Journal of
Information Technology,
Cybersecurity, and Artificial Intelligence

(ARTICLE)

MACHINE LEARNING (ML) TO EVALUATE GOVERNANCE, RISK, AND COMPLIANCE (GRC) RISKS ASSOCIATED WITH LARGE LANGUAGE MODELS (LLMs)

UPAKAR BHATTA

Assistant Professor, Information Technology Management Department
Central Washington University, Ellensburg, WA
Upakar.Bhatta@cwu.edu

Journal of Information Technology, Cybersecurity, and Artificial Intelligence, 2025, 2(2), 107-118

Article DOI: https://doi.org/10.70715/jitcai.2025.v2.i2.022

ABSTRACT

In today's AI-driven digital world, Governance, Risk, and Compliance (GRC) has become vital for organizations as they leverage AI technologies to drive business success and resilience. GRC represents a strategic approach that helps organization using Large Language Models (LLMs) automation tasks and enhances customer service, while maintaining regulatory complexity across various industries and regions. This paper explores a machine learning approach **to evaluate Governance, Risk, and Compliance (GRC) risks** associated with **Large Language Models (LLMs).** It utilizes Azure OpenAI Service logs to construct a representative dataset, with key features including response_time_ms, model_type, temperature, tokens_used, is_logged, data_sensitivity, compliance_flag, bias_score, and toxicity_score. These features are used to train a model that predicts GRC risk levels in LLM interactions, enabling organizations to improve efficiency, foster innovation, and deliver customer value, while maintaining compliance and regulatory requirements.

Keywords: Artificial Intelligence, Machine Learning, Large Language Model, Governance Risk Compliance

1. INTRODUCTION

Large language model (LLM), a subset of Generative AI, are rapidly evolving within the business landscape. They help modern organizations automate customer support, document summarization, software development and many other tasks. Their capacity to generate natural language at an unprecedented scale enables modern organization to enhance data-driven decision-making. However, deploying LLM-enabled system introduces a governance risk and compliance challenges. Since these systems use vast amounts of data for training, they raise data privacy concern such as unintentional data leaks and breaches. The number of organizations interested in deploying LLM-based AI systems is steadily increasing. However, the black-box nature of these systems has raised several ethical issues [1]. To address these concerns and promote responsible AI systems, various research groups across the world have defined principles for ethical AI usage. Expert research teams have developed guidelines emphasizing transparency and explainability in AI system design [5,7,8]. Organizations often utilize diverse machine learning models and algorithms in their decision-making processes. However, the outputs and the decisions of AI systems are usually difficult to understand and lack transparency [2]. This paper explores a machine learning approach to evaluate Governance, Risk, and Compliance (GRC) risks associated with Large Language Models (LLMs), enabling organizations to improve efficiency, foster innovation, and deliver customer value, while maintaining compliance and regulatory requirements. To assess the effectiveness of GRC-related risks in LLM, this paper aligns its conceptual model with NIST AI Risk Management Framework.

Figure 1: NIST Risk Management Framework overview. Adapted from *NIST*, 2023.

The above figure shows the four pillars of NIST AI Risk Management Framework include Govern, Map, Measure, and Manage

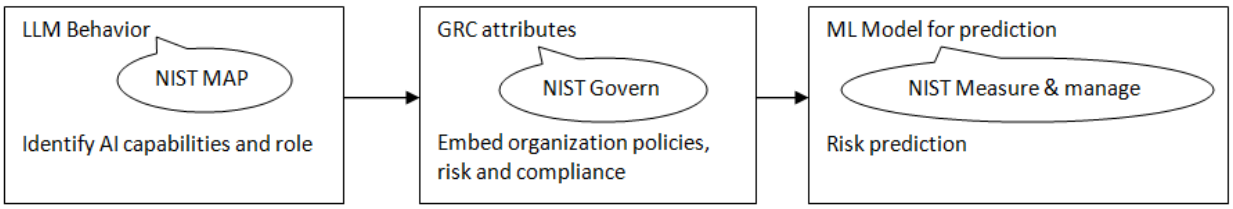

Figure 2: Conceptual linking diagram LLM behavior, GRC attributes, and ML model

The above figure shows the conceptual diagram linking LLM behavior, GRC attributes, and ML prediction model ensuring structured based approach to accessing the effectiveness of GRC-related risks. This paper maps **NIST AI Risk Management Framework** to LLM, GRC, and ML components to evaluate policy enforcement.

2. PRE-REQUISITE KNOWLEDGE

2.1. Cloud computing

Cloud computing is the on-demand delivery of virtual IT resources over the internet, providing a foundational infrastructure for deploying scalable AI systems. Cloud service models such as Infrastructure as a Service (IaaS), Platform as a Service (PaaS), and Software as a Service (SaaS) along with deployment models such as public, private, community, and hybrid clouds, allow organizations to align their technical requirements with business strategic objectives [14]. These cloud services and models play critical roles in hosting ML pipelines that evaluate LLM deployments. The cloud computing services that modern organizations are leveraging today include:

Infrastructure as a Service (IaaS): Provides the highest level of control over the computer, storage, and networking infrastructure. Deploying ML based agentic AI systems within IaaS enables organizations to monitor LLM behavior and enforce custom GRC policies.

Platform as a Service (PaaS): Provides secure ML deployment platform without managing underlying infrastructure. It enables customers to monitor LLM behavior in real-time against GRC thresholds.

Software as a Service (SaaS): Delivers vendor managed solutions on a subscription basis. However, organizations must integrate ML based monitoring to assess LLM interaction and ensure compliance with GRC standard.

2.2. Big data analytics

Big data analytics is vital for analyzing and validating the ML agent behavior in support for GRC enforcement. It enables data collection, processing, visualization of normal and abnormal pattern in LLMs interactions. Companies can apply advanced analytical tools and technologies to uncover hidden trends and extract valuable business intelligence [6]. By applying advanced analytical techniques, organizations can continuously monitor LLMs behavior and maintain visibility into GRC metrics.

2.3. Machine learning

Machine learning is a mathematical technique that is used to identify patterns and detect compliance violation. Machine learning is a field of Artificial Intelligence (AI) introduces a new approach for training algorithms with real-time datasets to identify specific patterns and anomalies in network traffic [11]. There are three main categories of machine learning: supervised learning, unsupervised learning, and reinforcement learning.

Supervised learning can be beneficial for predicting risk indicators using labeled datasets. In unsupervised learning, the algorithm is trained on unlabeled data to predict unexpected behaviors in LLM responses. In reinforcement learning, ml model is trained through trial and error learning to optimize ML behavior.

2.4. Governance, risk, and compliance framework

The GRC framework provides a structured approach that enables organizations to align AI solution with internal policies and external regulatory requirements. Governance ensures the responsible use of LLMs, risk management develops the risk assessment strategy to manage risk associated with LLMs deployment, and compliance ensures that LLMs output meets the legal and regulatory requirements.

3. RELATED WORK

LLM can reveal sensitive data that threatens privacy or security, raising concern about their compliance and regulatory implications. Several studies have previously explored transparency for LLM, yet a comparative evaluation of compliance risks associated with LLMs remains limited. In [16], ML communities are primarily aimed at supporting a mechanistic understanding of how the model or system functions by disclosing its components and processes. In [17], explanations of AI systems had been identified as contributing to greater system transparency. Furthermore, explanations of machine learning and AI outputs have been proposed as a means to mitigate transparency-related challenges [9,4]. In [13], the author outlined the importance of taking a human-centered perspective on transparency. However, previous research doesn't offer a comparative machine learning framework that evaluates regulatory risks considerations. This paper addresses this gap by employing machine learning model to assess the compliance risks associated with LLM. Furthermore, existing research lacks the analysis of governance models relevant to LLM such as **NIST AI Risk Management Framework** (2023), which offers structured approaches to AI governance and risk management [3]. AI auditing tools, such as **AuditPal AI, MindBridge,** and **Deloitte's Omnia platform,** demonstrate how LLM can support multistep audit processes and validate its behavior in regulated environments [10]. Recent works have outlined the importance of LLM transparency and privacy risk mitigation. Reports from the European Data Protection Board and Stanford's AI Index emphasized the importance of real-time auditing, bias detection, and regulatory compliance in LLM-powered systems [12]. Emerging models also emphasized **economic metrics for LLM** can be integrated into ML pipelines for predictive risk evaluation [15].

4. RESEARCH METHOD

This research paper explores the application of machine learning techniques to assess Governance, Risk, and Compliance (GRC) risks associated with Large Language Models (LLMs). The study demonstrates how various machine learning algorithms can be leveraged to evaluate security risks in alignment with compliance and regulatory requirements. The experimental methodology incorporates Azure services logs to construct a sample dataset capturing LLM interaction. The dataset undergoes preprocessing steps including cleaning, encoding and normalization, followed by exploratory data analysis (EDA) to identify risk patterns and compliance indicators. It is then split into training and testing segments, with Synthetic Minority Oversampling Technique (SMOTE) applied to address class imbalances. Machine learning model are selected, trained, tuned, and deployed to predict GRC violations. Model performance is evaluated using appropriate metrics to ensure that the outcomes align with compliance and regulatory standards.

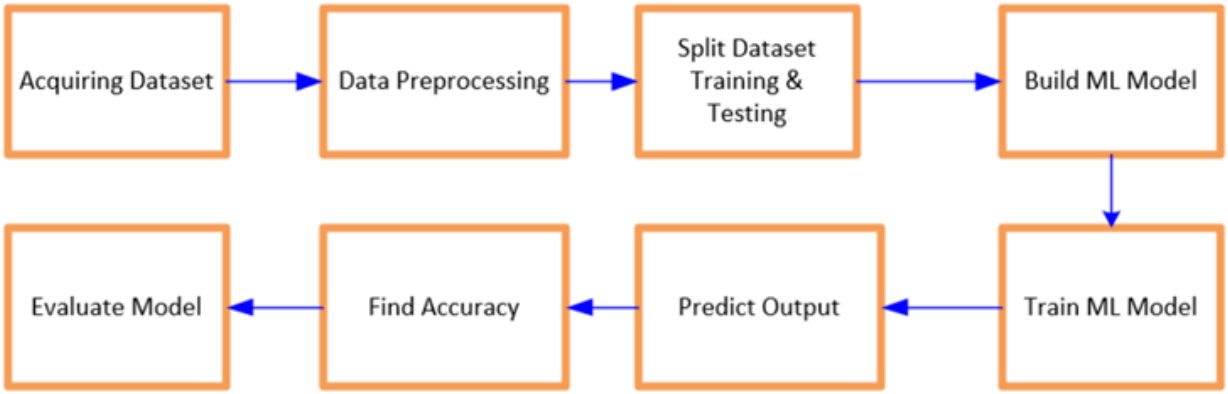

Figure 3: Machine learning implementation workflow

A modular ML pipeline was designed with the components such as feature extraction, model architecture, and assessing accuracy of machine learning for GRC (Governance, Risks, and Compliance evaluation).

Data Collection: Security log from Azure services were used to construct a sample dataset consisting of 5000 records.

Feature selection: A total of nine features were included: response_time_ms, model_type, temperature, tokens_used, logged, data_sensitivity, compliance_flag, bias_score, and toxicity_score.

Data preprocessing: Data cleaning, encoding, and normalization were performed.

Exploratory Data Analysis (EDA): Data visualization techniques such as Correlation heatmap was utilized to detect outliers, observe feature distributions, and identify relations between features.

Split the dataset: The dataset was divided into training (80%) and testing (20%) sets.

Synthetic Minority Over-sampling Technique (SMOTE): SMOTE was applied to address the class imbalance in the dataset.

Model selection: Mathematical algorithms were selected to train the machine learning models. The selected models include Random forest, Decision Tree, KNN, Logistic Regression, and Support Vector Machine. These models were selected based on their suitability for tabular data and their ability to handle classification tasks effectively.

Model training: Five machine learning models were trained. Hyperparameter tuning and cross-validation techniques were employed to select the optimal configuration optimize model performance, and ensure robustness.

Model Accuracy and evaluation: Machine learning models were evaluated using performance metrics including Accuracy (overall correctness), Precision (correctness of positive predictions), Recall (ability to detect true positives), and F1-score (balanced measure of precision and recall).

Interpretability and Transparency in ML prediction: To ensure trust and to demonstrate transparency in ML model predictions, important features were extracted from tree-based models and confusion matrix were analyzed to identify misclassification patterns.

5. DATA ANALYSIS

The dataset used in this research include numerical features, such as response_time_ms, model_type, temperature, tokens_used, logged, data_sensitivity, compliance_flag, bias_score, and toxicity_score. The exploratory data analysis methods applied in this research include:

Correlation Heatmap: The correlation matrix shown below highlights the numerical features used in the machine learning model for GRC evaluation. Dark red indicates a high positive correlation value, while data blue indicates a negative correlation. Features such as response_time_ms, and relationship between bias_score compliance_flag show very weak negative correlations. Overall, the heatmap reveals no strong correlation among the features. Features such as tokens_used, temperature, and model_type show low interdependence. The features independence indicates less redundancy and helps reduce the risk of overfitting.

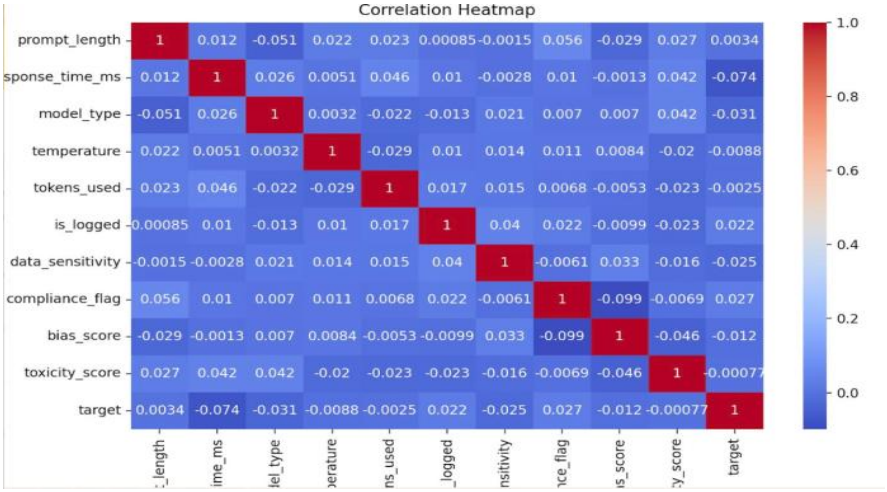

Figure 4: Correlation Heatmap

Feature distribution: This following figure shows the histograms of different features distributions in the dataset, including prompt_length, response_time_ms, temperature, tokens_used, bias_score, and toxicity_score for GRC evaluation. The histogram indicate that most features follow a uniform disctribution, suggesting that the dataset is well balance for model training. Features such as respons_time_ms exhibit a normal distribution, providing key predictive value. The absense of significant outliers across most features indicates consistent data quality, which is beneficial for training machine leaning models.

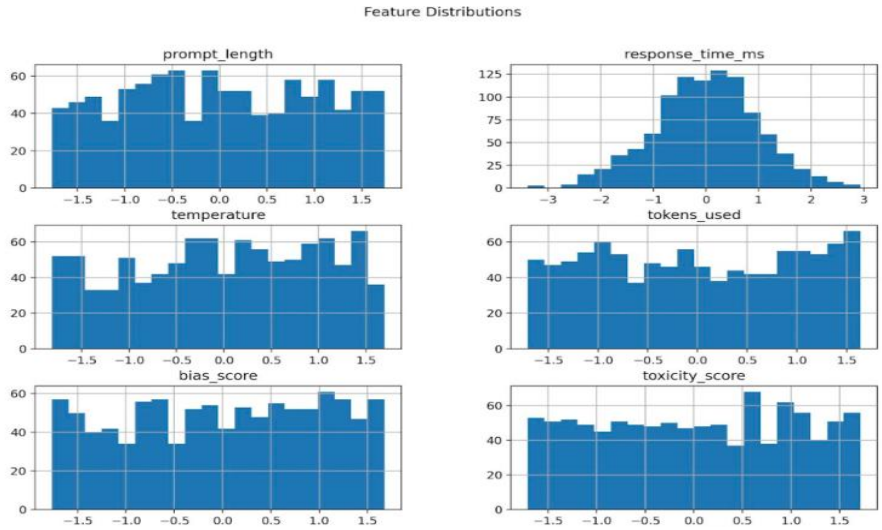

Figure 5: Histogram of feature distribution

Boxplot: The Boxplot of bias_score compares the distribution of bias score across two target classes, 0 and 1. It displays well distributed data with no bias values, indicating data consistency and fairness.

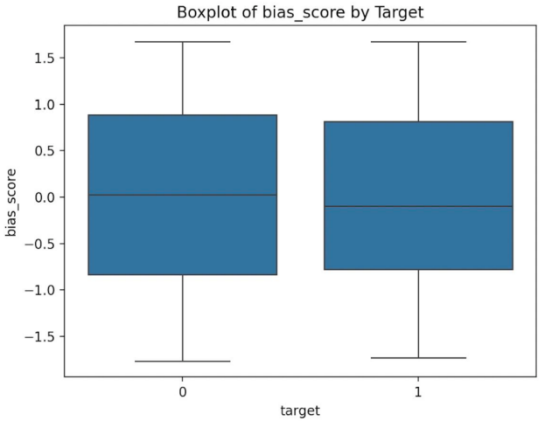

Figure 6: Boxplot of features used in the dataset

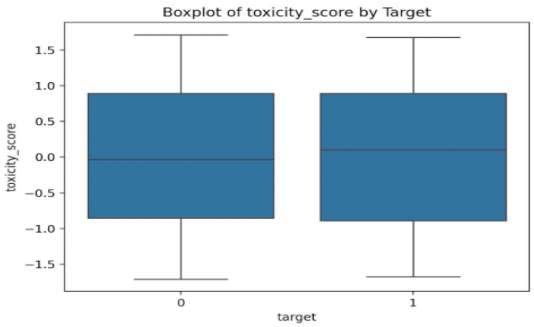

Figure 7: Boxplot of features used in the dataset

The Boxplot of toxicity score displays the minimal separation between target classes 0 and 1, indicating weak predicator on its own

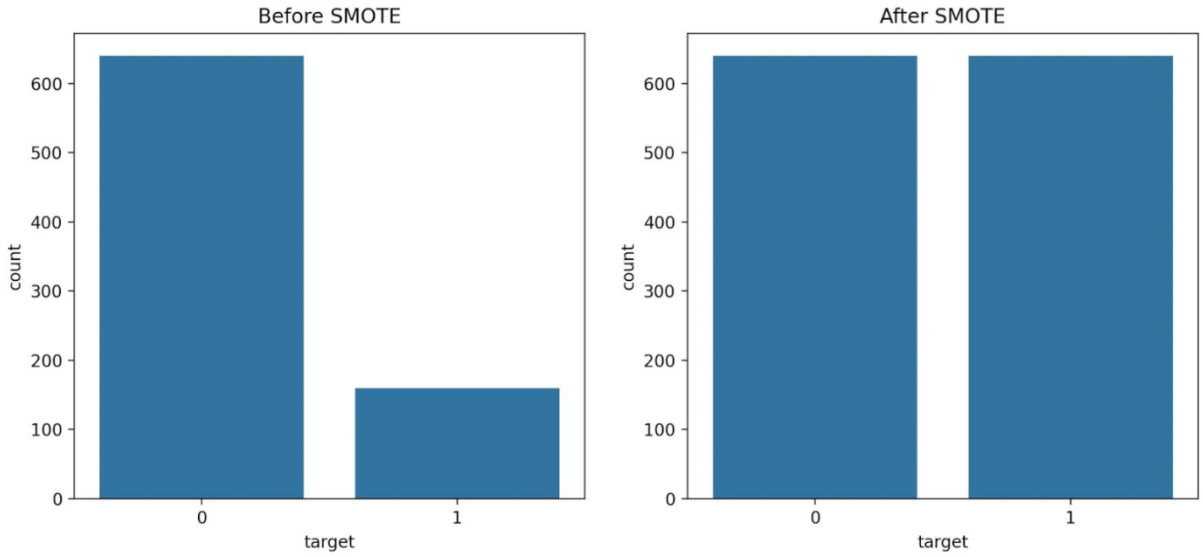

Figure 8: Class distribution before and after SMOTE

Synthetic Minority Over-Sampling Technique (SMOTE): SMOTE technique has been applied to address class imbalance issues, helping machine learning models avoid bias toward the majority class and improve overall performance.

6. RESULTS

This paper leverages Azure services logs to construct a sample dataset, **containing** 5,000 labeled instances across nine features: response_time_ms, model_type, temperature, tokens_used, is_logged, data_sensitivity, compliance_flag, bias_score, and toxicity_score. These features were used to train a model designed to help organizations to protect their systems and data. The machine learning model demonstrated good performance, achieving an accuracy of above 72% for evaluating **Governance, Risk, and Compliance (GRC) risks** associated with large language models (LLMs).

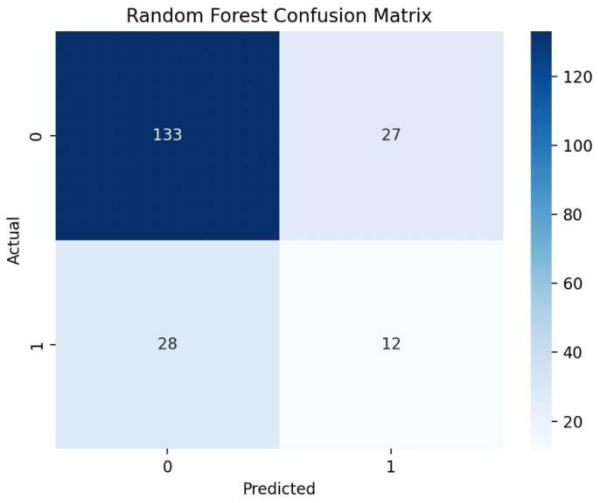

Figure 9: Confusion matrix random forest

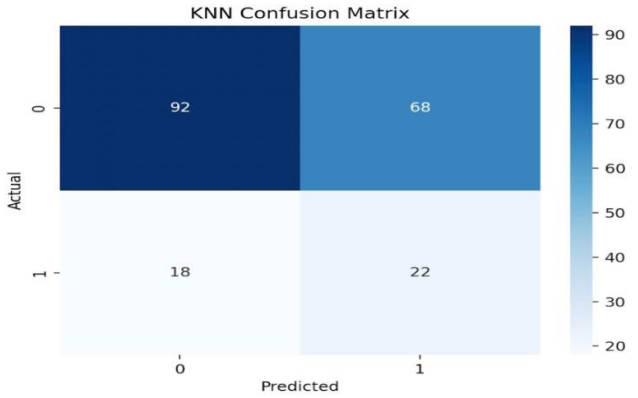

Figure 10: Confusion matrix k-nearest neighbors

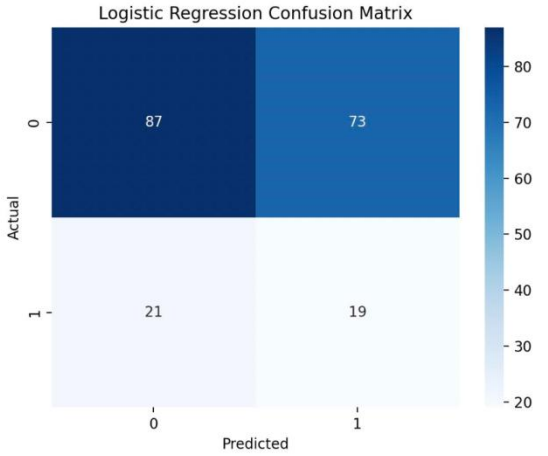

Figure 11: Confusion matrix logistic regression

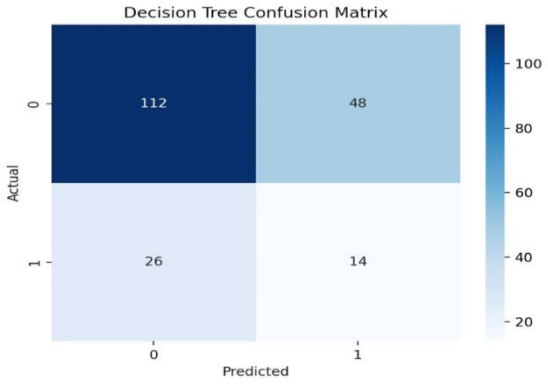

Figure 12: Confusion matrix decision tree

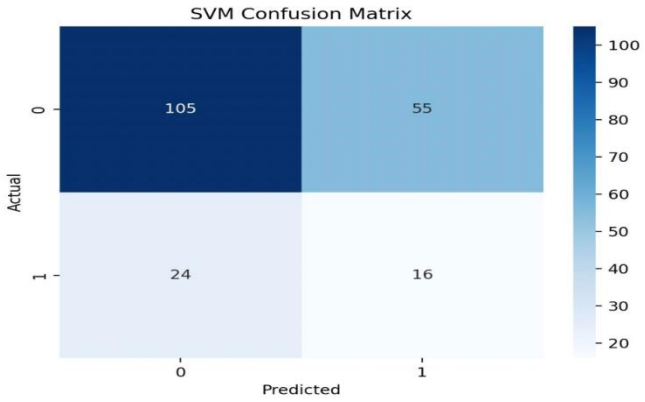

Figure 13: Confusion matrix SVM

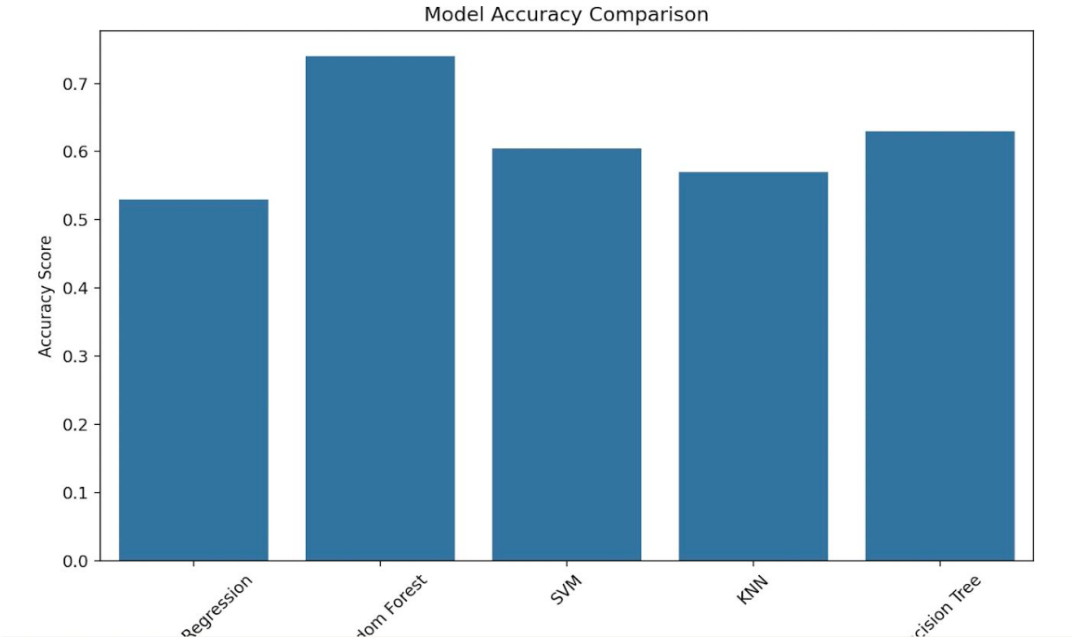

Figure 14: Model accuracy

7. DISCUSSIONS RESULTS

Based on the sample outlined in the confusion matrix, five machine learning models were evaluated for their effectiveness in predicting Governance, Risk, and Compliance (GRC) risks associated with Large Language Models (LLMs). The performance metrics are summarized in Table 1.

Table 1: Model metrics

Random forest	Decision Tree	KNN	Logistic Regression	Support Vector Machine

TP= 12 TN= 133 FP= 27 FN= 22 A= 74 P= 30.8 R= 30 F1-Score= 30	TP= 14 TN= 112 FP= 48 FN= 26 A= 63 P= 22.6 R= 35 F1-Score= 27.5	TP= 22 TN= 92 FP= 68 FN= 18 A= 57 P= 24.4 R= 55 F1-Score= 33.8	TP= 19 TN= 87 FP= 73 FN= 21 A= 53 P= 20.7 R= 47.5 F1-Score= 28.9	TP= 16 TN= 105 FP= 55 FN= 24 A= 60.5 P= 22.5 R= 40 F1-Score= 28.9

$$Accuracy\ (A) = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

$$Precision\ (P) = \frac{TP}{TP+FP} \quad (2)$$

$$Recall\ (R) = \frac{TP}{TP+FN} \quad (3)$$

$$F1 - Score = \frac{2*(P*R)}{P+R} \quad (4)$$

Where TP is true positive, TN is true negative, FP is false positive, and FN is false negative.

Random forest performance was better than other models across most metrics. It combines multiple decisions tress to reduce variance which makes it suitable for noisy datasets. Random Forest was chosen for evaluating **Governance, Risk, and Compliance (GRC) risks** associated with large language models (LLMs) as it outperformed all other tested models. Additional reasons for selecting random forest algorithm includes its ability to combines multiple weak learners to achieve high predictive accuracy and its reduced sensitivity to overfitting compared to other models.

7.1. Model Selection Rationale

Random Forest outperformed other models across most metrics, particularly in accuracy and precision. Its ensemble nature—combining multiple decision trees—helps reduce variance and improve generalization. This makes it well-suited for noisy compliance datasets where signal-to-noise ratios vary across features. However, when applied to high-dimensional compliance data, it has limitations in interpretability. Tools such as SHAP (SHapley Additive Explanations) can be beneficial for enhancing explainability in regulatory audits.

7.2. Implications for Organizational AI Governance

The proposed ML pipeline can be beneficial for managers or compliance officers in several ways. It enables automated risk by identifying anomalous LLM behavior based on operational metadata. It support audit process, as Confusion matrix can help justify decision during internal and external audits. Furthermore, it facilitates policy monitoring, as model aligns with NIST AI Risk Management Framework.

7.3. Link to Responsible AI and Bias Mitigation

The outcomes align with recent studies on responsible AI and bias mitigation. For instance, In [15], the authors outline the role of explainable AI in auditing; in [16], the report highlights privacy risk mitigation in LLMS; and in [17], author outlines the key indicators for LLM evaluation, that can be integrated in ML pipeline for predictive GRC scoring.

8. CONCLUSION

In today's evolving threat landscape, it is vital for organizations to proactively assess the risks posed by latest advancements in technology, such as Large Language Models (LLMs). This paper leverages machine learning to evaluate Governance, Risk, and Compliance (GRC) risks associated with LLM deployment. It utilizes Azure services to construct a dataset for predictive analysis. Key features in this dataset include response_time_ms, model_type, temperature, tokens_used, is_logged, data_sensitivity, compliance_flag, bias_score, and toxicity_score. These features are used to train a model that enables organizations to assess and mitigate GRC violations. With an accuracy of over 72 percent, this paper demonstrates how effective machine learning techniques can be in transforming GRC risk evaluation for LLMs through deterministic AI.

9. LIMITATION AND FUTURE WORK

While the proposed ml pipeline shows promise, the research can be enhanced by applying larger dataset to improve robustness and by validating model across platform such as OpenAI API and Google Gemini to ensure cross-platform applicability. Future work should focus on integrating explainable AI (XAI) tools such as SHAP to support transparency in regulatory audits, embedding ML-based GRC risk assessment into enterprise dashboard for real-time decision making and enhancing real-world AI governance frameworks.

REFERENCES

[1] Balasubramaniam, N., Kauppinen, M., Kujala, S., & Hiekkanen, K. (2020). Ethical guidelines for solving ethical issues and developing AI systems. *Product-Focused Software Process Improvement*, *331*, 331–346. https://link.springer.com/chapter/10.1007/978-3-030-64148-1_21

[2] Chazette, L., & Schneider, K. (2020). Explainability as a non-functional requirement: Challenges and recommendations. *Requirements Engineering*, *25*(4), 493–514. https://doi.org/10.1007/s00766-020-00333-1

[3] Chazette, L., Brunotte, W., & Speith, T. (2021). Exploring explainability: A definition, a model, and a knowledge catalogue. In *Proceedings of the International Requirements Engineering Conference* (pp. 197–208).

[4] Chazette, L., Karras, O., & Schneider, K. (2019). Do end-users want explanations? Analyzing the role of explainability as an emerging aspect of non-functional requirements. In *Proceedings of the International Requirements Engineering Conference* (pp. 223–233).

[5] European Commission. (2021). *Ethics guidelines for trustworthy AI*. https://ec.europa.eu/futurium/en/ai-alliance-consultation/guidelines

[6] European Data Protection Board. (2025). *AI privacy risks & mitigations: Large language models (LLMs)*. https://www.edpb.europa.eu/news/news/2025/ai-privacy-risks-mitigations-large-language-models-llms_en

[7] Govindasamy, P. (2025, October 24). Measuring quality, risk and cost of action through the economics of agentic AI. *Forbes Technology Council*. https://www.forbes.com/councils/forbestechcouncil/2025/10/24/measuring-quality-risk-and-cost-of-action-through-the-economics-of-agentic-ai/

[8] Horkoff, J. (2019). Non-functional requirements for machine learning: Challenges and new directions. In *Proceedings of the International Requirements Engineering Conference* (pp. 386–391).

[9] IEEE. (2021). *Ethically aligned design* (1st ed.). https://ethicsinaction.ieee.org/

[10] Kumar, V., Sinha, D., Das, A. K., Pandey, S. C., & Goswami, R. T. (2020). An integrated rule-based intrusion detection system: Analysis on UNSW-NB15 data set and the real time online dataset. *Cluster Computing*, *23*, 1397–1418. https://doi.org/10.1007/s10586-019-03008-x

[11] Lombrozo, T. (2012). Explanation and abductive inference. https://ifilnova.pt/wp-content/uploads/2021/12/Lombrozo-2012.pdf

[12] National Institute of Standards and Technology. (2023, January). *NIST risk management framework aims to improve trustworthiness of artificial intelligence*. https://www.nist.gov/news-events/news/2023/01/nist-risk-management-framework-aims-improve-trustworthiness-artificial

[13] Russom, P. (2011). *Big data analytics*. *TDWI Best Practices Report*, *Fourth Quarter*, 4–5, 22.

[14] Software and Information Industry Association (SIIA). (2017). *Ethical principles for artificial intelligence and data analytics* (pp. 1–25).

[15] Vaughan, J. W., & Wallach, H. (2021). A human-centered agenda for intelligible machine learning. In M. Pelillo & T. Scantamburlo (Eds.), *Machines we trust: Perspectives on dependable AI* (pp. 32–47). MIT Press

[16] Zhang, Q., Cheng, L., & Boutaba, R. (2010). Cloud computing: State-of-the-art and research challenges. *Journal of Internet Services and Applications*, *1*, 7. https://doi.org/10.1007/s13174-010-0007-6

[17] Zhong, C., & Goel, S. (2024). Transparent AI in auditing through explainable AI. *Current Issues in Auditing*, *18*(2), A1–A14. https://doi.org/10.2308/CIIA-2023-009