [1,3,11,26,31]. However, most approaches focused on metrics to evaluate classification or regression models, often neglecting GPAI models such as LLMs. Furthermore, there is less research on designing and implementing a quality management system (QMS), which is mandatory for high-risk AI systems, as stated in Chapter 3, Section 3, Article 17 EU AIA, and is beneficial for GPAI models. Such a QMS should orchestrate AI development and use in compliance with EU AIA regulations to check and document the AI system's design, risk, and quality. However, no clear suggestion exists on how such a tool should look. Therefore, the idea is to provide a QMS design as a software as a service (SaaS) web application that assists and involves different stakeholders in the compliance management, verification, and documentation process for AI systems. This QMS should include i) mandatory compliance requirements derived from legal regulations of the EU AIA, ii) technical evaluation metrics to analyze AI systems, and iii) several sub-services, such as a risk management system (Article 9 EU AIA) and a data management and governance system (Article 10 EU AIA). Questions remain about translating legal regulations into software requirements and building a conceptual software architecture model for a QMS. The QMS SaaS should be generic enough to apply to any type of AI system. For that, a generic architecture must be chosen that can easily be extended for different exceptional needs depending on the type of AI system.

This paper aims to fill these gaps by i) constructing legal and structural requirements derived from the EU AIA [9] and developing a conceptual architecture and data model in UML for a QMS based on these requirements. ii) Implementing a first-version prototype QMS that focuses on the integration and adaptation of LLMs as an example of an AI system type. Third, iii) proposing several technical evaluation metrics tailored to LLMs to evaluate performance,

explainability, and consistency based on specialized tasks. The paper presents details of the EU AIA in Sect. 2. The method in Sect. 3 outlines the requirements, the QMS architecture, and details of the implemented prototype. The prototype is evaluated and discussed in Sect. 4. The conclusion is provided in Sect. 5.

2   EU Artificial Intelligence Act

The EU Artificial Intelligence Act (EU AIA) [9] is a comprehensive legislative framework regulating the development, deployment, and use of AI systems within the European Union. According to the EU AIA, an AI system *"means a machine-based system that is designed to operate with varying levels of autonomy and that may exhibit adaptiveness after deployment, and that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments;"* [1]. The EU AIA categorizes AI systems into four different risk classes. Depending on which risk category the AI system is in, specific requirements must be met to be authorized in the EU. The following AI system risk classes exist in the EU AIA:

– **Unacceptable Risk**: AI practices pose unacceptable safety, rights, and societal well-being risks. One example would be a social credit system, which evaluates human behavior and actions. These AI systems will be prohibited in the EU. [2]
– **High Risk**: These AI systems are allowed in the EU, but can cause risks to critical areas such as health, safety, and fundamental rights and therefore need to comply with a couple of regulations. [3]
– **Limited or Low Risk:** AI systems with limited risks have limited transparency regulations. An example of such an AI system is a chatbot answering customer requests employed in a non-high-risk domain. Low-risk AI systems such as spam filters, have no transparency regulations.

High-risk AI systems must adhere to the most challenging regulations and furnish the EU with documented evidence detailing risk identification, analysis, assessment, mitigation, data management, and data governance. They must be evaluated through their whole lifecycle before entering and post-market. Furthermore, they must demonstrate through technical evaluation that despite being classified as high-risk applications, measures are in place to limit risks effectively. Additionally, GPAI models present a unique challenge in classification under the risk categories. They are versatile and can be employed across various tasks, domains, and purposes, as mentioned by [27]. For that, the FLI came up in 2022 with an article that presents a short list of recommendations that suggested

---

[1] Chapter 1 Article 3 para. 1 EU AIA
[2] Chapter 2 Article 5 EU AIA
[3] Chapter 3 Section 1 Article 6 EU AIA

treating GPAI systems the same as high-risk AI systems [10]. They suggested establishing a QMS for GPAIs, the same as for high-risk AI systems, and checking as many regulations as possible from Chapter 3, Section 2. In [28], the authors also plead for categorizing GPAI systems as high-risk AI systems. The EU AIA partly adopted the recommendations from the FLI. It defined a GPAI model as a model *"... trained with a large amount of data using self-supervision at scale, that displays significant generality and is capable of competently performing a wide range of distinct tasks regardless of the way the model is placed on the market..."* [4]. Based on that definition, GPAI models can also be used in high-risk domains, for high-risk tasks, or integrated into high-risk systems. However, for GPAI models alone, no QMS system as for high-risk AI systems is mandatory; only transparency of used training, validation, and testing data, potential drawbacks, and risk should be documented, which is less strict but could still be changed in future adoptions of the EU AIA. Additionally, the EU AIA defined GPAI models with systematic risk, which is, *"... a risk that is specific to the high-impact capabilities of general-purpose AI models, having a significant impact on the Union market due to their reach, or due to actual or reasonably foreseeable negative effects on public health, safety, public security, fundamental rights, or the society as a whole..."* [5]. GPAI models with systematic risks are those that use 10^{25} FLOPS for training, such as GPT-4, which used $\sim 2*10^{25}$ FLOPS with 1.76 trillion parameters [6]. These models must perform more evaluations and risk assessments and adhere to transparency and documentation obligations. Because most open-source GPAI models are much smaller and trained with less data, they will probably not be treated as having systematic risk.

High-Risk AI Systems Regulations. A Summary of the key regulations for high-risk AI systems [7] is given. For detailed provisions, refer to the complete articles in the EU AIA [9].

- **Article 9 Risk management system** The article states that an RMS should be "established, applied, documented and maintained" (para. 1). The article describes the RMS as an "iterative process" that shall be documented and updated throughout the entire life cycle of the high-risk AI system (para. 2). In addition, the process is specified in the letters. Foreseeable risks shall be identified, analyzed, evaluated, and mitigated (lit. a). Risks that could result in a possible misuse of the AI system shall also be documented and mitigated (lit. b). Risks after placing the system on the market shall be identifiable (lit. c). Targeted risk management measures shall be applied (lit. d). Risk management measures shall be taken in such a way that as few interactions as possible occur (para. 4) but also that residual risks are still considered acceptable (para. 5). To identify risks, the underlying tests shall be carried out under real-world conditions before, during and after the development phase and post-market (para. 6, 7, 8).

---

[4] Chapter 1 Article 3 para. 63 EU AIA
[5] Chapter 1 Article 3 para. 65 EU AIA
[6] Chapter 5 Section 3 Article 55 EU AIA
[7] Chapter 3 Section 2 EU AIA

- **Article 10 Data and data governance** For data used to train high-risk AI systems, it shall be ensured that the training, validation, and testing dataset splits meet the specified quality criteria in para. 2 ff. (para. 1). These datasets shall be governed and managed according to practices suited to the AI system's intended purpose (para. 2). Key aspects include design choices, data collection and preparation, data suitability, bias examination, and mitigation, and data gaps (para. 2 lit. a - h). Training, validation, and testing datasets shall be relevant, representative, to the best extent, free of errors, and as complete as possible for their intended purpose (para. 3).
- **Article 11 Technical documentation** The technical documentation shall be prepared and submitted before the high-risk AI system is placed on the market and shall be continuously updated (para. 1). It shall contain evidence for at least all points listed in Annex IV EU AIA, including all regulations for high-risk AI systems (para. 2).
- **Article 12 Record-keeping** High-risk AI systems shall be designed to automatically record event logs throughout the AI system's entire lifetime (para. 1).
- **Article 13 Transparency and provision of information to deployers** High-risk AI systems shall be designed to provide sufficient transparency, allowing deployers to interpret and use the AI system's output correctly (para 1). Additionally, high-risk AI systems shall come with user instructions in a suitable digital format or another accessible form. These instructions shall be concise, complete, accurate, and clear, ensuring they are relevant and understandable to the deployers (para 2.).
- **Article 14 Human Oversight** High-risk AI systems shall be designed with appropriate human-machine interface tools to ensure they can be effectively monitored by humans throughout their use (para. 1). Human oversight shall aim to prevent or minimize risks to health, safety, or fundamental rights that may arise from the risk within the system's intended use or foreseeable misuse, primarily when other safeguards may not fully address these risks (para. 2). The human-machine interface shall also allow users to manually shut down the system in case of an emergency.
- **Article 15 Accuracy, robustness, and cybersecurity** High-risk AI systems shall be designed and developed to maintain appropriate levels of accuracy, robustness, and cybersecurity throughout their entire lifecycle (para. 1). To ensure these standards, the EU Commission, in collaboration with relevant stakeholders and organizations will promote the development of benchmarks and measurement methodologies for assessing accuracy, robustness, and other AI characteristic metrics (para. 2) [8].

Article 17 Quality Management System. The QMS aims to structure and plan a process to control at least all listed regulations for high-risk AI systems. The QMS aims to ensure compliance with all EU AIA regulations, verify the AI systems design, and assure its quality. The QMS must be appropriately documented and contain at least the points listed in para. 1 lit. a - m.

---

[8] No guidelines or further clarification have been provided yet by the EU Commission.

GPAI Models Regulations. A Summary of the key regulations for GPAI models [9] is provided. For detailed provisions, refer to the complete articles in the EU AIA [9]. It is important to note that GPAI models can be included in high-risk AI systems. The obligations for GPAI models and high-risk AI systems must be fulfilled in such cases.

- **Article 53 Obligations for providers of general-purpose AI models** Providers of GPAI models shall prepare and maintain up-to-date technical documentation of the model, including details of its training and testing processes and evaluation results (para 1.). This documentation, which shall include at least the information listed in Annex VI, should be available to the AI Office and national competent authorities upon request (lit. a). Prepare, update, and provide documentation to other AI system providers who intend to integrate the GPAI model into their systems. This documentation shall be made available while respecting intellectual property rights and trade secrets under Union and national laws (lit. b).
- **Article 55 Obligations for providers of general-purpose AI models with systemic risk** In addition to the obligations outlined in Articles 53 and 54, providers of GPAI models with systemic risk shall evaluate the model using standardized protocols and tools that reflect current best practices, including adversarial testing to identify and mitigate systemic risks (para. 1 lit. a). Assess and address potential systemic risks, including their sources, arising from these AI models' development, market placement, or use (lit. b). Track, document, and promptly report severe incidents and corrective actions to the AI Office (lit. c). Ensure cybersecurity protection for the GPAI model and the system in which it is integrated (lit. d).

Comments and Reviews. The EU AIA will be legally binding for companies offering AI systems on the European market. However, according to some German legal experts, many uncertainties remain. According to [4], the EU AI Regulation needs more specificity in several areas, making concrete implementation unclear for many requirements. In [29], the author highlights the high and numerous bureaucratic demands for high-risk AI systems. Due to this, compliance checks can require significant human and financial resources [4]. Without innovative approaches to drafting these regulations, only large tech companies will likely have the resources to develop high-risk and generative AI systems. This situation could force small and medium-sized enterprises to leave Europe or avoid developing high-risk AI systems, contradicting the EU AIA Regulation's objectives and stifling innovation. Additionally, determining if an AI system is classified as high-risk can be very challenging, as described by [8], for example, in the legal tech domain. Another concern is the EU AIA's lack of alignment and consistency with other legal acts. In [19], it points out that the EU AIA is not fully coordinated with, for example, the Medical Device Regulation, which governs medical devices, including medical software. Additionally, in [5] a German legal expert provides a more comprehensive summary of the EU AIA.

---

[9] Chapter 5 Section 2 & Section 3 EU AIA

3   Quality Management System for AI Systems

The proposed quality management system (QMS) is designed as a SaaS web application to ensure compliance with EU AIA regulations. It connects directly to an AI system, allowing stakeholders to perform technical quantitative and qualitative checks. The overall idea is to have a single system that verifies the design and quality of the AI system and automatically creates the necessary documentation. The QMS allows for performing technical tests and qualitative assessments from various domain experts. The generated documentation is then sent to national authorities to prove compliance. The first version of the prototype QMS includes two sub-services: a risk management system (RMS) and a data management and governance system (DMDGS). For this prototype implementation, the QMS adapts only LLMs. LLMs are chosen as AI models because of i) their immense popularity, ii) their frequent and domain-independent use, and iii) their open-source availability and easy integration from Hugging Face [10]. Even if LLMs are not necessarily high-risk AI systems, in any case, GPAI models, the idea and design of the QMS remain unchanged and can be used for all other types of AI systems. The legal regulations of the EU AIA mostly guide the design and functionality of each sub-service within the QMS. Particular emphasis is placed on the functionality of the RMS sub-service, based on insights from a conducted literature review for designing such a system. In addition to the RMS, the QMS incorporates a DMDGS sub-service. Although the development of the DMDGS was not the primary focus of this first version of prototype QMS, its inclusion should demonstrate that the QMS is designed to integrate multiple different sub-services. The goal is to integrate a separate sub-service for each article for high-risk AI systems. The chapter is structured as follows: It starts with the elicited high-level requirements on how to develop a QMS (cf. Sect. 3.1). Then the architecture and data model of the prototype QMS is described (cf. Sect. 3.2). Lastly, the implementation outcome and the user interface of the QMS and the two sub-services are presented (cf. Sect. 3.3).

3.1   Requirements

The QMS comprises functional (FR) and non-functional (NFR) requirements grouped into three distinct types. As defined in [6], FRs describe the system's functionalities, such as the feature of evaluating an AI system's performance. In contrast, NFRs describe how the system should perform the functionality, such as ensuring that the AI system's performance evaluation is done in under 2 seconds. Firstly, the QMS incorporates legal requirements derived from interpreting legal regulations set out in the EU AIA for high-risk AI and GPAI systems to ensure comprehensive safety and compliance. Secondly, system design requirements specify the technical aspects of the QMS. These requirements cover human involvement, architecture, and computational needs. All requirements are high-level and will be subdivided into more concrete sub-requirements

---

[10] Hugging Face: https://huggingface.co, accessed on 29 July 2024

in future iterations. However, they represent the minimum necessary functionality for such a QMS and illustrate the extensive range of features that need to be integrated.

Legal Requirements. Several requirements are derived from the given legal regulations. Articles 53 and 55 EU AIA provide information on requirements specifically for GPAI models, shown in Table 1.

Table 1: Legal Requirements (GPAI Systems)

Legal Requirements
LR01: Article 53 Obligations for GPAI (FR):   Draw up and maintain up-to-date technical documentation and provide transparency regarding the data used for the training, validation, and testing of the GPAI model
LR02: Article 55 Obligations for GPAI with systematic risk (FR):   Implement evaluation metrics, using standardized protocols, incorporating robustness and security checks to assess and mitigate systematic risk.

The GPAI-specific requirements align closely with the requirements constructed based on the regulations for high-risk AI systems. The idea is to integrate all requirements for high-risk AI systems into the QMS. According to the EU AIA, the QMS [11] should generally encompass strategies for regulatory compliance, including conformity assessment and management for modifications of legal guidelines or technical features. It should also provide techniques or procedures for the model's design, design control, verification, quality control, and quality assurance. Specifically, the QMS should contain an RMS to identify, analyze, assess, and mitigate potential risks and a DMDGS to demonstrate the data quality used for training, validation, and testing. The results should be stored in technical documentation. The software requirements derived from the obligation of high-risk AI systems are listed in Table 2 and must be ensured within the QMS.

Table 2: Legal Requirements (High-Risk AI Systems)

Legal Requirements
LR03: Article 9 Risk Management System (FR):   Incorporate into the QMS a module for risk identification, analysis, and assessment functionalities, ensuring comprehensive coverage throughout the entire lifecycle of the AI system.
LR04: Article 10 Data and Data Governance (FR):   Incorporate into the QMS a module to provide evidence that the data used for training, validation, and testing is unbiased, non-discriminatory, and compliant with privacy and data ownership regulations.
LR05: Article 12 Record keeping (FR):   Develop an integration within the QMS to link and log the usage of the AI system in use, recording details such as the timestamp of usage, user identification, purpose of use, and the specific task for which the AI system is employed.
LR06: Article 13 Transparency Provision (FR):   Incorporate into the QMS a transparency metric to assist users in interpreting the output of the AI system. This metric should provide clear and understandable insights into the decision-making process and underlying factors influencing the output.
LR07: Article 14 Human Oversight (FR):   The QMS shall incorporate a UI for deployers and end-users of the AI system to document in-use risks and misuses and to have the control to shut up the system in emergencies.
LR08: Article 15 Accuracy, Robustness and Cybersecurity (FR):   Implement into the QMS metrics to measure the system's accuracy, robustness, and cybersecurity.
LR9: Article 11 Technical Documentation (FR):   The QMS shall allow to create and maintain up-to-date technical documentation for the AI system before its market release and after the market release.
LR10: Article 61 Post-Market Monitoring (FR):   The QMS should be used to continuously assess the AI system's compliance with the requirements outlined in Chapter 2 after market release.

---

[11] Check for details and all obligations: Chapter 3, Section 3 Article 17 Quality Management System EU AIA

System Design Requirements. The system design requirements are categorized into three distinct types. First, The User Interface (UI) design and interaction specifications should encourage human involvement in checking and documenting AI systems. Second, architectural requirements should ensure a modular design, a smooth flow of data and communication between the services, and easy maintenance. The design should apply to multiple AI systems across various domains and tasks. Third, computational requirements should guarantee the efficient execution of technical evaluation metrics, even for large AI systems such as LLMs, to maintain optimal performance. According to [28], the risk assessment process should adopt a human-centered design approach. Moreover, besides human involvement in the RMS, users should be able to include references and descriptions of the data utilized for training, validation, and testing, consolidating all necessary information required by legal standards within the DMDGS. Additionally, users should be able to directly access and download technical documentation from the QMS to verify the test and validation proceedings on the AI system and ensure compliance with EU AIA regulations. Integrating all these functionalities into several sub-modules within a single tool aims to reduce effort, time, and costs for AI system providers and deployers. Table 3 lists all UI and human involvement requirements.

Table 3: System Design Requirements (Human Involvement)

System Design Requirements
SDR01: User Interface (FR):
The QMS shall provide a UI that actively engages users in the verification process of the AI system.
SDR02: Human-based Risk Management (FR):
Pages to empower users to participate in the risk identification, analysis, assessment, and mitigation processes shall be included.
SDR03: Data Page (FR):
Pages shall be included to upload or reference the data utilized for training, validation, and testing, accompanied by evidence of compliance.
SDR04: Downloadable Technical Documentation (FR):
The QMS shall allow users to view and download the technical documentation for the AI system assessment.

The QMS should have a modular design to react quickly to future changes in the legislative framework. The modular design is ensured by designing independent sub-services for each article in the EU AIA, which can be plugged into the QMS and maintained and updated separately at any time. Key modules include the RMS (cf. Article 9 EU AIA) and the DMDGS (cf. Article 10 EU AIA), which are integrated into the first version of the prototype QMS. Potential additional modules, such as the AI system's event logging (cf. Article 12 EU AIA), can be added in future work. The modules should be designed to apply generically to any type and architecture of AI system, except for the specific technical evaluation metrics, which need to align with the type and architecture of the underlying AI system. Additionally, the QMS allows users to persistently store technical documentation, past risk assessment processes, and data check references. This feature not only ensures comprehensive record-keeping but also facilitates easy access to historical data, which is crucial for maintaining compliance and continuous improvement. The goal is to enable the QMS to be utilized throughout the entire lifecycle of the AI system, including post-market. All requirements of the system design architecture are detailed in Table 4.

Table 4: System Design Requirements (Architecture)

System Design Requirements
SDR05: Modular Design (NFR): Enable users to customize and modify the QMS modules, add various AI systems, and incorporate technical metrics tailored to domains, purposes, and tasks.
SDR06: Communication and Data Flow (NFR): Implement reliable and secure data communication across the various modules of the QMS.
SDR07: Persistent Storage (NFR): Set up a database to persistently store all user information, technical documentation, and assessment processes.

Identifying the required computational resources is the third type of system design requirement. LLMs, for example, require significant computational power for tasks like calculating the gradients needed for some technical evaluation metrics. These gradients are necessary for certain adversarial attacking and explainability techniques. The specific GPU and CPU resources need to be tested, evaluated, and defined for a second version of the prototype QMS. Table 5 details the computational requirements for the system design.

Table 5: System Design Requirements (Computation)

System Design Requirements
SDR08: Computational Resource (NFR): Allocate sufficient computational resources, including GPU resources, to ensure that cost-intensive computations, even on large GPAI models, can be computed within an acceptable timeframe.
SDR09: Performance (NFR): Implement high-performance technical evaluation metrics for cost-intensive computations that minimize GPU storage usage and improve execution time.

3.2   Architecture and Data Models

The QMS architecture is based on a microservice design (cf. [25]). The current prototype QMS consists of an RMS module, a dedicated DMDGS module, and a user authentication module, which only stores user and login information. All modules are independent services containing their backend and database. The architecture of the QMS and the data models for the RMS, DMDGS, and user authentication databases as UML class diagrams are illustrated in Fig. 1.

Users can interact with the UI implemented on the frontend to participate in the verification and documentation process within the QMS by executing risk assessments and ensuring data management and governance compliance. An API Gateway backend service orchestrates all user requests from the frontend and forwards each request to the corresponding backend system. The API Gateway loads an environment file containing the names or pseudonyms for each sub-service root. The backend of the corresponding sub-service then sends the response back to the API Gateway, which forwards the message to the frontend, where it is loaded into the UI for the user. This structure, data transfer, and orchestration reduce complexity and improve modularity. Complexity is reduced, and modularity is ensured because adding sub-services to the QMS does not alter existing data transfers. To add a new service to the communication and data transfer pipeline, only an additional entry in the environment file read by the API Gateway is required. This prototype QMS focuses on the functionalities of the RMS rather than those in the DMDGS, as already mentioned. The RMS backend consists of two different components. It contains a verification component that can load different language models available from Hugging Face and implements all technical evaluation metrics presented in the
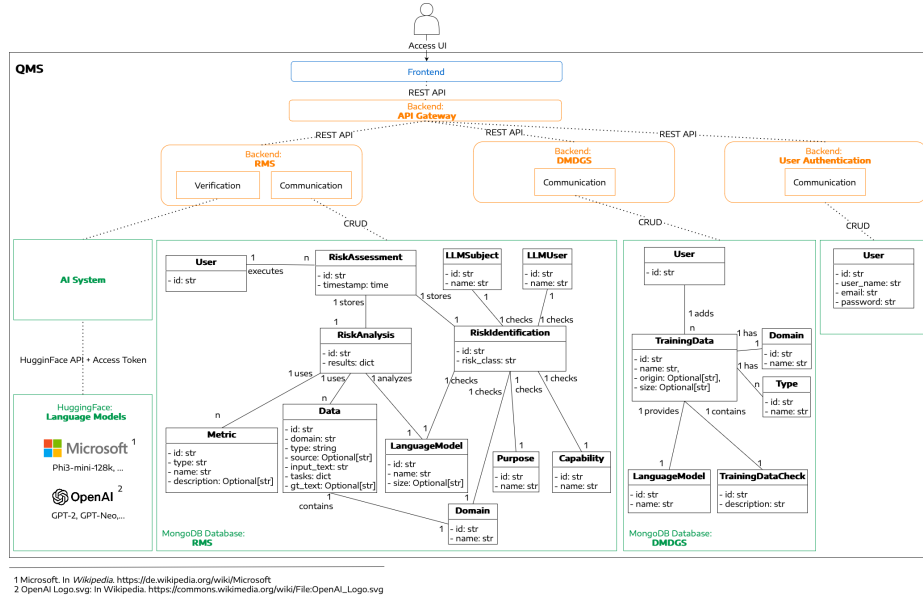
Fig. 1. Architecture and Data Models of the QMS

next section (cf. Sect. 3.3), which are used to analyze and quantitatively assess the LLM's performance, consistency, and explainability for specialized tasks in selected high-risk domains. Furthermore, the RMS backend contains a communication component responsible for frontend-backend-database communication and data transfer. Each sub-service, including the DMDGS and user authentication service, contains a communication component with the same structure and functionality. In contrast to the RMS, the DMDGS, and the user authentication services only include a communication component and no further components. The RMS data model consists of three main classes: risk identification, analysis, and assessment. A risk assessment object stores one risk identification and one risk analysis object. Users can execute multiple risk assessments, with the risk assessment IDs stored in the user object. This setup ensures that users can quickly review past risk assessment results in the UI. The DMDGS data model follows the same principle. The user class stores a list of data IDs, where each data object is linked to an LLM object and a data check object that verifies the compliance of the added data reference object. It is important to note that the RMS risk analysis requires data, which has a domain and task to assess the LLM's characteristics, efficiency, and risks (performance, explainability, and consistency). On the other hand, the data referenced in the DMDGS are the training, validation, and testing data to train and develop the LLM which will be assessed in the RMS after it is ready to use. The user authentication data model only contains a user class to store user IDs, personal data, and login

information. This database encrypts and hashes sensitive data and is made secure as the only database storing sensitive personal data. Whenever a new user signs up, the user is stored in the database, and the created user ID (random string created by MongoDB) is distributed to each user collection in the other sub-services databases, which ensures consistency, with each user collection storing identical user IDs. The following technology stack is used to implement the prototype QMS: The frontend is developed using JavaScript and the React.js [12] library. All backend services are written in Python, utilizing various libraries and packages such as PyTorch [13], Transformers [14] in the verification component to i) load the LLMs from Huggingface into the GPU and to ii) perform computations on the loaded LLMs, and FastAPI [15] and PyMongo [16] to implement all REST and CRUD methods to provide a communication pipeline from the frontend to the backend and from the backend to the database. MongoDB [17], a non-SQL database, is employed for data storage, allowing for rapid and straightforward modifications and design changes. MongoDB uses so-called collections (database tables) which can be modeled and designed using UML class diagrams and can always be extended or changed, keeping them flexible. The documents are the elements of a collection (rows of a table) and are stored in the collections as JSON data, maintaining a unified structure and language used in the frontend. A document can contain different collections of document data, similar to object-oriented programming, making MongoDB easy to understand.

3.3   Prototype QMS - Version 1

The first version of the prototype QMS can be accessed under the following URL: https://power.bpm.cit.tum.de/qmsAIA/.

Main Service: Quality Management System. The user accesses the QMS home page after signing up and signing in. The prototype QMS home page design is depicted in Fig. 2. To sign in, the user must enter his/ her username, email address, and password. If the user has no account, he/ she can click on the sign-up button to access the sign-up page and create a new account. The homepage is subdivided and structured into the prototype's several sub-services. Each functional sub-service (RMS and DMDGS) is listed in a separate row, one below the other, and contains two boxes each, one for performing risk assessments or data checks and one for viewing past risk assessments or data checks. The user authentication sub-service contains no UI components and is not classified as a functional sub-service, as its aim is solely to securely store user data. Additionally, the UI provides access to various sections with details about the LLMs, a page to add verification data for risk assessments, and relevant about

---

[12] React.js: https://react.dev, accessed on 7 May 2024.

[13] PyTorch: https://pytorch.org, accessed on 7 May 2024.

[14] Transf.: https://huggingface.co/docs/transformers/, accessed on 15 May 2024.

[15] FastAPI: https://fastapi.tiangolo.com, accessed on 7 May 2024.

[16] PyMongo: https://pymongo.readthedocs.io/en/, accessed on 24 July 2024.

[17] MongoDB: https://www.mongodb.com, accessed on 7 May 2024.

the EU AIA regulations. All these pages can be accessed through the buttons on the tab bar in the header.
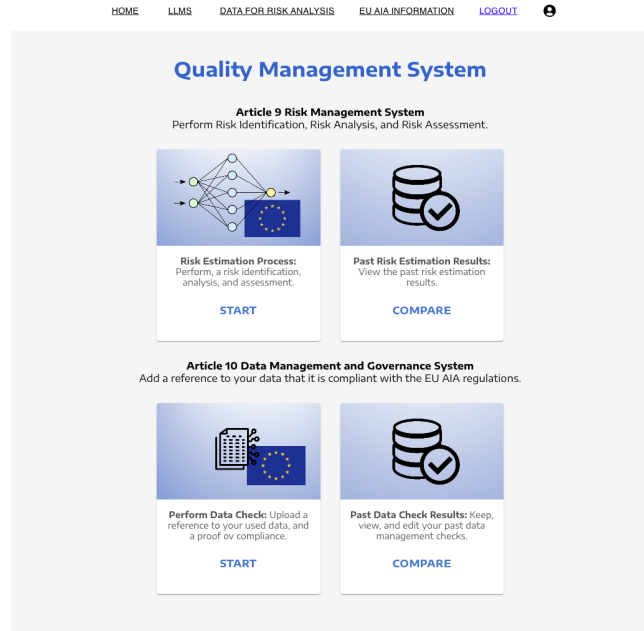

Fig. 2. QMS - Home

Sub-Service 01: Risk Management System. As mentioned, the main focus of the prototype QMS is on the RMS sub-service. The overall structure of the RMS is based on the ISO 31000 standard [15] which is also applied to the AI system RMS presented in [33]. It follows a *"Plan-Do-Act-Check"* principle: Plan involves planning the RMS based on the specific use case and defining the processes' sub-activities, Do encompasses conducting the designed RMS process, Act entails inspecting and evaluating the entire RMS process, and output, and Check involves refining and optimizing the RMS process. Additionally, [26] establishes five design principles for RMS for AI systems, which should also be integrated. These principles include i) incorporating multi-perspective expert assessment, such as drawing insights from various domains and involving AI experts in the process, ii) encouraging participation from diverse stakeholders in the risk assessment process, iii) identifying risks based on real-life scenarios, and iv) analyzing risks using metrics beyond accuracy, v) ensuring "human in the loop" processes for black-box models. This RMS can be seen as a prototype within the prototype QMS. The construction process of this RMS, according to ISO 31000, consists of an iterative process containing six sub-activities: component selection, risk identification, verification data selection, risk analysis, risk

assessment, and risk mitigation. In the component selection, the users should be encouraged to select the AI system (in this case, the LLM) that is analyzed, and the task the AI system performs. In the second step, the user conducts a risk identification based on the risk categorization strategy for AI systems presented by [12]. A vocabulary-based approach is used where the user adds values for the LLM's domain, purpose, capabilities, LLM user, and LLM subject. Based on these selected values, the risk class of the AI system is determined. A basic algorithm was implemented to evaluate the risk class according to the categorization presented by [12]. Future work will explore changing the risk identification process to a more stakeholder-oriented approach, as mentioned in design principle i) by [26]. Additionally, it will be considered whether risk identification should be designed to identify foreseeable risks and misuses, assuming the AI system is already classified as high-risk, instead of categorizing the AI system into a risk class. In the third step, the user can quantitatively analyze the AI system's performance, explainability, and consistency risks.

Fig. 3. QMS - RMS - Analysis

As shown in the UI screenshot in Fig. 3, the user can choose between different technical evaluation metrics applied in the risk analysis to the selected LLM. For performance metrics, the user can choose: i) *Accuracy score* (cf. [2]), ii) *Rouge-n score* (cf. [21]), and iii) *Perplexity* (cf. [14]). For explainability, a *gradient-based saliency-map* metric can be employed (cf. [20], initially invented for image classifiers: [30]). For consistency, a *gradient-based adversarial example input modification* method can be selected (cf. [34], initially invented for image classifiers: [32,13]). In the fourth step, the model calculates a risk assessment documentation based on the risk identification and analysis results. This documentation aims to verify compliance with EU AIA regulations. The current documentation will not serve as proof, but the basic principle and the feasibility of automatically creating and downloading a risk and technical documentation from the prototype QMS can be demonstrated. In the final step, the user has the option to add strategies to mitigate assessed risks. This feature still needs to be fully developed and integrated into the prototype's second version.

To provide an example of how such a risk analysis and assessment result looks, the process was demonstrated with test verification data, and all presented metrics were selected to be computed. As LLM the Phi-3-mini-128k-instruct [18] from Microsoft containing between 4 and 5 billion parameters was analyzed on the domain *"Industry Process Description"*, and the task *"Summarization"*. The model input was: *"Extract the Actor and Activity pairs from the text. Return only the list of JSON documents in the following format: ['actor': 'example_actor_1', 'activity': 'example_activity_1', 'actor': 'example_actor_2', 'activity': 'example_activity_2', ...] without any further explanation: The user creates a new process instance, then the system can execute the instance and stop it afterward."*.

Performance Metrics Results

Accuracy-based Performance:
For each task, data type, LLM combination, the ROUGE-N score and the Accuracy is calculated.

Task	Data Type	Rouge N-Score (F1)	Accuracy Score	Perplexity Score
Extract the Actor and Activity pairs from the text. Return only the list of JSON documents in the following format: [{'actor': 'example_actor_1', 'activity': 'example_activity_1'}, {'actor': 'example_actor_2', 'activity': 'example_activity_2'}, ...] without any further explanation.: The user creates a new process instance, then the system can execute the instance and stop it afterwards.	Legal Guideline	0.29	0.5	8.6

Fig. 4. QMS - RMS - Assessment - Performance Result

In Fig. 4, the performance results of the model from the risk assessment UI page are shown. The performance results display the numerical values for the model's accuracy, Rouge score, and perplexity for the given input tasks. In Fig.

---

[18] Phi-3-mini: https://huggingface.co/microsoft/Phi-3-mini-128k-instruct, accessed on 30 July 2024

5, the explainability result is visualized as a saliency map, where each input token is colored between red (unimportant, less sensitive) and green (important, sensitive) to the model's generated output.
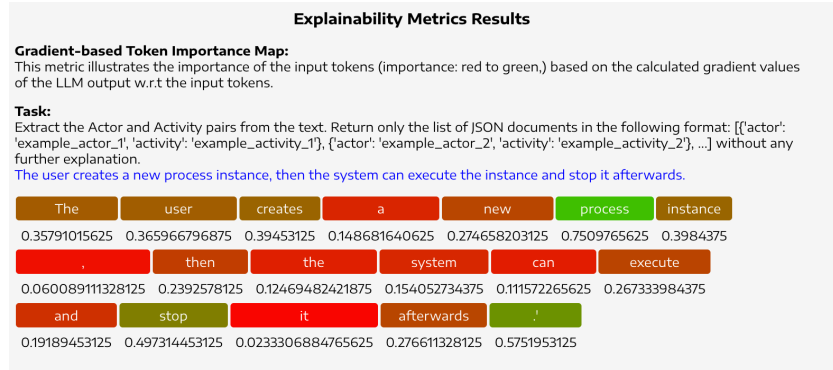

Fig. 5. QMS - RMS - Assessment - Explainability Result

The consistency result is depicted in Fig. 6. The consistency results display the ground-truth output (green), the adversarial output (red) generated by modifying the input tokens in the gradient direction using a hyperparameter similar to the learning rate, and the number of iterations to fool the model.
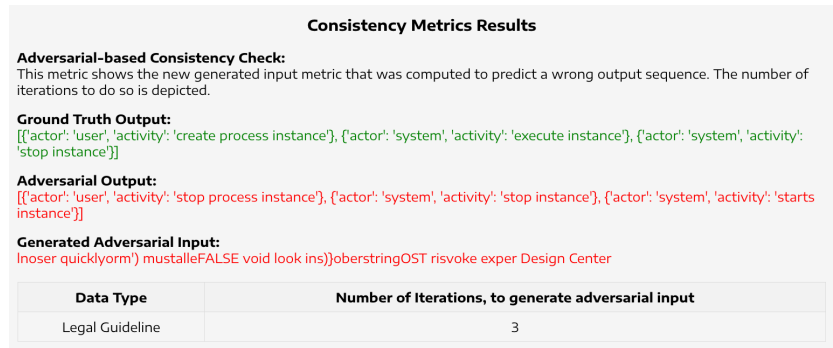

Fig. 6. QMS - RMS - Assessment - Consistency Result

Sub-Service 02: Data Management and Governance System. The DMDGS is the second sub-service in the QMS. In the current first version of the prototype QMS, the DMDGS is kept very basic. It consists of two components in the UI: the data check page and the past data check page. The former, as shown in Fig.

7, allows the user to select the AI system (in this case, LLM) for which training, validation, or testing data has been used. Furthermore, the user can specify the name of the dataset, the origin, the type, the domain, and the size. To provide proof that the data is checked and compliant under Article 10 EU AIA, the user has a text field in which he/ she can enter a textual reference to the proof.

Fig. 7. QMS - DMDGS - Data Check

In future work, the goal is to enhance the Data Management and Governance System (DMDGS) to provide functionality for analyzing and assessing datasets for training, validation, and testing quantitatively within the QMS. This includes evaluating datasets for biases and errors. Additionally, the QMS should verify that dataset splits are well-determined and that the data aligns with the goals of AI system development. It is also essential to check whether other EU data regulations, such as the GDPR[19], need to be considered within the QMS.

---

[19] GDPR - Website: https://gdpr-info.eu, accessed on 30 July 2024

4   Evaluation

The implementation and realization of functional (FR) and non-functional requirements (NFR) in the QMS are evaluated technically through a prototypical implementation in Sect. 3.3. In the sequel, FRs are evaluated based on user story scenarios and NFRs are qualitatively assessed based on technical evaluations and calculations regarding the required GPU storage and performance for the LLMs used.

FR Evaluation – User Stories:. A user story has the form: "*As a [provider, deployer, user. . .] [I want to. . .], [ensure, guarantee, provide. . .]. . .*". **LR01 – LR10** and **SDR01 – SDR04** are evaluated along with ten user story scenarios that exemplify potential stakeholders' needs within the QMS. The user stories are mainly structured to evaluate the constructed requirements chronologically. The evaluation also includes potential limitations, unimplemented requirements, and future work.

(1) "*As a provider or deployer of a high-risk AI or GPAI system, I want a QMS in the form of a web application where I can directly perform technical evaluations on the AI system (model), document these evaluations, and include data references and checks for used train, validate and test datasets of the AI system (model).*" (**LR01**, **LR02**). A prototype has been developed to provide an initial version of this QMS as a web application. While the prototype meets the requirements specified in the user story, several optimizations are needed. For example, these include improving the risk identification process, enabling simultaneous participation from multiple stakeholders in the RMS, and incorporating technical checks on datasets used for AI (training) development.

(2) "*As an AI Expert, I want to compute, download, and assess the technical evaluation results of the AI system in charge, discuss with legal experts the sufficiency of the model results, and see the current results in a certain domain or task for potential improvement.*" (**LR02**, **LR08**, **LR09**). The latter requirement mandates performing several technical evaluations on the AI system. For this QMS, LLMs have been used as an example of AI systems because they are already pre-trained and freely available on Hugging Face, making it easy to access and integrate them into the prototype QMS. Five different metrics have been implemented: three to evaluate the model's performance, one to evaluate its explainability, and one for its consistency (as a general term for robustness and reliability). However, no standard protocols such as the ISO 4213 [17] for AI performance evaluation, ISO CD TS 6254 [18] for AI explainability evaluation, and ISO TR 24029 for AI robustness evaluation [16] have been used as a guideline. The metrics and calculations are currently included in the risk analysis component. However, a different sub-service can also be created for LR08 (Article 15 EU AIA) to prove the AI system's accuracy, robustness, and cybersecurity separately. For that, the already implemented verification component can be reused. The download feature to comply with LR09 (Article 11 EU AIA) already exists. However, implementing requirements outlined by AI and legal experts to define the scope and design of such technical documentation would further enhance its

quality.

③ "*As a provider or deployer of a high-risk AI or GPAI system, I want at least one AI expert from the development team, a domain expert, and a legal expert to perform a risk assessment together in a certain domain for specific tasks, to identify foreseeable risks or misuses and to effectively mitigate the risk*" (**LR03**). An RMS is implemented using the ISO 31000 risk management standard [15] to guide and structure the process. The implementation includes essential risk identification, risk analysis with technical evaluation metrics on the LLM, and a risk assessment page. However, improvements are needed in the risk identification process, a solution to incorporate multiple stakeholders must be established, and the mitigation strategy page must still be implemented.

④ "*As an AI expert of the development team, I want to have a sub-service included in the QMS where I can reference and check the data used for the AI system's development, or include a data check, to ensure that my training, validation, and testing data is compliant with all privacy and data ownership regulations defined in legal frameworks. This will provide transparency and proof that can be shared with legal authorities and internally with all AI development members.*" (**LR04**). The QMS includes the presented DMDGS, but the user can only add textual descriptions of the data and the corresponding checks. Further improvements are required, such as uploading certified data checks and more in-depth analysis capabilities. Additionally, establishing and implementing a dedicated requirements section would enhance functionality.

⑤ "*As a provider or deployer of a high-risk AI system, I need to have a sub-service in the QMS that shows and stores the logs of the AI system's use to comply with the regulations of the EU AIA and to ensure that no decisions are made with the AI system in critical situations.*" (**LR05**). This requirement and the corresponding sub-service still need to be implemented.

⑥ "*As a high-risk AI or GPAI system provider, I want to better understand how the created AI system works by applying explainability metrics. This will help me improve the system in future updates and create the best possible and most transparent instructions for end-users. These instructions shall inform them about good practices, limitations, and risks as comprehensively as possible.*" (**LR06**). One explainability metric optimized for generative language models is implemented, helping users understand how much each input token influences the model's generated output. This is useful for analyzing which types of tokens and grammar styles work better to achieve the best possible outcome. In future work, more types of explainability metrics will be added, not only for generative models but also for classification and regression models, allowing users of the QMS to select the most suitable metrics. Additionally, more visualizations will be researched and developed to better interpret the explainability metrics. Instructions on how to interpret these results will be provided through the QMS UI. However, no considerations have been made yet on how to automatically create or collaborate with humans to generate instructions for using the AI system. Such a feature will also be integrated into future versions of the QMS.

⑦ "*As a user of the high-risk AI system, I want to have a UI to document risks*

and misuses and the potential errors and limitations of the AI system. In emergency cases, I want to be able to shut down the system." (**LR07**). This is not yet implemented. The aim is to have the QMS available not only for providers of AI systems and deployers who modify the system and must comply with all regulations for high-risk AI systems but also the end-users of these AI systems. End-users can document limitations and risks within the QMS, which will be directly sent to the provider. The provider can then add this feedback to their backlog and address the issues in future updates. The QMS will be one extensive system with different functionalities, depending on the license, for example, provider or end-user.

(8) "*As an AI Expert, I want to evaluate the AI system, specifically the high-risk AI or GPAI system, on performance, robustness, explainability, security, and fairness to verify its use according to the EU AIA regulations and to identify its limitations.*" (**LR08**). Five technical evaluation metrics for performance, explainability, and consistency have been implemented, allowing the user to evaluate the LLM in a specific domain for specific tasks. More metrics will be implemented, including those for classification and regression models. The design of the risk analysis component follows an object-oriented approach using the strategy software design pattern (cf. [6]). Additionally, no cybersecurity metrics have been implemented yet, as cybersecurity is only relevant for systems in production. However, integrating cybersecurity metrics will be addressed in future versions.

(9) "*As a provider or deployer of a high-risk AI or GPAI system, I want to check potential risks and perform model evaluations even aftermarket release to comply with the EU AIA and to detect potential new risks that have not been documented in the development phase.*" (**LR10**). This requirement is fully fulfilled due to the design choice of building the QMS as a web application and integrating the AI system into the QMS. The risk assessment, creation of technical documentation, and all other features provided by the QMS can be accessed and used anytime without additional expenses.

(10) "*As an AI expert, domain expert, legal expert, or any other person involved in the assessment and documentation process of the high-risk AI or GPAI system, I want to have an application and user interface to easily interact with my colleagues, the model, and the tasks such as risk management and data governance, to ensure a transparent and reliable compliance management process, reduce costs, improve communication through a platform, and to create the proof of conformity according to the EU AIA regulations.*" (**SDR01**, **SDR02**, **SDR03**). The presented design of the QMS is based on a SaaS approach. The idea is to have one system connected to an AI system that contains multiple sub-services, ideally one for each EU AIA regulation. The results from each sub-service are stored persistently in a database and can be printed as PDF documentation. However, some limitations still exist, such as the absence of user roles and the lack of support for collaboration, communication, and interaction among multiple users within the same process, such as in a risk assessment scenario. These features will be implemented in the second version of the prototype QMS.

NFR Evaluation - Design, Memory, and Performance. The QMS utilizes a microservice architecture. Each sub-service can be adopted, extended, and maintained independently of the others, guaranteeing a loosely coupled, and modular design (**SDR05**). All sub-services communicate via REST through the API Gateway with the frontend. Additionally, each sub-service backend communicates with its database using CRUD operations, ensuring persistent storage of all user and QMS data. (**SDR06** and **SDR07**). Analyzing LLMs as a type of AI system is a good choice for determining the required GPU VRAM (memory), GPU, and CPU performance because these AI systems (models) are significantly larger than other AI systems. If the GPU and CPU resources are sufficient for LLMs, they should be more than adequate for most other AI systems and models. GPU VRAM is crucial for computing the gradients of LLMs and obtaining results for the presented explainability and consistency metrics. This is because all gradients are calculated at once, meaning that all parameters must be stored in the GPU's VRAM simultaneously. The following calculation determines the required GPU VRAM for inferencing, gradient calculations, and one optimization step using a batch size of one. For instance, a 7B model comprises seven billion parameters (each including a weight and a bias term). Calculating the logits and gradients with full precision, meaning in float32, requires that each parameter be represented by 32 bits or 4 bytes. Therefore, a single forward pass necessitates: $7 * 10^9 \ parameters * 4 \ bytes = 28 * 10^9 \ bytes$ (= 28GB). The same amount is again required for gradient computation, meaning an additional 28 GB is needed. Thus, approximately 56 GB of GPU VRAM is required for inference and gradient calculation. Switching to float16 precision from float32 has a significant impact on memory requirements. With the memory requirement for each parameter halved, only 28GB of memory is needed to compute a forward step and the gradients for a batch size of one. This underscores the potential for memory optimization and its impact on the overall system performance. By implementing optimizations such as gradient accumulation and other techniques, memory usage can be further reduced. The prototype QMS uses an NVIDIA RTX 4090 GPU with 24GB of GPU VRAM. This amount is insufficient for 7B LLMs like Meta's LLama2-7B, and all larger Meta LLMs such as LLama3-8B. Therefore, a smaller model, such as Microsoft's Phi3-mini, OpenAI's GPT-2, and GPT-neo are integrated into the QMS. Phi-3-mini, which stores between 4 and 5 billion parameters, performs well in test cases, providing good results for all five technical evaluation metrics in 10 to 20 seconds. The Phi3-mini requires at most 20GB of GPU VRAM, making it suitable for the NVIDIA RTX 4090. The smaller language models can also be used as needed but do not deliver good results in summarization tasks. (**SDR08**, **SDR09**). In future work, memory and performance optimization will be further explored. Additionally, tests with real-life data will be conducted to evaluate the effectiveness of the metrics and the performance of different LLMs. Furthermore, other types of AI systems will be integrated such as neural networks for classification tasks. These models are significantly smaller than LLMs and can therefore be well-evaluated using the NVIDIA RTX4090 GPU.

5   Conclusion

The presented quality management system (QMS) is a tool for ensuring compliance with the EU AIA regulations for high-risk AI and GPAI systems, guided by legal and system design functional (FR) and non-functional requirements (NFR). The QMS is based on a microservice architecture, directly connecting AI systems and containing several sub-services, each with a different purpose based on the regulations for high-risk AI or GPAI systems of the EU AIA. This SaaS web application aims to map the compliance management processes for AI systems (especially high-risk AI and GPAI systems) into one tool and to carry them out efficiently. Quantitative tests can be directly applied to the adopted AI system within the QMS. Although the presented first version prototype QMS is optimized to check and document LLMs, its concept, design, and architecture can be applied to various types of AI systems. Currently, the QMS integrates a risk management system (RMS) (cf. Article 9 EU AIA) and a data management and governance system (DMDGS) (cf. Article 10 EU AIA) as sub-services. Further sub-services, such as logging the use of the AI system (Article 12 EU AIA), transparency approaches on how to create instructions for use (cf. Article 13 EU AIA), and a human-machine interface for users of the AI system (cf. Article 14 EU AIA), will be added in future versions of the prototype QMS. Additionally, each sub-service sub-component can be improved in the future, particularly the identification and mitigation components requiring stakeholder and AI user involvement. Currently, risk identification is not linked with risk analysis, such as by suggesting technical evaluation metrics or tools for addressing potential risks manually or automatically. Future work will focus on better integrating risk identification results with the risk analysis and assessment components. The idea of this paper is related to previous compliance management work (cf. [22,23]). The goal for future versions of the prototype QMS is to integrate findings and technical features from earlier automated compliance verification research and insights from business process management research, such as modeling larger processes by combining regulations for AI development and post-development when the system is in use.

References

1. André Steimers, T.B.: Sources of risk and design principles of trustworthy artificial intelligence (2021). https://doi.org/10.1007/978-3-030-77820-0_18
2. Banerjee, D., Singh, P., Avadhanam, A., Srivastava, S.: Benchmarking LLM powered chatbots: Methods and metrics. CoRR abs/2308.04624 (2023). https://doi.org/10.48550/ARXIV.2308.04624, https://doi.org/10.48550/arXiv.2308.04624
3. Bhaumik, D., Dey, D.: An audit framework for technical assessment of binary classifiers. In: International Conference on Agents and Artificial Intelligence. vol. 2, pp. 312–324 (2023). https://doi.org/10.5220/0011744600003393
4. Bormhard/ Siglmüller: Ai act – das trilogergebnis. RDi 2024 45 (2024)
5. Bronner: Die ki-verordnung (ki-vo) der eu. juris PR-ITR 13/2024 Anm.2 (2024)
6. Brügge, B., Dutoit, A.H.: Object-oriented software engineering - using UML, patterns and Java (2. ed.). Prentice Hall (2004)

7. Clarke, R.: Principles and business processes for responsible ai. Computer Law and Security Review **35**(4), 410–422 (2019). https://doi.org/10.1016/j.clsr.2019.04.007
8. Ebers: Die ki-verordnung ante portas: Ein neuer rechtsrahmen für legal tech? LTZ 2024 **1** (2024)
9. EU: Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act) , http://data.europa.eu/eli/reg/2024/1689/oj
10. Future of Life Institute (FLI): General purpose ai and the ai act (2022)
11. Giudici, P., Centurelli, M., Turchetta, S.: Artificial intelligence risk measurement. Expert Systems with Applications **235** (2024). https://doi.org/10.1016/j.eswa.2023.121220
12. Golpayegani, D., Pandit, H.J., Lewis, D.: To be high-risk, or not to be - semantic specifications and implications of the ai act's high-risk ai applications and harmonised standards. In: ACM International Conference Proceeding Series. pp. 905–915 (2023). https://doi.org/10.1145/3593013.3594050
13. Goodfellow, I.J., Shlens, J., Szegedy, C.: Explaining and harnessing adversarial examples. In: 3rd International Conference on Learning Representations, ICLR (2015), http://arxiv.org/abs/1412.6572
14. Hu, J., Gauthier, J., Qian, P., Wilcox, E., Levy, R.: A systematic assessment of syntactic generalization in neural language models. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020. pp. 1725–1744. Association for Computational Linguistics (2020). https://doi.org/10.18653/V1/2020.ACL-MAIN.158
15. International Organization for Standardization: ISO 31000:2018 - Risk management — Guidelines (2018)
16. International Organization for Standardization: ISO TR 24029 - Artificial Intelligence (AI) — Assessment of the robustness of neural networks (2021)
17. International Organization for Standardization: ISO 4213:2022 - Artificial Intelligence — Assessment of machine learning classification performance (2022)
18. International Organization for Standardization: ISO CD TS 6254 - Artificial Intelligence — Objectives and approaches for explainability and interpretability of ML models and AI systems (under development)
19. Jaeckel: Künstliche intelligenz im europäischen datenraum am beispiel der medizinprodukte. SächsVBl 2023 pp. 194–202 (2023)
20. Krishna, S., Ma, J., Slack, D., Ghandeharioun, A., Singh, S., Lakkaraju, H.: Post hoc explanations of language models can improve language models. In: Advances in Neural Information Processing Systems 36 (NeurIPS 2023) (2023), http://papers.nips.cc/paper_files/paper/2023/hash/ce65173b994cf7c925c71b482ee14a8d-Abstract-Conference.html
21. Lin, C.Y.: Rouge: A package for automatic evaluation of summaries. In: Text summarization branches out. pp. 74–81 (2004)
22. Mustroph, H., Barrientos, M., Winter, K., Rinderle-Ma, S.: Verifying resource compliance requirements from natural language text over event logs. In: Business Process Management - 21st International Conference, BPM. pp. 249–265. Springer (2023). https://doi.org/10.1007/978-3-031-41620-0_15
23. Mustroph, H., Winter, K., Rinderle-Ma, S.: Social network mining from natural language text and event logs for compliance deviation detection. In: Cooperative In-

formation Systems - 29th International Conference, CoopIS. pp. 347–365. Springer (2023). https://doi.org/10.1007/978-3-031-46846-9_19

24. Mökander, J., Schuett, J., Kirk, H.R., Floridi, L.: Auditing large language models: a three-layered approach. AI and Ethics (2023). https://doi.org/10.1007/s43681-023-00289-2

25. Nadareishvili, I., Mitra, R., McLarty, M., Amundsen, M.: Microservice architecture: aligning principles, practices, and culture. " O'Reilly Media, Inc." (2016)

26. Nagbøl, P.R., Müller, O., Krancher, O.: Designing a risk assessment tool for artificial intelligence systems. In: Lecture Notes in Computer Science. vol. 12807 LNCS, pp. 328–339 (2021). https://doi.org/10.1007/978-3-030-82405-1_32

27. Novelli, C., Casolari, F., Rotolo, A., Taddeo, M., Floridi, L.: Taking ai risks seriously: a new assessment model for the ai act. AI and Society (2023). https://doi.org/10.1007/s00146-023-01723-z

28. Ortega, E., Tran, M., Bandeen, G.: Ai digital tool product lifecycle governance framework through ethics and compliance by design†. In: IEEE Conference on Artificial Intelligence (CAI). pp. 353–356 (2023). https://doi.org/10.1109/CAI54212.2023.00155

29. Schallbruch, M.: Eu-regulierung der künstlichen intelligenz: Informationstechnische systeme im fokus neuer rechtlicher anforderungen. Datenschutz und Datensicherheit-DuD **45**, 438–443 (2021)

30. Simonyan, K., Vedaldi, A., Zisserman, A.: Deep inside convolutional networks: Visualising image classification models and saliency maps. In: 2nd International Conference on Learning Representations, ICLR 2014, Workshop Track Proceedings (2014), http://arxiv.org/abs/1312.6034

31. Sovrano, F., Sapienza, S., Palmirani, M., Vitali, F.: A survey on methods and metrics for the assessment of explainability under the proposed ai act. In: Frontiers in Artificial Intelligence and Applications. vol. 346, pp. 235–242 (2021). https://doi.org/10.3233/FAIA210342

32. Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I.J., Fergus, R.: Intriguing properties of neural networks. In: 2nd International Conference on Learning Representations, ICLR (2014), http://arxiv.org/abs/1312.6199

33. Tjoa, S., Temper, P.K.M., Temper, M., Zanol, J., Wagner, M., Holzinger, A.: Airman: An artificial intelligence (ai) risk management system. In: International Conference on Advanced Enterprise Information System (AEIS). pp. 72–81 (2022). https://doi.org/10.1109/AEIS59450.2022.00017

34. Yao, J., Ning, K., Liu, Z., Ning, M., Yuan, L.: LLM lies: Hallucinations are not bugs, but features as adversarial examples. CoRR **abs/2310.01469** (2023). https://doi.org/10.48550/ARXIV.2310.01469