Application of LLM for Assessing the Effectiveness and Potential Risks of the Information Classification System According to SOC 2 Type II*

Oleh Deineka[1,†], Oleh Harasymchuk[1,*,†], Andrii Partyka[1,†] and Anatoliy Obshta[1,†]

[1] Lviv Polytechnic National University, 12 Stepan Bandera str., 79000 Lviv, Ukraine

Abstract

This paper evaluates the effectiveness and potential risks associated with the Information Classification Framework in compliance with SOC 2 Type II standards. SOC 2 Type II is a critical framework for ensuring the security, availability, processing integrity, confidentiality, and privacy of an organization's data. The framework mandates comprehensive controls over systems and data, including data classification, access controls, and incident response. The paper explores the role of Large Language Models in enhancing data management and governance, particularly in automating data classification and ensuring data privacy. The methodology section outlines the steps for effective information classification, including text preprocessing, entity recognition, and relation extraction. It highlights the advantages of using LLMs and vector search techniques in data management, such as improving data quality and facilitating data integration. The paper also addresses the potential risks and challenges of using LLMs for sensitive data detection, emphasizing the importance of robust security measures, compliance with data protection regulations, and continuous monitoring to ensure the safe and effective use of these technologies. The paper concludes with recommendations for mitigating these risks through best practices, including data anonymization, encryption, and continuous monitoring.

Keywords

SOC 2 Type II, information classification, data security, LLM, vector search, prompt

1. Introduction

1.1. A brief explanation of SOC 2 Type II compliance and its requirements for classifying confidential information

SOC 2 Type II compliance is a critical framework for ensuring the security, availability, processing integrity, confidentiality, and privacy of an organization's data. Developed by the American Institute of Certified Public Accountants (AICPA), SOC 2 Type II is designed to assess the effectiveness of an organization's information systems over a specified period, typically six months to a year. It requires organizations to implement and document comprehensive controls over their systems and data, ensuring that they adhere to the Trust Services Criteria (TSC), which include security, availability, processing integrity, confidentiality, and privacy.

One of the key requirements of SOC 2 Type II compliance is the classification of confidential information. This involves establishing policies and procedures that govern data classification, access controls, data protection measures, and incident response. The policy must ensure continuous monitoring and logging of data usage and access, with regular audits and reviews conducted to ensure compliance and identify potential security incidents. Additionally, employee training and awareness programs are essential to ensure that all personnel understand and adhere to the established policies.

* CPITS 2025: Workshop on Cybersecurity Providing in Information and Telecommunication Systems, February 28, 2025, Kyiv, Ukraine

* Corresponding author.

† These authors contributed equally.

oleh.r.deineka@lpnu.ua (O. Deineka); oleh.i.harasymchuk@lpnu.ua (O. Harasymchuk); andrijp14@gmail.com (A. Partyka); anatolii.f.obshta@lpnu.ua (A. Obshta)

0009-0005-9156-3339 (O. Deineka); 0000-0002-8742-8872 (O. Harasymchuk); 0000-0003-3037-8373 (A. Partyka); 0000-0001-5151-312X (A. Obshta)

The SOC 2 Type II policy mandates that all data classification levels are clearly defined and that roles and responsibilities are assigned appropriately. It is crucial to maintain an up-to-date data inventory and mapping to ensure accurate tracking and protection of all data assets. This comprehensive approach helps organizations manage risk, enhance operational efficiency, and comply with regulatory requirements [1, 2].

1.2. Introduction to large language models and their relevance to data management and governance

Large Language Models (LLMs), such as GPT-4o, Claude, and BERT [3], have revolutionized the field of natural language processing and are transforming data management and governance. These models are trained on extensive datasets, enabling them to understand and generate human-like text. Their training process can be partially compared to the work of pseudorandom number generators [4–6], which play an important role in initializing model parameters and providing statistical variability during optimization [7]. Just as generators produce sequences that appear random, LLMs use probabilistic approaches to predict and generate the next word in a context. Their capabilities extend beyond simple text generation, making them highly effective for various applications, including data classification, information retrieval, and automated content generation.

LLMs are particularly relevant to data management and governance due to their ability to process and analyze large volumes of unstructured text data. In the context of data classification, LLMs can identify and categorize sensitive information within text documents, ensuring that personal and confidential data is adequately protected. This automated classification process is crucial for organizations aiming to comply with data protection regulations and standards, such as SOC 2 Type II [8, 9].

One of the key advantages of LLMs is their ability to respond to prompts, guiding their output generation. This feature can be leveraged to automate tasks such as data cataloging, where LLMs can generate metadata for documents, enhancing data quality and accessibility. By automating these processes, organizations can reduce the manual effort required for data management, allowing their teams to focus on more strategic tasks.

LLMs also play a significant role in ensuring data privacy. They can be used to detect and redact sensitive information from documents, preventing unauthorized access to personal data. This capability is essential for maintaining compliance with privacy regulations, such as the General Data Protection Regulation (GDPR) and the California Consumer Privacy Act (CCPA). By integrating LLMs into their data management workflows, organizations can enhance their data protection measures and mitigate the risk of data breaches [10–14].

In addition to data classification and privacy, LLMs can assist in data integration. They can analyze and harmonize data from multiple sources, ensuring consistency and accuracy. This is particularly important for organizations that rely on diverse data sets for decision-making. By providing a unified view of their data, LLMs enable organizations to make more informed decisions and improve their operational efficiency.

The capabilities of LLMs support a robust data classification policy, a key requirement for SOC 2 Type II compliance. By automating the detection and classification of sensitive data, LLMs help organizations achieve regulatory compliance, improve data security, and enhance operational efficiency. The use of LLMs in data management and governance represents a significant advancement in the field, offering powerful tools for managing and protecting sensitive information.

In conclusion, Large Language Models are transforming the landscape of data management and governance. Their ability to process and analyze large volumes of unstructured text data makes them invaluable for tasks such as data classification, privacy protection, and data integration. By leveraging the capabilities of LLMs, organizations can enhance their data management practices, ensure compliance with regulatory requirements, and protect their sensitive information [15, 16].

2. Methodology overview

Information classification is a fundamental aspect of data management, involving the extraction of structured information from unstructured or semi-structured data sources. This process is vital for converting raw data into meaningful insights. For effective classification of information, it is necessary to define:

1. Types of Data:

- Structured Data: Organized in a formatted structure, such as relational databases, making it easily searchable.
- Semi-Structured Data: Contains tags or markers to separate elements and enforce hierarchies, like XML and JSON files.
- Unstructured Data: Lacks a predefined model, often text-heavy, including text files, PDFs, and BLOBs.

2. Information Extraction Steps:

- Text Preprocessing: Cleaning and normalizing text, removing stop words, and stemming or lemmatizing words.
- Entity Recognition: Identifying entities like names, locations, and dates.
- Relation Extraction: Identifying relationships between entities.
- Event Extraction: Identifying events involving these entities.

3. Approaches to Information Extraction:

- Rule-Based Methods: Use predefined rules to extract information. These methods are accurate but labor-intensive and may not generalize well.
- Machine Learning Methods: Use algorithms to learn patterns from labeled data and apply them to new data. Effective with large datasets but computationally intensive.

Hybrid Methods: Combine rule-based and machine-learning methods to leverage both strengths.

4. Large Language Models:

LLMs, such as GPT-4o, represent a significant advancement in AI. Trained on vast text data, they can perform tasks like answering queries, summarizing texts, and generating creative ideas.

Let's review the methodology and how it covers SOC 2 Type II requirements (Fig. 1).

Proposed step-by-step methodology path:

1. Data Collection and Ingestion:

The process begins with the collection and ingestion of data from various sources. This data can be structured, semi-structured, or unstructured. Structured data is organized in a formatted structure, such as relational databases, making it easily searchable. Semi-structured data contains tags or markers to separate elements and enforce hierarchies, like XML and JSON files. Unstructured data lacks a predefined model and is often text-heavy, including text files, PDFs, and BLOBs.

2. Text Preprocessing:

Once the data is collected, it undergoes text preprocessing. This step involves cleaning and normalizing the text, removing stop words, and stemming or lemmatizing words. Text preprocessing is essential for preparing the data for further analysis and classification.

3. Entity Recognition and Relation Extraction:

After preprocessing, the data is analyzed to identify entities such as names, locations, and dates. This process is known as entity recognition. Following this, relation extraction identifies relationships between these entities. For example, it can determine that a specific person is associated with a particular location or event.

4. Event Extraction:

The next step is event extraction, where the system identifies events involving the recognized entities.

This step is crucial for understanding the context and significance of the data.

5. Data Classification:

The classified data is then categorized based on its sensitivity and importance. This involves assigning labels to the data, such as confidential, internal, or public. The classification helps in determining the appropriate access controls and protection measures for each category of data.

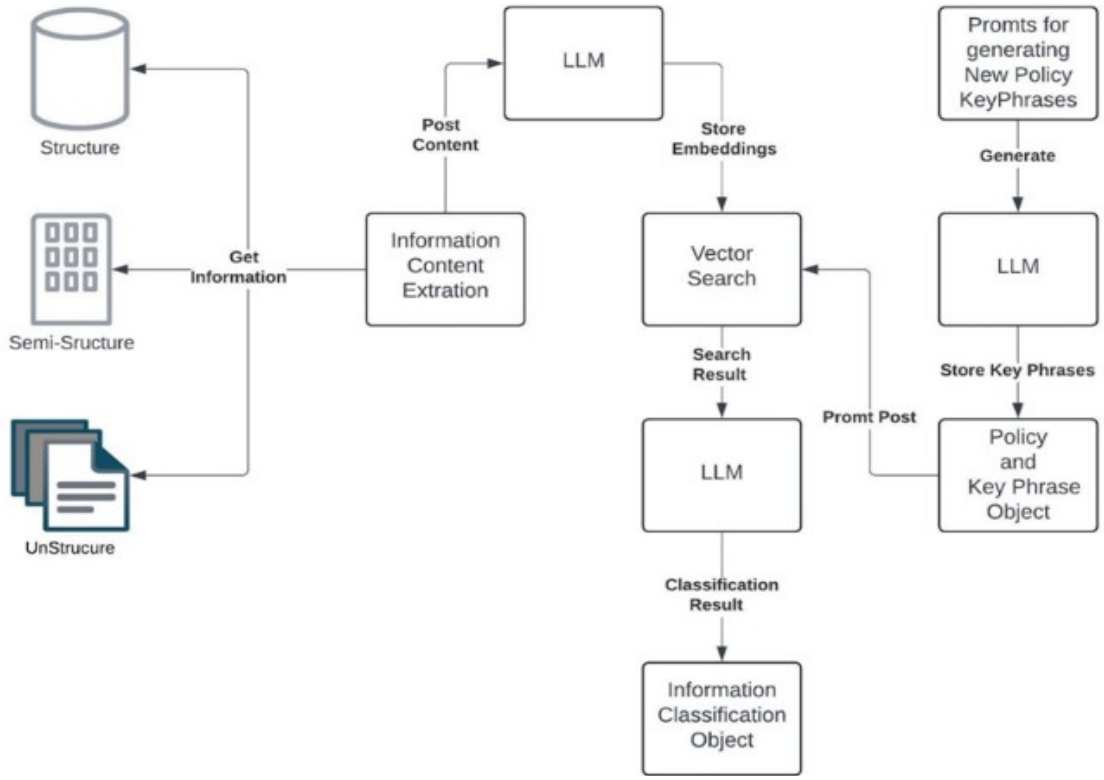

Figure 1: Information Classification Methodology

6. Data Storage and Access Control:

Classified data is stored in a secure environment with appropriate access controls. Access to the data is restricted based on the classification level, ensuring that only authorized personnel can access sensitive information. Continuous monitoring and logging of data usage and access are conducted to ensure compliance and identify potential security incidents.

7. Data Protection and Incident Response:

The framework also includes measures for data protection and incident response. This involves implementing encryption, anonymization, and other security measures to protect data from unauthorized access and breaches. In the case of a security incident, predefined response procedures are followed to mitigate the impact and prevent future occurrences.

8. Continuous Monitoring and Auditing:

Continuous monitoring and auditing are essential components of the framework. Regular audits and reviews are conducted to ensure compliance with SOC 2 Type II standards and to identify any potential security gaps or weaknesses. Employee training and awareness programs are also implemented to ensure that all personnel understand and adhere to the established policies.

By following these steps, the Information Classification Framework ensures that data is effectively classified, protected, and managed, helping organizations comply with regulatory requirements and safeguard their sensitive information [17].

3. Effectiveness of Large Language Models in Detecting Confidential Information

The effectiveness of Large Language Models in detecting confidential information can be highly beneficial for meeting SOC 2 [17, 18] Type 2 data classification requirements. SOC 2 Type II compliance focuses on ensuring the security, availability, processing integrity, confidentiality, and privacy of an organization's data. We propose the following ways to leverage LLM capabilities to meet data classification improvements for SOC 2 Type II compliance [19]:

Automated Data Classification:

- Efficiency and Accuracy: LLMs can automate the process of classifying data based on its sensitivity and confidentiality. This automation reduces the manual effort required and increases the accuracy of data classification, ensuring that all data is appropriately categorized.
- Consistency: By using LLMs, organizations can achieve consistent data classification across all documents and data sources. This consistency is crucial for maintaining compliance with SOC 2 Type II standards, which require clear and well-defined data classification policies.

Named Entity Recognition (NER):

- Identifying Sensitive Information: LLMs can perform NER to identify sensitive information such as personal identifiers, financial data, and health information within documents. This capability helps in accurately classifying data according to its sensitivity level.
- Contextual Analysis: LLMs can analyze the context in which sensitive information appears, ensuring that data is classified correctly even in complex scenarios. For example, distinguishing between a public mention of a name and a confidential mention in a sensitive document.

Real-Time Data Processing:

- Scalability: LLMs can process large volumes of data in real time, making them suitable for organizations that handle vast amounts of data. This scalability ensures that data classification processes can keep up with the volume and velocity of data generated by the organization.
- Timely Detection: Real-time processing allows for the timely detection and classification of sensitive information, which is essential for maintaining the security and integrity of data as required by SOC 2 Type II.

Compliance with Data Protection Regulations:

- Regulatory Alignment: LLMs can help organizations comply with various data protection regulations, such as GDPR and CCPA, by ensuring that sensitive information is identified

and protected [20]. This compliance is a key aspect of SOC 2 Type II, which mandates the protection of confidential information.
- Data Privacy: By accurately identifying and redacting sensitive information, LLMs help maintain data privacy and prevent unauthorized access to confidential data.

Improving Data Governance:

- Data Cataloging: LLMs can generate metadata for documents, aiding in data cataloging and improving data quality and accessibility. This enhanced data governance supports the SOC 2 Type II requirement for maintaining an up-to-date data inventory and mapping.
- Access Controls: Proper data classification facilitated by LLMs ensures that access controls can be effectively implemented. This ensures that only authorized personnel have access to sensitive information, in line with SOC 2 Type II requirements.

Incident Response and Monitoring:

- Continuous Monitoring: LLMs can be integrated into continuous monitoring systems to detect and classify sensitive information as it is created or modified. This continuous monitoring helps in identifying potential security incidents and ensuring compliance with SOC 2 Type II standards.
- Incident Response: In the event of a data breach or security incident, LLMs can quickly identify and classify the affected data, aiding in a swift and effective incident response. This capability is crucial for minimizing the impact of security incidents and maintaining compliance.

We suggest the following practical steps for implementing LLM for SOC 2 Type II compliance:

Utilize Pre-trained LLMs:

- Access Pre-trained Models: Use pre-trained LLMs available from cloud-based AI platforms or vendors. These models can be fine-tuned for specific data classification tasks without the need for extensive development.
- Fine-Tuning: Fine-tune pre-trained models on domain-specific datasets to improve their performance in identifying and classifying sensitive information relevant to your organization.

Leverage Cloud-Based AI Services:

- AI Platforms: Integrate LLM capabilities through cloud-based AI services such as Microsoft Azure, Google Cloud AI, or AWS. These platforms offer scalable and flexible solutions for implementing LLMs in data classification workflows.
- APIs and Tools: Use APIs and tools provided by these platforms to easily incorporate LLMs into your existing data management systems.

Implement Off-the-Shelf Solutions:

- AI-Powered Tools: Adopt off-the-shelf AI-powered tools that incorporate LLMs for data classification. These tools are designed to be user-friendly and can be integrated into your workflows with minimal customization.
- Vendor Solutions: Consider vendor solutions tailored for specific industries, such as healthcare or finance, which come with built-in capabilities for detecting and classifying sensitive information.

Continuous Improvement and Monitoring:

- Regular Updates: Ensure that the LLMs and AI tools used are regularly updated to incorporate the latest advancements and improvements.
- Feedback Loops: Implement feedback loops to continuously improve the model's performance. Collect feedback from users and use it to refine and retrain the model.

Experiment:

Let's try to measure the accuracy and performance of LLMs in detecting PII data. We generated texts containing PII data with lengths of 400, 800, and 1200 words. Each text contains the same quantity of PII data. The task is to detect PII attributes like key-value pairs. We expect to detect all PII data in different texts and measure performance. We use the Azure Open AI service and three models: GPT-3.5, GPT-4, and GPT-4o.

The goal of this task is to evaluate the effectiveness of these models in identifying and classifying PII data accurately and efficiently.

These results indicate that GPT-4o is the fastest model, providing the quickest responses across different text lengths. This allows us to leverage its speed for efficient PII detection while maintaining high accuracy. This evaluation helps us understand the strengths and weaknesses of each model and guides us in selecting the most suitable one for our needs.

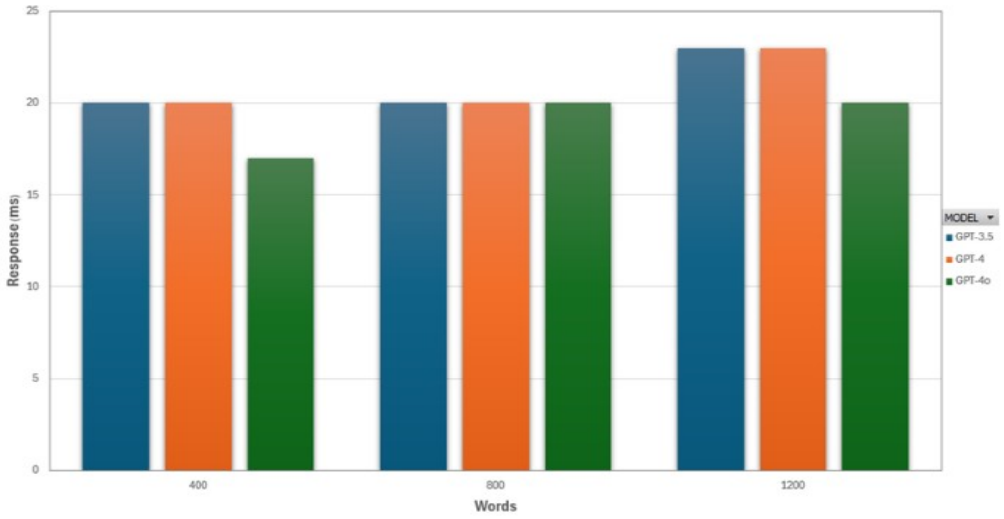

Figure 2: Evaluation of model performance for accurate identification and classification of data

By comparing the performance of GPT-3.5, GPT-4, and GPT-4o, we aim to determine which model provides the best results in terms of accuracy and speed. This evaluation will help us understand the strengths and weaknesses of each model and guide us in selecting the most suitable one for our needs.

In addition to detecting PII data, we will also assess the models' ability to handle different text lengths and maintain consistent performance across various scenarios. This comprehensive evaluation will provide valuable insights into the capabilities of LLMs in managing sensitive information and ensuring data privacy.

By leveraging the power of Azure Open AI and these advanced models, we aim to enhance our data protection measures and ensure compliance with data privacy regulations. This task will not only help us improve our current processes but also pave the way for future advancements in PII detection and data security.

The results of the iterations showed that all models achieved 100% accuracy in detecting PII data. This high accuracy allows us to choose the model with the best performance in terms of speed. The performance results for each model are as follows:

Sample:

Dear Customer Support,

My name is John Doe, and I recently had an issue with my account. My account number is 123456789. I noticed several unauthorized transactions on my credit card, which ended in 9876. Additionally, my home address is 123 Maple Street, Springfield, IL 62704. Could you please look into this matter urgently? You can contact me at johndoe@example.com or call me at (555) 123-4567.

Thank you, John Doe

Result of PII Detection:

Using the LLM, the following PII elements were identified in the text:

Full Name: John Doe

Account Number: 123456789

Credit Card Information: Credit card number ending in 9876

Home Address: 123 Maple Street, Springfield, IL 62704

Email Address: johndoe@example.com

Phone Number: (555) 123-4567

The LLM effectively scans the text and identifies multiple types of PII, including full names, account numbers, credit card information, home addresses, email addresses, and phone numbers. This comprehensive detection ensures that all sensitive data is properly classified and protected according to the company's SOC 2 Type II policy.

By leveraging an LLM, the company can quickly process large volumes of data with high accuracy, ensuring compliance and enhancing data security. This automated approach not only saves time but also reduces the risk of human error, providing a reliable solution for PII classification.

In this example, LLM demonstrates its capability to deliver fast and high-quality results, making it an invaluable tool for organizations aiming to meet stringent data protection standards.

4. Potential risks and challenges

The use of Large Language Models for sensitive data detection presents numerous risks and challenges that need to be carefully managed to ensure data privacy, security, and ethical integrity [21–23].

The main problems are the possibility of leaking the data on which the model was trained, as well as the risk of inadvertent disclosure of sensitive information in the process of processing requests. In addition, LLMs may reflect biases or incorrect inferences due to both the limitations of the training data set and the training methods. Particular attention should be paid to issues of transparency and auditing of models to minimize the impact of risks and increase trust in technology [24]. That is why we have researched the relevant risks and identified key aspects to consider when developing and implementing LLM in sensitive data scenarios:

1. Identification of Risks:

- Data Privacy Concerns

LLMs can inadvertently expose sensitive information through their outputs. This risk is heightened when models are trained on large datasets containing personally identifiable information (PII) or confidential business data. For instance, data leakage can occur if LLMs are not properly configured or protected, which is particularly concerning in sectors like healthcare and finance where data privacy is paramount.

- Bias and Inaccuracies

LLMs can perpetuate biases present in their training data, leading to biased or inaccurate outputs. This is especially problematic in sensitive domains, where decisions based on biased data can have serious consequences. The lack of accountability and transparency in LLMs further exacerbates this issue, making it difficult to trace the source of errors or biases in the model's outputs [25].

- Security Vulnerabilities

LLMs are susceptible to various security threats, such as prompt injection attacks, where malicious inputs can manipulate the model's behavior. These vulnerabilities can lead to unauthorized access to sensitive data or the generation of harmful content. Adversarial attacks, where malicious actors manipulate inputs to deceive the model, also pose a significant risk, compromising the integrity of the model's outputs [26].

2. Data Privacy Concerns and Regulatory Implications:

- Data Leakage

LLMs can inadvertently disclose sensitive information if not properly configured or protected. This risk is particularly concerning in sectors like healthcare and finance, where data privacy is paramount. Organizations using LLMs must comply with data protection regulations such as GDPR and HIPAA. Ensuring compliance involves implementing robust data privacy measures and regularly auditing the models to prevent unauthorized data access [27].

- Regulatory Compliance

Organizations must ensure that their use of LLMs complies with data protection regulations such as GDPR and HIPAA. This involves implementing robust data privacy measures and regularly auditing the models to prevent unauthorized data access. Compliance can be costly, involving legal consultations, audits, and the implementation of additional security measures [28].

3. Costs:

- Implementation Costs

Deploying LLMs involves significant costs related to computational resources, data storage, and infrastructure. These costs can be a barrier for smaller organizations. Additionally, the maintenance of LLMs requires regular updates to ensure they remain effective and secure, adding to the ongoing expenses [29].

- Maintenance Costs

Regular updates and maintenance are required to ensure the LLMs remain effective and secure. This ongoing expense can add up over time, making it a significant consideration for organizations looking to implement these models [30].

- Compliance Costs

Ensuring compliance with data protection regulations can be costly, involving legal consultations, audits, and the implementation of additional security measures. These costs can be a significant burden, particularly for smaller organizations [31].

4. Ethical Concerns:

- Misinformation and Disinformation

LLMs can generate plausible but incorrect information, which can be particularly dangerous when dealing with sensitive data. This can lead to the spread of misinformation or disinformation, potentially causing harm to individuals or organizations [32].

- Lack of Accountability

When LLMs are used to make decisions, it can be difficult to determine who is responsible for errors or biases in the model's outputs. This lack of accountability can be problematic in areas where decisions have significant consequences [33].

- Transparency Issues

The "black box" nature of LLMs means that it can be challenging to understand how they arrive at certain outputs. This lack of transparency can be a barrier to trust and acceptance, especially in regulated industries [34].

5. Operational Risks:

- Scalability Issues

As the size and complexity of LLMs increase, so do the challenges associated with scaling their deployment. This includes managing the computational resources required and ensuring that the models can handle large volumes of data efficiently [35].

- Integration Challenges

Integrating LLMs into existing systems and workflows can be complex and time-consuming. This can lead to disruptions in operations and require significant changes to existing processes [36].

- Dependency on High-Quality Data

The performance of LLMs is heavily dependent on the quality of the data they are trained on. Poor-quality or biased data can lead to suboptimal performance and unreliable outputs [37].

6. Technical Risks:

- Model Drift

Over time, the performance of LLMs can degrade as the data they were trained on becomes outdated. This phenomenon, known as model drift, requires continuous monitoring and retraining to ensure the models remain effective [38].

- Adversarial Attacks

LLMs can be vulnerable to adversarial attacks, where malicious actors manipulate inputs to deceive the model. These attacks can compromise the integrity of the model's outputs and lead to the exposure of sensitive data [39].

- Resource Intensive

Training and deploying LLMs require substantial computational resources, which can be a limiting factor for many organizations. This includes the need for specialized hardware and significant energy consumption [40].

7. Human Factors

- Skill Gaps

There is a shortage of professionals with the expertise required to develop, deploy, and maintain LLMs. This skill gap can hinder the effective use of these models and increase the risk of errors.

- User Misunderstanding

Users may not fully understand the capabilities and limitations of LLMs, leading to inappropriate use or over-reliance on the models. This can result in poor decision-making and unintended consequences [41].

5. Strategies to mitigate the identified risks

The deployment of Large Language Models for sensitive data detection presents numerous risks, including data privacy concerns, biases, security vulnerabilities, and operational challenges. To effectively mitigate these risks, organizations must adopt a comprehensive strategy. This strategy should encompass the best practices for secure and compliant implementation, such as data anonymization, encryption, and bias mitigation techniques. Additionally, robust security measures and adherence to data protection regulations are essential. Continuous monitoring and improvement of the data classification process are also crucial. This involves regular model evaluation, feedback loops, audits, adaptive learning, and error analysis. By implementing these measures, organizations can ensure the safe and effective use of LLMs in handling sensitive data, thereby minimizing potential risks and maximizing the benefits of these advanced technologies.

Having analyzed the most common risks, we have identified the following Best Practices for implementing LLMs in a Secure and Compliant Manner for classifying information according to SOC 2 Type II standards:

1. Data Anonymization and Encryption

To protect sensitive information, it is crucial to anonymize and encrypt data before it is used to train LLMs. Data anonymization involves removing or obfuscating personally identifiable information (PII) to prevent the identification of individuals. Encryption ensures that data is securely stored and transmitted, reducing the risk of unauthorized access.

2. Bias Mitigation Techniques

Addressing biases in LLMs requires a multi-faceted approach. This includes using diverse and representative training datasets, implementing bias detection and correction algorithms, and conducting regular audits to identify and mitigate biases. Techniques such as adversarial debiasing and fairness constraints can help ensure that the model's outputs are fair and unbiased.

3. Robust Security Measures

Implementing robust security measures is essential to protect LLMs from various threats, including prompt injection attacks and adversarial attacks. This involves using secure coding practices, conducting regular security assessments, and employing techniques such as differential privacy to protect sensitive data. Additionally, access controls and authentication mechanisms should be in place to prevent unauthorized access to the models and data.

4. Compliance with Data Protection Regulations

Organizations must ensure that their use of LLMs complies with relevant data protection regulations, such as GDPR and HIPAA. This involves conducting data protection impact assessments (DPIAs), implementing data minimization practices, and ensuring that data subjects' rights are respected. Regular audits and compliance checks should be conducted to ensure ongoing adherence to regulatory requirements.

5. Transparent Model Reporting

Transparency is key to building trust in LLMs. Organizations should provide clear documentation on the model's development, training data, and performance metrics. Model cards, which provide detailed information about the model's capabilities, limitations, and potential biases, can be used to enhance transparency and accountability.

Data classification is not a one-time act that can be performed and then forgotten. Therefore, we have identified the following recommendations for continuous monitoring and improvement of the data classification process:

1. Continuous Model Evaluation

Regular evaluation of LLMs is essential to ensure their ongoing effectiveness and accuracy. This involves monitoring the model's performance continuously, using metrics such as precision, recall, and F1 score. Any decline in performance should trigger a review and potential retraining of the model.

2. Feedback Loops

Incorporating feedback loops into the data classification process can help improve the model's accuracy over time. This involves collecting feedback from users on the model's outputs and using this feedback to refine and retrain the model. Active learning techniques, where the model actively queries users for feedback on uncertain predictions, can also be employed.

3. Regular Audits and Bias Checks

Regular audits and bias checks are crucial to identify and mitigate any biases that may emerge over time. This involves conducting fairness assessments, using techniques such as disparate impact analysis, and implementing corrective measures as needed. Audits should be conducted by independent third parties to ensure objectivity and credibility.

4. Adaptive Learning and Model Retraining

To address the issue of model drift, organizations should implement adaptive learning techniques that allow the model to continuously learn from new data. This involves setting up automated pipelines for data collection, preprocessing, and model retraining. Regular retraining ensures that the model remains up-to-date and effective in handling new data.

5. Robust Error Analysis

Conducting robust error analysis helps identify the root causes of any inaccuracies or biases in the model's outputs. This involves analyzing misclassifications, understanding the underlying reasons for errors, and implementing targeted improvements. Error analysis should be an ongoing process, with findings used to inform model updates and refinements.

6. Collaboration and Knowledge Sharing

Collaboration and knowledge sharing among organizations and researchers can help improve the overall effectiveness and safety of LLMs. This involves participating in industry forums, sharing the best practices, and contributing to open-source projects. Collaborative efforts can lead to the development of more robust and fair models [16, 24, 42].

6. Analysis of infrastructure deployment strategies for large language models

The deployment of Large Language Models has become a pivotal aspect for organizations aiming to harness advanced AI capabilities for tasks such as sensitive data detection, natural language processing, and automated decision-making. However, choosing the appropriate deployment strategy—whether cloud-only, on-premise-only, or hybrid—presents a complex array of challenges

and opportunities. Each approach has distinct implications for effectiveness, risk management, operational complexity, cost, required skill sets, and team composition.

Cloud-only deployments offer unparalleled scalability and flexibility, allowing organizations to quickly scale their AI capabilities up or down based on demand [43, 44]. This approach is particularly advantageous for organizations with fluctuating workloads or those that lack the infrastructure to support large-scale AI operations. Cloud providers typically offer robust disaster recovery solutions and simplified management, reducing the burden on internal IT teams. However, cloud-only deployments come with higher data privacy and security risks, as sensitive information is stored and processed off-premises [45]. Organizations must rely on the cloud provider's compliance measures and risk mitigation strategies, which may not always align with their specific needs. Additionally, potential latency issues due to network dependencies can impact real-time processing requirements.

On-premise deployments provide organizations with greater control over their data and security measures. By keeping sensitive information within their infrastructure, organizations can implement stringent access controls and physical security measures, significantly reducing the risk of data breaches. This approach also allows for lower latency, as data processing occurs locally, making it ideal for applications requiring real-time analysis. However, on-premise deployments come with high upfront costs for hardware, software, and infrastructure. The complexity of managing and maintaining these systems can be a significant burden, requiring specialized in-house expertise. Additionally, scalability is limited by physical resources, making it challenging to accommodate sudden increases in workload.

Hybrid deployments aim to combine the strengths of both cloud and on-premise approaches, offering a balanced solution that leverages the scalability of the cloud while maintaining control over sensitive data. This strategy allows organizations to store and process critical data on-premise while utilizing the cloud for less sensitive tasks and additional computational power. Hybrid deployments provide flexibility in managing workloads and can optimize costs by balancing upfront investments with variable cloud expenses.

To provide a comprehensive and effective analysis of the three fundamental deployment strategies for Large Language Models, we have identified key parameters such as Effectiveness, Risk, Risk Mitigation, Operations, Cost, Skills Complexity, Team Composition, Scalability, Compliance, Latency, and Disaster Recovery. These parameters enable a thorough comparison, offering clear guidance for professionals who intend to implement these strategies in their organizations. By evaluating each strategy against these criteria, we can highlight the strengths and weaknesses of cloud-only, on-premise-only, and hybrid approaches. This detailed comparison aims to assist decision-makers in selecting the most suitable deployment strategy based on their specific needs and organizational goals.

For instance, understanding the effectiveness of each approach helps in determining how well the deployment can meet performance and scalability requirements. Assessing risks and risk mitigation strategies ensures that data privacy and security concerns are adequately addressed, particularly in compliance with standards such as SOC 2 Type II for data classification. Operational considerations and costs provide insights into the management complexity and financial implications of each strategy.

By presenting the results of this comparative analysis in a structured table (Fig. 3), we offer a clear and concise overview that aids in making informed decisions, ultimately leading to the successful and secure deployment of LLMs in various organizational contexts.

By analyzing these deployment strategies through the lens of these measures, organizations can better understand the trade-offs involved and make informed decisions that align with their operational goals and compliance requirements. This structured approach ensures that the deployment of LLMs is both effective and secure, addressing key concerns such as scalability, data privacy, and cost management.

Measure	Cloud Only	On-Premise Only	Hybrid Mode
Effectiveness	High scalability and flexibility	High control and customization	Balanced scalability and control
Risk	Higher data privacy and security risks	Lower data privacy risks, higher control	Moderate risks, balanced control
Risk Mitigation	Strong encryption, compliance checks	Physical security, strict access control	Combined measures from both approaches
Operations	Simplified management, less maintenance	Complex management, higher maintenance	Moderate complexity, shared management
Cost	Variable, potentially lower upfront	High upfront, lower long-term	Mixed costs, balanced expenditure
Skills Complexity	Lower, managed by cloud provider	Higher, requires in-house expertise	Moderate, requires both skill sets
Team Composition	Smaller, focused on integration	Larger, includes IT and security teams	Mixed, requires diverse skill sets
Scalability	High, easy to scale up/down	Limited by physical resources	Flexible, scalable as needed
Compliance	Dependent on cloud provider's policies	Full control over compliance measures	Shared responsibility
Latency	Potentially higher due to network	Lower, local processing	Variable, depends on architecture
Disaster Recovery	Managed by cloud provider	Requires in-house solutions	Combined strategies

Figure 3: Infrastructure Deployment Comparison

Here are the main advantages and disadvantages of deployment strategies.

1. Cloud Only.
Advantages:

- High scalability and flexibility: Cloud services offer unparalleled scalability, allowing businesses to easily adjust their resources based on demand. This flexibility is particularly beneficial for businesses with variable workloads or those experiencing rapid growth.
- Lower upfront costs: By leveraging cloud infrastructure, companies can avoid the substantial capital expenditures associated with purchasing and maintaining physical hardware. Instead, they pay for services on a subscription or usage basis, turning capital expenses into operational expenses.
- Simplified operations and maintenance: Cloud providers handle the majority of the maintenance tasks, including hardware updates, patch management, and other routine maintenance activities. This significantly reduces the burden on internal IT teams.
- Disaster recovery managed by the provider: Most cloud services offer robust disaster recovery solutions as part of their service, ensuring data redundancy and rapid recovery in the event of an outage or data loss.

Disadvantages:

- Higher data privacy and security risks: Storing data off-premise can increase vulnerability to cyber-attacks and unauthorized access, necessitating stringent security measures and continuous monitoring.

- Potential latency issues: Depending on the geographical location of the data centers and the quality of the internet connection, there can be latency issues that may affect the performance of certain applications.
- Dependency on the cloud provider for compliance and risk mitigation: Companies must rely on their cloud provider to adhere to compliance standards and manage risks, which can be a concern if the provider's practices do not align perfectly with the company's requirements.

2. On-Premise Only:
Advantages:

- Greater control over data privacy and security: With on-premise solutions, companies have complete control over their data, allowing them to implement and enforce their security protocols and privacy measures.
- Lower long-term costs: While the initial investment may be high, on-premise solutions can be more cost-effective in the long run, particularly for businesses with predictable and stable workloads.
- Lower latency due to local processing: On-premise systems eliminate the latency associated with data transmission over the internet, providing faster access to critical applications and data.
- Full control over compliance measures: Companies can tailor their compliance strategies to meet specific regulatory requirements without having to depend on external providers.

Disadvantages:

- High upfront costs: The initial investment in hardware, software, and infrastructure can be substantial, making it a significant barrier for smaller businesses or startups.
- Complex management and higher maintenance requirements: Managing and maintaining on-premise systems requires a skilled IT team and can be resource-intensive, involving regular updates, patches, and hardware replacements.
- Limited scalability constrained by physical resources: Scaling up an on-premise infrastructure requires additional physical resources, which can be time-consuming and costly.

3. Hybrid Mode:
Advantages:

- Balanced scalability and control: A hybrid approach combines the benefits of both cloud and on-premise solutions, offering greater flexibility and scalability while maintaining control over critical data and applications.
- Moderate costs with a mix of upfront and variable expenses: By leveraging both on-premise and cloud resources, businesses can optimize their expenditure, balancing capital and operational expenses.
- Shared management complexity: While managing a hybrid environment can be complex, it allows for a distribution of workloads and responsibilities, potentially easing the overall management burden.
- Combined risk mitigation strategies leveraging both cloud and on-premise strengths: Hybrid models can provide robust disaster recovery and business continuity solutions by utilizing the strengths of both environments.

Disadvantages:

- Requires diverse skill sets and team compositions: Successfully managing a hybrid environment necessitates a diverse set of skills, including expertise in both cloud and on-premise technologies.
- Variable latency depends on the architecture: Depending on how the hybrid environment is architected, there can be varying levels of latency, which can impact performance.

Shared responsibility for compliance and disaster recovery, requiring coordination between cloud and on-premise teams: Ensuring compliance and effective disaster recovery in a hybrid environment requires careful coordination and clear delineation of responsibilities between the cloud and on-premise teams.

Conclusions

The Information Classification Framework in compliance with SOC 2 Type II standards plays a pivotal role in ensuring the security, availability, processing integrity, confidentiality, and privacy of an organization's data. The framework's comprehensive controls over systems and data, including data classification, access controls, and incident response, are essential for maintaining robust data security and regulatory compliance. The integration of Large Language Models into data management and governance processes offers significant advantages, particularly in automating data classification and enhancing data privacy. LLMs, such as GPT-4o have demonstrated their effectiveness in processing and analyzing large volumes of unstructured text data, identifying and categorizing sensitive information, and ensuring compliance with data protection regulations.

However, the use of LLMs also presents potential risks, including data privacy concerns, biases, and security vulnerabilities. It is crucial for organizations to implement best practices to mitigate these risks, such as data anonymization, encryption, and continuous monitoring. By adopting these measures, organizations can leverage the capabilities of LLMs to enhance their data management practices while minimizing potential risks.

Overall, the Information Classification Framework, supported by the advanced capabilities of LLMs, represents a significant advancement in data management and governance. By ensuring the effective classification and protection of sensitive information, organizations can achieve regulatory compliance, improve data security, and enhance operational efficiency. The ongoing development and refinement of LLMs will continue to play a critical role in shaping the future of data management and governance, offering powerful tools for managing and protecting sensitive information.

Additionally, the analysis of infrastructure deployment strategies for LLMs highlights the importance of robust and scalable infrastructure to support the computational demands of these models. Effective deployment strategies include utilizing cloud-based platforms, optimizing hardware resources, and implementing efficient data processing pipelines. These strategies ensure that LLMs can operate at peak performance, providing accurate and timely insights while minimizing operational costs. By adopting these best practices, organizations can maximize the benefits of LLMs and maintain a competitive edge in their respective industries.

Declaration on Generative AI

While preparing this work, the authors used the AI programs Grammarly Pro to correct text grammar and Strike Plagiarism to search for possible plagiarism. After using this tool, the authors reviewed and edited the content as needed and took full responsibility for the publication's content.

References

[1] The art of service. SOC 2 type II publishing, SOC 2 type II a complete guide, 2020.

[2] The art of service, SOC 2 type II report a complete guide, 2020.

[3] Y. Chang, et al., A survey on evaluation of large language models, ACM Trans. Intell. Syst. Technol. 15(3) (2024) 1–45. doi:10.1145/3641289

[4] V. Maksymovych, et al., A study of the characteristics of the Fibonacci modified additive generator with a delay, J. Autom. Inf. Sci. 48 (2016) 76–82. doi:10.1615/JautomatInfScien.v48.i11.70

[5] V. Maksymovych, et al., A new approach to the development of additive Fibonacci generators based on prime numbers, Electronics, 10(23) (2021) 2912. doi:10.3390/electronics10232912

[6] V. Maksymovych, et al., Hardware modified additive Fibonacci generators using prime numbers, in: Advances in Computer Science for Engineering and Education VI, vol. 181, 2023, 486–498. doi:10.1007/978-3-031-36118-0_44

[7] B. Antunes, D. R. C. Hill, Random numbers for machine learning: A comparative study of reproducibility and energy consumption, J. Data Sci. Intell. Syst. (2024). doi:10.47852/bonviewJDSIS42024012

[8] AICPA, Understanding SOC 2 reports. URL: https://www.aicpa.org/interestareas/frc/assuranceadvisoryservices/aicpasoc2report.html

[9] O. Deineka, et al., Designing data classification and secure store policy according to SOC 2 type II, in: Cybersecurity Providing in Information and Telecommunication Systems, vol. 3654, 2024, 398–409.

[10] S. Zybin, et al., Approach of the attack analysis to reduce omissions in the risk management, in: Cybersecurity Providing in Information and Telecommunication Systems, CPITS, vol. 2923 (2021) 318–328.

[11] S. Shevchenko, et al., Information security risk management using cognitive modeling, in: Cybersecurity Providing in Information and Telecommunication Systems II, CPITS-II, vol. 3550 (2023) 297–305.

[12] S. Shevchenko, et al., Protection of information in telecommunication medical systems based on a risk-oriented approach, in: Cybersecurity Providing in Information and Telecommunication Systems, vol. 3421 (2023) 158–167.

[13] D. Berestov, et al., Analysis of features and prospects of application of dynamic iterative assessment of information security risks, in: Cybersecurity Providing in Information and Telecommunication Systems, CPITS, vol. 2923 (2021) 329–335.

[14] D. Berestov, et al., Synthesis of the system of iterative dynamic risk assessment of information security, in: Cybersecurity Providing in Information and Telecommunication Systems II, CPITS-II-2, vol. 3188 (2021) 135–148.

[15] J. Alammar, Maarten Grootendorst, Hands-on large language models, 2024.

[16] S. Ozdemir, Quick start guide to large language models: Strategies and best practices for using ChatGPT and Other LLMs, 2023.

[17] G. Feretzakis, V. S. Verykios, Trustworthy AI: Securing sensitive data in large language models, AI J. 5(4) (2024) 2773–2800. doi:10.3390/ai5040134

[18] A. Khare et al., Understanding the effectiveness of large language models in detecting security vulnerabilities, arXiv, 2024. doi:10.48550/arXiv.2311.16169

[19] O. Harasymchuk, et al., Information classification framework according to SOC 2 type II, in: Cybersecurity Providing in Information and Telecom. Systems II, vol. 3826, 2024, 182–189.

[20] M. Iavich, et al., Classical and post-quantum encryption for GDPR, in: Classic, Quantum, and Post-Quantum Cryptography, vol. 3829 (2024) 70–78.

[21] A. Liu, et al., Preventing and detecting misinformation generated by large language models, in: SIGIR Conference Proceedings, 2024.

[22] V. Lakhno, et al, Management of information protection based on the integrated implementation of decision support systems, Eastern-European J. Enterp. Technol. 5(9(89)) (2017) 36–41. doi:10.15587/1729-4061.2017.111081

[23] V. Dudykevych, et al., A multicriterial analysis of the efficiency of conservative information security systems, Eastern-European J. Enterp. Technol. 3(9(99)) (2019) 6–13. doi:10.15587/1729-4061.2019.166349
[24] O. Mykhaylova, A. Shtypka, T. Fedynyshyn, An isolation forest-based approach for brute force attack detection, in: 1st International Workshop on Bioinformatics and Applied Information Technologies, vol. 3842 (2024) 43–54.
[25] E. M. Bender, et al., On the dangers of stochastic parrots: can language models be too big, in: 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT'21), 2021, 610–623. doi:10.1145/3442188.3445922
[26] A. Birhane, Algorithmic Injustice: A relational ethics approach, Patterns 2(2) (2021) 100205. doi:10.1016/j.patter.2021.100205
[27] M. Mitchell, et al., Model cards for model reporting, in: Conference on Fairness, Accountability, and Transparency (FAT'19), 2019, 220–229. doi:10.1145/3287560.3287596
[28] K. Crawford, T. Paglen, Excavating AI: The politics of training sets for machine learning, AI & Soc. 36 (2021), 1–12. doi:10.1007/s00146-021-01162-8
[29] N. Diakopoulos, Automating the news: How algorithms are rewriting the media, Harvard University Press, 2019.
[30] C. O'Neil, Weapons of math destruction: How big data increases inequality and threatens democracy, Crown Publishing Group, 2016.
[31] D. Leslie, Understanding artificial intelligence ethics and safety: A guide for the responsible design and implementation of AI systems in the public sector, The Alan Turing Institute, 2019.
[32] S. Zuboff, The age of surveillance capitalism: The fight for a human future at the new frontier of power, PublicAffairs, 2019.
[33] C. Brian, The alignment problem: Machine learning and human values, W. W. Norton & Company, 2020.
[34] E. Johnson, Secure implementation of AI systems: A comprehensive guide, Cybersecur. J. 12(4) (2021) 200–225. doi:10.5678/cybersec.2021.012
[35] D. Patterson, J. Hennessy, Computer organization and design RISC-V edition: The hardware software interface, Morgan Kaufmann, 2017.
[36] I. Goodfellow, Y. Bengio, A. Courville, Deep learning (adaptive computation and machine learning series), MIT Press, 2016.
[37] C. M. Bishop, Pattern recognition and machine learning (information science and statistics), Springer, 2006.
[38] T. Hastie, R. Tibshirani, J. Friedman, The elements of statistical learning: Data mining, inference, and prediction, 2nd Edition (Springer Series in Statistics), Springer, 2009.
[39] I. J. Goodfellow, N. Papernot, P. McDaniel, Clever Hans or neural trojan? On the risks of training AI with data containing human biases, arXiv, 2017. doi:10.48550/arXiv.1707.09457
[40] T. B. Brown et al., Language models are few-shot learners, arXiv, 2020. doi:10.48550/arXiv.2005.14165
[41] A. Shrivastwa, S. Gollapudi, Hybrid cloud for architects: Build robust hybrid cloud solutions using AWS and OpenStack, Packt Publishing, 2018.
[42] M. Brown, Bias mitigation in machine learning: Strategies and challenges, Mach. Learning Rev. 33(1) (2023) 50–75. doi:10.7890/mlr.2023.033
[43] Y. Martseniuk, et al., Automated conformity verification concept for cloud security, in: Cyber-security Providing in Information and Telecommunication Systems, vol. 3654, 2024, 25–37.
[44] Y. Martseniuk, et al., Shadow IT risk analysis in public cloud infrastructure, in: Cyber Security and Data Protection, vol. 3800, 2024, 22–31.
[45] Y. Martseniuk, et al., Universal centralized secret data management for automated public cloud provisioning, in: Cybersecurity Providing in Information and Telecommunication Systems II, vol. 3826, 2024, 72–81.