Information classification framework according to SOC 2 Type II

Oleh Deineka[1,†], Oleh Harasymchuk[1,†], Andrii Partyka[1,†] and Valerii Kozachok[2,*,†]

[1] Lviv Polytechnic National University, 12 Stepana Bandery str., 79000 Lviv, Ukraine
[2] Borys Grinchenko Kyiv Metropolitan University, 18/2 Bulvarno-Kudryavska str., 04053 Kyiv, Ukraine

Abstract

Large Language Models (LLMs) like GPT-3 and BERT, trained on extensive text data, are transforming data management and governance, areas crucial for SOC 2 Type II compliance. LLMs respond to prompts, guiding their output generation, and can automate tasks like data cataloging, enhancing data quality, ensuring data privacy, and assisting in data integration. These capabilities can support a robust data classification policy, a key requirement for SOC 2 Type II. Vector search, another important method in data management, finds similar items to a given item by representing them as vectors in a high-dimensional space. It offers high accuracy, scalability, and flexibility, supporting efficient data classification. Embeddings, which convert categorical data into a form that can be input into a model, play a key role in vector search and LLMs. Prompt engineering, the crafting of effective prompts, is crucial for guiding LLMs' output, and further enhancing data management and governance practices.

Keywords

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

1. SOC 2 Type II

In today's digital age, the exponential growth of information assets, a significant portion of which is critical, is a defining characteristic. The sheer volume of this information necessitates its classification based on various parameters and features, secure storage and transmission, and protection against unauthorized access. The frequency of potential attacks on information resources is on the rise [1–3]. To counteract these threats, cybersecurity experts are continually developing new standards, strategies, and techniques, as well as advancing infrastructure [4–9]. A key focus is the creation and research of standards for secure data storage [10–14]. These standards provide insight into how an organization controls data access and ensures its security and confidentiality.

The standards and requirements for data storage can differ for organizations based on factors such as geographical location, industry, sensitivity of the information, and more. Specific organizations may have unique standards and requirements based on their needs and legal obligations.

Most organizations formulate their security policies based on international standards, often with the involvement of external auditing firms that certify standard compliance. However, professionals dealing with secure storage of large data volumes still face numerous challenges, including data integrity, confidentiality, and accessibility. Ensuring the information remains unchanged from creation through storage and retrieval can be a complex task. Additionally, professionals must ensure confidentiality, allowing only authorized individuals to access the data and guarantee data accessibility when needed, a task that becomes increasingly challenging with growing data volumes.

Despite the existence of various effective strategies, methods, and systems for organizing big data storage, certain problems persist. One significant issue is the difficulty of searching for required information in unstructured data.

ISO 27001 [11] is a standard aimed at ensuring the proper management of a company's digital assets, including financial information, intellectual property, employee data, and trusted third-party information. Meanwhile, SOC 2 certification [10] is more recognized and typically preferred by American and Canadian companies.

SOC is divided into SOC 1, SOC 2, and SOC 3. The first pertains exclusively to financial control, and the third is primarily used for marketing purposes, allowing SaaS providers to focus solely on SOC 2.

The Service and Organization Controls 2 standard, developed by the American Institute of Certified Public Accountants using the Trust Services Criteria reliability criteria, provides an independent evaluation of risk management control procedures in IT companies that provide services to users. The standard emphasizes data privacy and confidentiality, making it a choice for giants

CPITS-II 2024: Workshop on Cybersecurity Providing in Information and Telecommunication Systems II, October 26, 2024, Kyiv, Ukraine
* Corresponding author.
† These authors contributed equally.
oleh.r.deineka@lpnu.ua (O. Deineka);
garasymchuk@ukr.net (O. Harasymchuk);
andrijp14@gmail.com (A. Partyka);
v.kozachok@kubg.edu.ua (V. Kozachok)

0009-0005-9156-3339 (O. Deineka);
0000-0002-8742-8872 (O. Harasymchuk);
0000-0003-3037-8373 (A. Partyka);
0000-0003-0072-2567 (V. Kozachok)

like Google and Amazon, for whom high-security levels and transparent data processing processes are crucial. External auditors are engaged for certification. Their role is to examine the implemented practices, verify the company's adherence to its procedures, and monitor changes in processes.

SOC 2 Type II is a significant certification in the data security and compliance landscape. It serves as an attestation by an independent auditor that a service organization's systems are not only designed to meet the Trust Services Criteria but also operate effectively over time. The Trust Services Criteria cover several critical areas: security, availability, processing integrity, confidentiality, and privacy.

The value of SOC 2 Type II lies in its ability to foster trust with clients and stakeholders. By demonstrating a commitment to stringent data management practices, companies can assure clients that their sensitive data is managed responsibly. This is particularly important in sectors where data privacy and security are crucial, such as financial services, healthcare, and cloud computing.

Furthermore, the audit process for SOC 2 Type II helps organizations identify and mitigate potential security risks, ensuring they maintain a robust security posture. This proactive approach to risk management is vital in an era where cyber threats are continually evolving, and data breaches can have devastating consequences. Hence, there is a constant search for new strategies and methods to ensure reliable data storage and user and device authentication where this data is stored [15–18].

In an increasingly regulated environment, SOC 2 Type II compliance can also support adherence to legal and regulatory requirements, helping organizations avoid expensive penalties and legal issues associated with non-compliance.

From a business perspective, SOC 2 Type II compliance can serve as a competitive differentiator. It signals to the market that an organization is a reliable and secure partner, which can be instrumental in winning new business and retaining existing customers [19].

The outcome of implementing SOC 2 is a report based on the AICPA Attestation Standards, section 101, Attest Engagement.

2. SOC 2 Type II information classification policy

2.1. Requirements

According to the document: SOC 2 Type II, while not prescribing specific data classification policies, mandates that organizations effectively manage and safeguard the confidentiality, privacy, and security of information in line with the Trust Services Criteria (TSC).

A Data Classification Policy is essential in meeting these criteria, especially the Security criterion, which is common to all SOC 2 audits. A SOC 2 audit evaluates the effectiveness of an organization's processes and systems based on the Trust Service Criteria and checks compliance with information security standards and regulations, including Common Criteria standards. To support SOC 2 Type II compliance, a Data Classification Policy should address several general requirements, including the identification of data types. The policy should define the types of data the organization handles, including sensitive data subject to SOC 2 considerations, such as personal identifiable information (PII), business confidential data, and intellectual property. The policy must also establish clear classification levels that reflect the sensitivity of the data, with common levels including Public, Internal Use Only, Confidential, and Highly Confidential. Additionally, the policy should define roles and responsibilities for data classification, including data owners, custodians, and users, and outline their responsibilities in maintaining data classification.

A Data Classification Policy for SOC 2 Type II compliance should specify handling requirements for each classification level, including storage, transmission, access controls, encryption standards, and end-of-life procedures. The policy should also provide guidelines on how data should be labeled or marked according to its classification to ensure that it is easily identifiable and handled appropriately. Access controls must be addressed, ensuring that access to data is based on the principle of least privilege and that only authorized individuals can access sensitive data. The policy should outline data retention periods and secure disposal methods for each classification level, ensuring data is not kept longer than necessary and is disposed of securely. Regular training and awareness programs for employees should be mandated to understand the importance of data classification and their role in it. The policy should include provisions for regular auditing and monitoring to ensure that classification controls are effective and being followed. The policy should be linked to an incident response plan that addresses potential data breaches or loss, with procedures tailored to the classification level of the data involved. The policy should specify intervals for reviewing and updating data classification procedures to ensure they remain relevant and effective as the organization evolves, data volumes increase, and new threats emerge. According to the document: If data is shared with or handled by third-party vendors, the data classification policy must extend to these vendors, often requiring them to adhere to similar or compatible classification and handling standards. To ensure alignment with SOC 2 Type II requirements, developing a Data Classification Policy usually demands a comprehensive understanding of the AICPA's TSC and the unique data protection requirements of the organization. Engaging with seasoned compliance experts or auditors who can give tailored advice and oversee compliance with the standard's stipulations is highly recommended. The AICPA's guidance and frameworks such as ISO 27001, when consulted and utilized, can offer invaluable inputs for the creation and sustenance of a strong data classification policy. It is crucial to identify and categorize data based on its sensitivity, importance, and regulatory mandates. Moreover, regular reviews and updates of the policy should be conducted to ensure its efficiency and continued compliance with SOC 2 Type II requirements [20–24].

2.2. Design

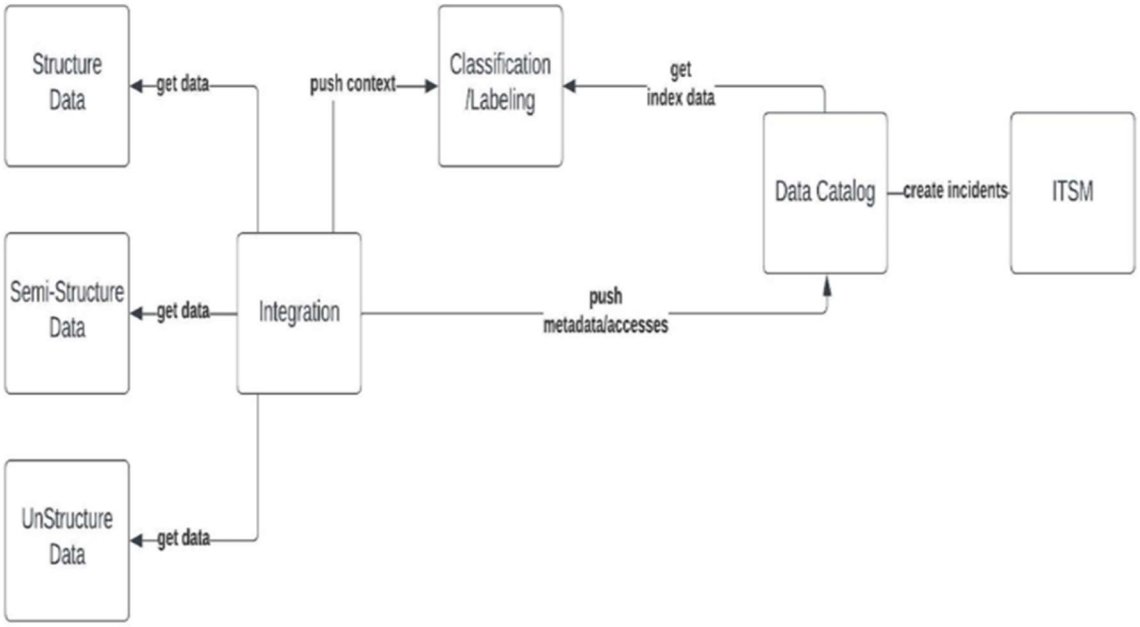

Figure 1: Data flow diagram

So, we offer a Data Flow Diagram. Time Creating a Data Flow Diagram necessitates an initial comprehensive grasp of the various data types that your company possesses. Typically, data can be divided into three main categories: structured, semi-structured, and unstructured.

Data that is organized in a prearranged manner, such as the data stored in a relational database, is referred to as structured data. Its consistent format makes structured data easy to search, analyze, and manipulate. On the other hand, semi-structured data has a certain level of organization but lacks a strict format. XML and JSON files, which house data in a hierarchical format without a fixed schema, are examples of semi-structured data.

Unstructured data is characterized by its lack of inherent structure or organization. This category includes text documents, images, and videos. The inconsistent format of unstructured data can pose challenges when it comes to searching, analyzing, and manipulating it [25–26].

After the identification of the company's data types, the subsequent phase involves gaining an understanding of the metadata linked to that data. Metadata is essentially data that offers information about other data. For example, the metadata linked to a text document could include details like the author, the date of creation, and the file size. A deep understanding of the metadata associated with your data can facilitate better organization, management, and analysis of your data [27].

The process of creating a Data Flow Diagram continues with the utilization of integration tools to manage and store your data, once you have identified the types of data your company owns and the metadata associated with that data. Integration tools facilitate the extraction of data from various sources, its transformation into a common format, and its loading into a data store. This process, known as Extract, Transform, Load (ETL), consolidates your data into a single location, simplifying its management and analysis [28–32].

Following the extraction, transformation, and loading of your data into a data store, the subsequent phase involves creating a data model. A data model is a visual depiction of the relationships between different data elements. It provides a structure for organizing and structuring your data and can assist in identifying patterns and trends within your data [33].

Once a data model has been created, the next step involves classifying your data and linking it to the associated metadata. This involves assigning a sensitivity level to your data, based on its importance and the potential impact if it were to be lost or stolen. After your data has been classified, it can be linked to the associated metadata, providing additional context and information about the data [34].

The final phase in creating a Data Flow Diagram involves creating an application that enables you to visualize and manage your data. This application should offer a user-friendly interface for accessing, analyzing, and manipulating your data. It should also incorporate logic for managing access, requests, and incidents, and should be integrated with your ITSM system to ensure that data is handled according to your company's policies and procedures [35].

This solution offers numerous advantages over traditional product-based offerings from various companies. One of the primary benefits is the flexibility to choose the hosting environment that best suits your needs, whether on-premise or cloud-based. This allows you to align the solution with your operational requirements and infrastructure capabilities.

Additionally, you have the liberty to select the technology stack that best fits your project. This means that you're not confined to a predetermined set of technologies but can customize the solution to leverage the most relevant and efficient tools for your specific needs.

In terms of team composition, you have the flexibility to assemble a team that is uniquely suited to the project at hand. This flexibility ensures that the right expertise and skills are applied to deliver the best possible outcomes.

Another advantage is the flexibility in budgeting. Unlike vendor-specific solutions that may come with fixed licensing costs, the budget for this solution can be adjusted according to your financial capacity and project requirements. This can result in significant cost savings without compromising on quality or performance.

Lastly, this solution offers robust change and feature management capabilities. This means that it can easily adapt to evolving business needs, with the ability to incorporate new features and make necessary changes in a timely and efficient manner. This flexibility ensures the solution remains relevant and continues to deliver value over time [13].

3. Information classification

3.1. Overview

Information classification is a critical process in data management that involves categorizing data based on its sensitivity, importance, and regulatory requirements. This process is essential for organizations to effectively protect their data and comply with various legal, regulatory, and contractual obligations.

The primary goal of information classification is to facilitate appropriate levels of protection for different types of data. By classifying data such as public, internal, confidential, or highly confidential, organizations can apply suitable security measures to each category, ensuring that sensitive and critical data receives the highest level of protection.

Information classification is not a one-time activity but a continuous process that needs to be integrated into the organization's data lifecycle. It involves identifying the types of data the organization handles, defining classification levels, assigning responsibilities for data classification, and implementing procedures for handling, storing, and disposing of data based on its classification.

In addition to enhancing data security, information classification also aids in risk management, regulatory compliance, and resource allocation. It helps organizations understand where their most sensitive and valuable data resides, who has access to it, and how it is being protected, enabling them to identify and mitigate potential risks. It also supports compliance with regulations such as GDPR, HIPAA, and SOC 2, which require organizations to implement appropriate safeguards for sensitive data. Furthermore, by identifying less sensitive data that requires lower levels of protection, organizations can optimize their use of resources.

In today's data-driven world, where vast volumes of data are generated and processed every day, information classification has become more important than ever. It is a fundamental step in ensuring that all data is given the appropriate level of protection and handled responsibly throughout its lifecycle.

3.2. Importance

The importance of Information Classification in the context of SOC 2 Type II compliance cannot be overstated. It serves as the foundation for data security and privacy controls, helping organizations identify and protect their most sensitive data.

Firstly, Information Classification helps in identifying the types of data an organization handles, including sensitive data subject to SOC 2 considerations, such as personally identifiable information (PII), confidential business data, and intellectual property. This identification is the first step towards implementing appropriate security measures.

Secondly, Information Classification aids in establishing clear classification levels that reflect the sensitivity of the data. These levels, which commonly include Public, Internal Use Only, Confidential, and Highly Confidential, guide the implementation of access controls, encryption standards, and other security measures.

Thirdly, Information Classification supports the assignment of roles and responsibilities for data classification, including data owners, custodians, and users. This clear delineation of responsibilities ensures accountability and promotes adherence to data security policies.

Lastly, as suggested by us Information Classification facilitates compliance with legal and regulatory requirements, including those stipulated by SOC 2 Type II.
Information Content Extraction is a crucial process in data management that involves retrieving structured information from unstructured or semi-structured data sources. This process is essential for transforming raw data into meaningful and actionable insights.

Structured data is data that is organized into a formatted structure, often a relational database. This type of data is readily searchable by simple, straightforward search engine algorithms or other search operations.

Semi-structured data is a form of structured data that does not adhere to the formal structure of data models associated with relational databases or other forms of data tables but contains tags or other markers to separate semantic elements and enforce hierarchies of records and fields within the data. Examples of semi-structured data include XML and JSON files.

Unstructured data is information that either does not have a pre-defined data model or is not organized in a pre-defined manner. This type of data is typically text-heavy but may contain data such as dates, numbers, and facts as well. Examples of unstructured data include text files, PDFs, and BLOBs (Binary Large Objects).

Information extraction from these types of data involves several steps, including text preprocessing, entity recognition, relation extraction, and event extraction. Text preprocessing involves cleaning and normalizing the text, removing stop words, and stemming or lemmatizing words.

3.3. Framework

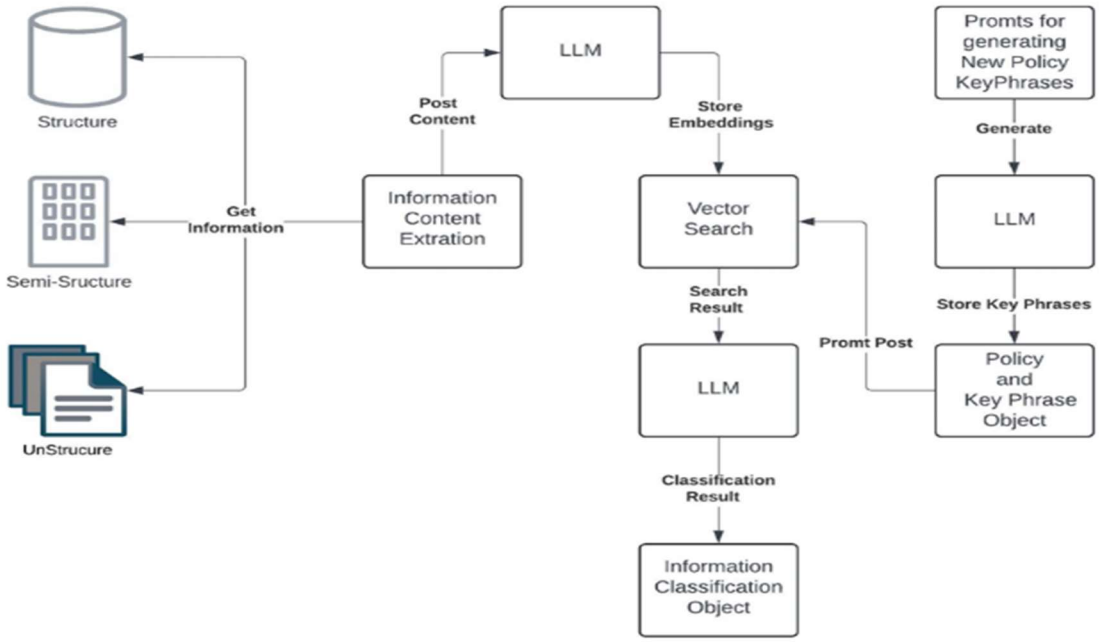

Figure 2: Information classification

Entity recognition identifies entities such as names, locations, and dates in the text. Relation extraction identifies relationships between these entities, and event extraction identifies events in which these entities are involved.

There are several approaches to information extraction:

1. Rule-based methods: These methods use a set of predefined rules or patterns to extract information. For example, a rule might specify that if a word is capitalized and followed by a certain verb, it is likely a person's name. While rule-based methods can be very accurate, they are also labor-intensive and may not generalize well to new data.
2. Machine learning methods: These methods use algorithms to learn patterns from labeled training data and apply these patterns to new data. For example, a machine learning model might learn that words that often appear in the same context as known person names are likely to be person names themselves. Machine learning methods can be very effective, especially with large amounts of training data, but they can also be complex and computationally intensive.
3. Hybrid methods: These methods combine rule-based and machine-learning methods to leverage the strengths of both. For example, a hybrid method might use rules to extract easy-to-identify information and machine learning to extract more complex information [36–37].

However, there are no clear recommendations regarding the implementation of a specific method, and the choice should be made considering a large number of factors.

Large Language Models (LLMs) [38–39] represent a significant advancement in the field of artificial intelligence. These models are trained on extensive volumes of text data, enabling them to generate text that closely resembles human writing. Notable examples of LLMs include GPT-3 by OpenAI and BERT by Google [40–42]. These models can perform a wide range of tasks, such as answering queries, crafting essays, summarizing texts, translating languages, and even generating creative ideas.

In the realm of data management and data governance, LLMs can be leveraged in several innovative ways:

1. Data Cataloging: LLMs can streamline the process of data cataloging. They can read and comprehend the metadata associated with various data assets and generate descriptions or tags for these assets, thereby automating a traditionally manual process.
2. Data Quality: LLMs can play a pivotal role in enhancing data quality. They can be trained to identify and flag potential errors or inconsistencies in data, facilitating proactive data quality management.
3. Data Privacy: LLMs can contribute to data privacy efforts by identifying and redacting sensitive information in datasets, thereby helping organizations comply with data privacy regulations.
4. Data Integration: LLMs can aid in data integration tasks. They can understand the context and semantics of different data sources and assist in mapping them to a common model, simplifying the integration process.

Choosing the right LLM for data management and data governance depends on various factors, including the specific requirements of the tasks, the size and complexity of the data, the computational resources available, and the expertise of the team.

Vector search, or nearest neighbor search, is a powerful technique utilized in machine learning and data science to identify items that are most similar to a given item. This method operates by representing items as vectors in a multi-dimensional space. Each point in this space corresponds to a potential item, and the position of that point is determined by the characteristics of the item.

The principle behind vector search is that similar items will be located near each other in this space, while dissimilar items will be further apart. When a new item is introduced, it is also converted into a vector and placed into this space. The algorithm then searches for vectors that are close to the new vector, with the "closeness" being determined by a distance metric such as Euclidean distance or cosine similarity.

This technique is particularly useful when dealing with large datasets, as it allows for efficient searching and retrieval of items. It's commonly used in recommendation systems, image recognition, and natural language processing among other applications.

For instance, in a movie recommendation system, each movie could be represented as a vector where each dimension corresponds to a different genre. A romance movie would be located closer to other romance movies and further from action movies. When a user rates a movie, the system can look for other movies that are close in the vector space to recommend to the user.

In essence, vector search is a method of transforming complex, abstract items into a format that can be easily and efficiently compared, enabling the rapid retrieval of similar items from large datasets.

Advantages of Vector Search:

1. High Accuracy: Vector search can provide highly accurate results because it considers the relationships between different features of the data. By representing data in a high-dimensional space, it captures the nuances and complexities of the data that might be missed by other methods.
2. Scalability: Vector search is highly scalable and can handle large amounts of data efficiently. This makes it suitable for big data applications where traditional search methods may be impractical.
3. Flexibility: Vector search is highly flexible and can be used with any data that can be represented as a vector. This includes text, images, audio, and more, making it applicable to a wide range of tasks and industries.

Disadvantages of Vector Search:

1. Computational Complexity: Vector search can be computationally intensive, especially when dealing with high-dimensional data or large datasets. This can make it slower than other methods, particularly for real-time applications.
2. Difficulty in Choosing the Right Distance Metric: The effectiveness of vector search heavily depends on the choice of distance metric, which can be challenging to determine. The choice of metric can significantly impact the results, and there is often no one-size-fits-all solution.
3. Sensitivity to Noise: Vector search can be sensitive to noise in the data. Outliers or errors in the data can affect the distance calculations and lead to inaccurate results.

Embeddings are a key component of vector search. In machine learning, an embedding is a learned representation for some specific type of data, such as words, users, or products, where similar items have a similar representation. They are used to convert categorical data into a form that can be input into a model. Embeddings are particularly useful for dealing with high-dimensional data, as they can reduce the dimensionality of the data while preserving its structure and relationships [43–45].

Prompts play a crucial role in the functioning of Large Language Models (LLMs) like GPT-3. A prompt is essentially an input that is given to the model to guide its output. It can be a question, a statement, or any piece of text. The LLM generates a response to the prompt based on the patterns it learned during its training on a large corpus of text data.

Prompts are valuable because they allow us to direct the model's output. By carefully crafting our prompts, we can guide the model to generate useful and relevant responses. For instance, if we're using an LLM to write an email, we might prompt it with "Dear [Recipient's Name], I am writing to inform you that..." and the model could generate the rest of the email.

In the context of data management, prompts can be used to extract or generate specific pieces of information from or about our data. For example, we could prompt an LLM with a question about our data, such as "What is the average value of column X?" or "How many entries in column Y are above Z?". The LLM could then generate a response based on its understanding of the data.

Prompts can also be used to generate metadata for our data. For instance, we could prompt the LLM with a piece of data and ask it to generate a description or a set of tags for that data. This could be particularly useful for tasks like data cataloging, where we need to generate human-readable descriptions or annotations for large amounts of data.

However, it's important to note that the effectiveness of prompts depends on the quality of the LLM's training. If the LLM has not been trained on relevant data, or if it has not been trained to understand the specific format or context of the prompts, it may not generate useful responses. Therefore, careful prompt design and model training are crucial for getting the most value out of LLMs in data management [46].

4. Conclusions

According to the document: In conclusion, the paper discusses the importance of information classification in the context of SOC 2 Type II compliance. Information classification serves as the foundation for data security and privacy controls, helping organizations identify and protect their most sensitive data. By effectively classifying their

data, organizations can ensure its security, meet regulatory requirements, and ultimately, safeguard their reputation and business continuity. To optimize and increase efficiency in the classification and organization of data by SOC 2 Type II standards, it is proposed to apply Large Language Models in this model. LLMs like GPT-3 and BERT, trained on extensive text data, are transforming data management and governance, areas crucial for SOC 2 Type II compliance. LLMs respond to prompts, guiding their output generation, and can automate tasks like data cataloging, enhancing data quality, ensuring data privacy, and assisting in data integration. These capabilities can support a robust data classification policy, a key requirement for SOC 2 Type II.

Vector search, another important method in data management, finds similar items to a given item by representing them as vectors in a high-dimensional space. It offers high accuracy, scalability, and flexibility, supporting efficient data classification. Embeddings, which convert categorical data into a form that can be input into a model, play a key role in vector search and LLMs.

Prompt engineering, the crafting of effective prompts, is crucial for guiding LLMs' output, and further enhancing data management and governance practices.

References

[1] B. Matturdi, et al., Big Data security and privacy: A review, *China Communications*, 11(14) (2014) 135–145. doi: 10.1109/CC.2014.7085614.

[2] V. Susukailo, I. Opirskyy, S. Vasylyshyn, Analysis of the attack vectors used by threat actors during the pandemic, IEEE 15th International Scientific and Technical Conference on Computer Sciences and Information Technologies, CSIT 2020 - Proceedings, 2 (2020) 261–264.

[3] M. N. Islam, et al., Security threats for big data: An empirical study, Int. J. Inf. Commun. Technol. Human Dev. (IJICTHD) 10(4) (2018) 1–18.

[4] A. Singh, A. Kumar, S. Namasudra: DNACDS: Cloud IoE big data security and accessing scheme based on DNA cryptography, Frontiers Comput. Sci. 18(1) (2024) 181801.

[5] O. I. Harasymchuk, et al., Generator of pseudorandom bit sequence with increased cryptographic security, Metallurgical and Mining Industry: Sci. Tech. J. 5 (2014) 25–29.

[6] V. Dudykevych, H. Mykytyn, K. Ruda, The concept of a deepfake detection system of biometric image modifications based on neural networks, in: 3rd KhPI Week on Advanced Technology (KhPIWeek) (2022) 1–4. doi: 10.1109/KhPIWeek57572.2022.9916378.

[7] O. Vakhula, I. Opirskyy, O. Mykhaylova, Research on Security Challenges in Cloud Environments and Solutions based on the "Security-as-Code" Approach, in: Cybersecurity Providing in Information and Telecommunication Systems, vol. 3550 (2023) 55–69.

[8] V. Maksymovych, et al., Development of Additive Fibonacci Generators with Improved Characteristics for Cybersecurity Needs, Appl. Sci. 12(3) (2022) 1519. doi: 10.3390/app12031519.

[9] V. Maksymovych, et al., Combined Pseudo-Random Sequence Generator for Cybersecurity, Sensors 22 (2022) 9700. doi: 10.3390/s22249700.

[10] SOC 2 Compliance Documentation URL: https://secureframe.com/hub/soc-2/compliance-documentation

[11] ISO/IEC 27001:2022 URL: https://www.iso.org/standard/27001

[12] V. Maksymovych, et al., Simulation of Authentication in Information-Processing Electronic Devices Based on Poisson Pulse Sequence Generators. Electronics 11(13) (2022). doi: 10.3390/electronics11132039.

[13] 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.

[14] O. Mykhaylova, et al., Mobile Application as a Critical Infrastructure Cyberattack Surface, in: Cybersecurity Providing in Information and Telecommunication Systems, vol. 3550 (2023) 29–43.

[15] J. Yi, Y. Wen, An Improved Data Backup Scheme Based on Multi-Factor Authentication, in: IEEE 9th Intl Conference on Big Data Security on Cloud (BigDataSecurity), IEEE Intl Conference on High Performance and Smart Computing, (HPSC) and IEEE Intl Conference on Intelligent Data and Security (IDS) (2023). doi: 10.1109/BigDataSecurity-HPSC-IDS58521.2023.00041.

[16] D. Shevchuk, et al., Designing Secured Services for Authentication, Authorization, and Accounting of Users, in: Cybersecurity Providing in Information and Telecommunication Systems II, vol. 3550 (2023) 217–225.

[17] Y. Martseniuk, et al., Automated Conformity Verification Concept for Cloud Security, in: Cybersecurity Providing in Information and Telecommunication Systems, vol. 3654 (2024) 25–37.

[18] A. Horpenyuk, I. Opirskyy, P. Vorobets, Analysis of Problems and Prospects of Implementation of Post-Quantum Cryptographic Algorithms, in: Classic, Quantum, and Post-Quantum Cryptography, vol. 3504 (2023) 39–49.

[19] A. Calder, S. Watkins, IT Governance: An International Guide to Data Security and ISO27001/ISO27002 (2019).

[20] AICPA, SOC 2® - SOC for Service Organizations: Trust Services Criteria. URL: https://us.aicpa.org/interestareas/frc/assuranceadvisoryservices/soc-for-service-organizations

[21] IS Audit Basics: The Domains of Data and Information Audits URL: https://www.isaca.org/resources/isaca-journal /issues/2016/volume-6/is-audit-basics-the-domains-of-data-and-information-audits

[22] Practical Data Security and Privacy for GDPR and CCPA, ISACA J. 3 (2020).

[23] Boosting Cyber Security with Data Governance and Enterprise Data Management, ISACA J. 3 (2017).

[24] D. Cannon, IT Service Management: A Guide for ITIL Foundation Exam Candidates, BCS (2012).

[25] N. Karumanchi, Data Structures and Algorithms Made Easy: Data Structures and Algorithmic Puzzles (2011).
[26] R. T. Watson, Data Management: Databases and Organizations (2017).
[27] M. Rhodes-Ousley, Information Security: The Complete Reference, Second Edition (2012).
[28] Munawar, Extract Transform Loading (ETL) Based Data Quality for Data Warehouse Development, in: 1st International Conference on Computer Science and Artificial Intelligence (ICCSAI), Jakarta, Indonesia (2021) 373–378. doi: 10.1109/ICCSAI53272.2021.9609770.
[29] V. Khoma, et al., Comprehensive Approach for Developing an Enterprise Cloud Infrastructure in: Cybersecurity Providing in Information and Telecommunication Systems, vol. 3654 (2024) 201–215.
[30] S. Chauhan, Mastering Apache Airflow (2020).
[31] A. Gaikwad, Learning AWS Glue (2021).
[32] D. Anoshin, R. Avdeev, R. van Vliet, Azure Data Factory Cookbook (2020).
[33] S. Hoberman, Data Modeling Made Simple: A Practical Guide for Business and IT Professionals (2005).
[34] C. C. Aggarwal, Data Classification: Algorithms and Applications (2014).
[35] J. Sharp, Y. Duhamel, Microsoft Power Platform Enterprise Architecture (2020).
[36] B. Magnini, et al., From Text to Knowledge for the Semantic Web: the ONTOTEXT project, in: Proceedings of SWAP 2005 Workshop (2005).
[37] S. Chakrabarti, Mining the Web. Discovering Knowledge from Hypertext Data, Morgan Kaufmann (2002).
[38] X. Yang, et al., Exploring the Application of Large Language Models in Detecting and Protecting Personally Identifiable Information in Archival Data: A Comprehensive Study, in: *IEEE International Conference on Big Data (BigData),* Sorrento, Italy, (2023) 2116–2123. doi: 10.1109/BigData59044.2023.10386949.
[39] A. Piskozub, D. Zhuravchak, A. Tolkachova, Researching vulnerabilities in chatbots with LLM (Large language model), Ukrainian Sci. J. Inf. Secur. 29(9) (2023) 111–117. doi: 10.18372/2225-5036.29.18069.
[40] GPT-3by OpenAI. URL: https://openai.com/research/gpt-3/
[41] BERT by Google. URL: https://ai.googleblog.com/2018/11/open-sourcing-bert-state-of-art-pre.html/
[42] Amazon Bedrock – Automating Large-Scale, Fault-Tolerant Distributed Training in the Deep Learning Compiler Stack. URL: https://aws.amazon.com/blogs/aws/amazon-bedrock-automating-large-scale-fault-tolerant-distributed-training-in-the-deep-learning-compiler-stack/
[43] A. N. Papadopoulos, Y. Manolopoulos, Nearest Neighbor Search: A Database Perspective (2004).
[44] J. Leskovec, A. Rajaraman, J. D. Ullman, Mining of Massive Datasets (2014).
[45] I. Goodfellow, Y. Bengio, A. Courville, Deep Learning (2016).
[46] Teaching with AI. URL: https://openai.com/blog/teaching-with-ai