http://doi.org/10.35784/iapgos.6910 *received: 06.12.2024 | revised: 21.08.2025 | accepted: 04.09.2025 | available online: 30.09.2025*

DETECTION CONFIDENTIAL INFORMATION BY LARGE LANGUAGE MODELS

Oleh Deineka[1], Oleh Harasymchuk[1], Andrii Partyka[1], Yurii Dreis[2], Yuliia Khokhlachova[3], Yuriy Pepa[4]
[1]Lviv Polytechnic National University, Lviv, Ukraine, [2]Mariupol State University, Kyiv, Ukraine, [3]State University of Trade and Economics, Kyiv, Ukraine, [4]State University of Information and Communication Technologies, Kyiv, Ukraine

Abstract. *In today's digital age, the protection of personal and confidential customer data is paramount. With the increasing volume of data being generated and processed, organizations face significant challenges in ensuring that sensitive information is adequately protected. One of the critical steps in safeguarding this data is the detection and classification of personal and confidential information within text documents. This process involves identifying sensitive data, classifying it appropriately, and storing the results in a semi-structured format such for further analysis and action. The need for detecting and classifying sensitive data is driven by regulatory compliance, data security, risk management, and operational efficiency. Various methodologies, including rule-based systems, machine learning models, natural language processing (NLP), and hybrid approaches, are employed to detect and classify sensitive data. 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 2 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 2.*

Keywords: data security, prompt, confidence, quality, information classification

WYKRYWANIE INFORMACJI POUFNYCH ZA POMOCĄ DUŻYCH MODELI JĘZYKOWYCH

Streszczenie. *W dzisiejszej erze cyfrowej ochrona danych osobowych i poufnych informacji klientów jest niezwykle ważna. Wraz ze wzrostem ilości generowanych i przetwarzanych danych, organizacje stają przed znacznymi wyzwaniami w zapewnieniu odpowiedniej ochrony wrażliwych informacji. Jednym z kluczowych kroków w zabezpieczaniu tych danych jest wykrywanie i klasyfikacja danych osobowych i poufnych w dokumentach tekstowych. Proces ten obejmuje identyfikację wrażliwych danych, odpowiednią ich klasyfikację oraz przechowywanie wyników w ustrukturyzowanym formacie, takim jak JSON, w celu dalszej analizy i działań. Potrzeba wykrywania i klasyfikacji wrażliwych danych wynika z wymogów regulacyjnych, bezpieczeństwa danych, zarządzania ryzykiem i efektywności operacyjnej. Do wykrywania i klasyfikacji wrażliwych danych stosuje się różne metody, w tym systemy oparte na regułach, modele uczenia maszynowego, przetwarzanie języka naturalnego (NLP) oraz podejścia hybrydowe. Duże modele językowe (LLM), takie jak GPT-3 i BERT, szkolone na dużych zbiorach danych tekstowych, przekształcają zarządzanie danymi i ich nadzór, co jest kluczowe dla zgodności z SOC 2 Type 2. LLM odpowiadają na zapytania, kierując generowaniem ich wyników, i mogą automatyzować zadania, takie jak katalogowanie danych, poprawa jakości danych, zapewnienie prywatności danych i wspieranie integracji danych. Te możliwości mogą wspierać solidną politykę klasyfikacji danych, która jest kluczowym wymogiem dla SOC 2 Type 2.*

Słowa kluczowe: bezpieczeństwo danych, zapytanie, pewność, jakość, klasyfikacja informacji

Introduction

In today's digital environment, organizations continuously generate and process massive amounts of textual information that may contain personal and confidential data. Despite the availability of various data protection tools, there remains a critical gap in the reliable detection and classification of sensitive information within unstructured text documents.

Traditional approaches, such as rule-based systems and classical machine learning models, often demonstrate limitations:

- low adaptability to new or non-standard data formats;
- high maintenance cost, as they require constant updating of rules or labeled datasets;
- limited contextual understanding, which reduces their ability to correctly identify sensitive data embedded in complex text.

Large Language Models (LLMs) such as GPT-3, GPT-4, Claude, and Gemini provide fundamentally new opportunities for analyzing natural language. Their advanced contextual understanding and ability to generate structured responses position them as potential solutions for the automation of confidential data detection. However, their effectiveness, precision, and reliability in this specific task have not been sufficiently studied and evaluated.

Thus, the scientific problem addressed in this article is the development of methodologies and evaluation criteria for detecting confidential information in text using LLMs, taking into account factors such as detection quality, computational cost, and compliance with data security standards (e.g., SOC 2 Type 2).

In Large Language Models (LLMs) [12, 29] represent a significant advancement in the field of natural language processing (NLP). These models, such as OpenAI's GPT-3 and GPT-4, Google DeepMind's Gemini, Anthropic's Claude, and Meta's LLaMA, are designed to understand, generate, and manipulate human language with unprecedented accuracy and fluency. They are built on deep learning architectures, particularly transformer networks, which enable them to process vast amounts of text data. LLMs are trained on diverse datasets encompassing books, articles, websites, and other textual resources, allowing them to capture a wide range of linguistic patterns and contextual nuances.

The primary strength of LLMs lies in their ability to generate human-like text based on the input they receive. They can perform a multitude of tasks, including text completion, translation, summarization, question answering, and more. The size of these models, often measured in billions of parameters, contributes to their effectiveness but also poses challenges in terms of computational resources and ethical considerations.

Azure OpenAI's GPT-3 and GPT-4: These models are among the most well-known LLMs, capable of generating coherent and contextually appropriate text across various domains. They have been widely adopted in applications ranging from chatbots to content creation [18, 24].

Google DeepMind's Gemini: Gemini is another advanced LLM that excels in understanding and generating human language. It is particularly noted for its ability to handle complex tasks and generate high-quality text, making it suitable for a range of applications, including research and development [13].

AWS Anthropic's Claude: Claude is designed with a focus on safety and ethical considerations. It aims to generate text that is not only accurate but also aligns with ethical guidelines, making it a preferred choice for applications where responsible AI use is paramount [22, 23].

Meta's LLaMA: LLaMA (Large Language Model Meta AI) is another prominent LLM known for its ability to process and generate text efficiently. It has been utilized in various applications, including social media analysis and content moderation [33].

The versatility of these models allows them to be integrated into numerous applications, enhancing productivity and enabling new possibilities in fields such as customer support, healthcare, education, and more. However, the deployment of LLMs must

be carefully managed to address potential ethical concerns, such as bias and misuse, ensuring that their benefits are realized responsibly [20, 28, 31].

1. Large Language Model – LLM

1.1. Core architecture of LLMs

A Large Language Model (LLM) is an advanced artificial intelligence system designed to understand, interpret, and generate human language. These models are built upon sophisticated neural network architecture and are trained in extensive corpora of text data, enabling them to learn the intricate statistical relationships between words, phrases, and sentences. This training allows LLMs to predict and generate coherent and contextually relevant text based on the input they receive.

Core architecture: At the heart of an LLM is a neural network, often comprising multiple layers of transformers. Transformers are specialized components of neural networks that excel at handling sequential data, such as text. They employ mechanisms like self-attention, which allows the model to weigh the importance of different words in a sentence. This capability is crucial for capturing long-range dependencies and contextual information, making the model adept at understanding the nuances of human language.

Training process: The training process for an LLM is both computationally intensive and resource demanding. It involves feeding the model vast amounts of text data and iteratively adjusting its parameters to minimize prediction errors. This is typically done using powerful GPUs and extensive datasets, which can range from general internet text to specialized corpora tailored for specific applications. The goal is to enable the model to generalize from the training data, allowing it to perform well on unseen text.

Fine-tuning and applications: Once the initial training phase is complete, the LLM can be fine-tuned for specific tasks or domains. Fine-tuning involves additional training on a smaller, task-specific dataset, which helps the model adapt its general language understanding to particular applications. This process enhances the model's performance in targeted areas, such as sentiment analysis, machine translation, or question-answering systems.

Practical implications: The capabilities of LLMs have far-reaching implications across various fields. In natural language processing (NLP), they are used to power chatbots, virtual assistants, and automated content generation tools. In academia and research, LLMs assist in data analysis, literature reviews, and even hypothesis generation. In business, they are employed for customer service automation, market analysis, and personalized marketing strategies.

Challenges and considerations: Despite their impressive capabilities, LLMs are not without challenges. The training process is resource-intensive, requiring significant computational power and large datasets. Additionally, the models can sometimes generate biased or inappropriate content, reflecting the biases present in the training data. Ethical considerations, such as data privacy and the potential misuse of generated content, are also critical issues that need to be addressed.

In summary, Large Language Models represent a significant advancement in artificial intelligence, offering powerful tools for understanding and generating human language. Their development and deployment, however, require careful consideration of computational resources, ethical implications, and potential biases [30].

1.2. Area of use

Large Language Models (LLMs) have demonstrated their versatility and broad applicability across various domains. Here are some key areas where LLMs are making a significant impact.

Customer support: LLMs power chatbots and virtual assistants, providing automated customer service and support. They can handle inquiries, troubleshoot issues, and guide users through processes, thereby improving efficiency and user satisfaction. These models can understand and respond to a wide range of customer queries, making them invaluable for businesses looking to enhance their customer service operations.

Content creation: In the media and publishing industries, LLMs assist in generating articles, summaries, and creative writing. They can draft reports, emails, and marketing content, save time and enhance productivity. By automating routine writing tasks, LLMs allow human writers to focus on more complex and creative aspects of content creation.

Healthcare: LLMs support medical professionals by summarizing patient records, generating medical reports, and assisting in diagnosing conditions based on textual data. They also enable the creation of personalized health advice and reminders. This can lead to more efficient patient care and better utilization of medical resources.

Education: In educational settings, LLMs provide personalized tutoring, generate educational content, and assist in grading and feedback. They can facilitate language learning by offering translations and practice exercises. By tailoring educational experiences to individual needs, LLMs can enhance learning outcomes and make education more accessible.

Research and development: Researchers use LLMs to analyze large datasets, generate hypotheses, and draft research papers. They assist in literature reviews and summarizing academic articles, making the research process more efficient. LLMs can also help identify trends and insights that might be missed through manual analysis.

Legal: In the legal field, LLMs aid in drafting legal documents, summarizing case law, and conducting legal research. They assist in contract analysis and identifying relevant legal precedents. By automating routine legal tasks, LLMs can help legal professionals focus on more complex and strategic aspects of their work.

Finance: LLMs support financial analysts by generating market reports, summarizing financial news, and providing insights based on textual data. They assist in automating customer interactions in banking and insurance, improving service efficiency and customer satisfaction. LLMs can also help in risk assessment and fraud detection by analyzing large volumes of financial data [8, 15, 17, 19].

Cybersecurity: Large Language Models (LLMs) have become a powerful tool in the field of software vulnerabilities and cybersecurity, opening up broad prospects for threat identification and analysis [1, 2, 9, 11, 32]. Their ability to process significant volumes of data allows for faster and more efficient threat detection, helping to enhance overall security in the digital environment. In particular, they are actively used for tasks such as:

- threat detection and analysis: helping to identify malicious code, phishing, and other threats by processing large volumes of data and logs;
- incident response automation: recognizing anomalies and breaches, accelerating the response to cyber incidents;
- vulnerability assessment: identifying potential vulnerabilities in code by analyzing its structure and content;
- phishing and spam detection: identifying suspicious messages and detecting phishing and malicious elements;
- natural language analysis: predicting new threats by analyzing textual data, such as from social networks or forums;
- training specialists: models support the training of specialists by simulating cyber threats.

These capabilities accelerate cybersecurity processes, allowing teams to focus on strategic aspects of security.

The aim of the research is to determine the effectiveness of detecting sensitive data using different language models.

2. SOC2 type 2 policy of information classification

Requirements

The SOC 2 Type 2 policy [10, 16, 27, 34] is a critical framework for ensuring the security, availability, processing integrity, confidentiality, and privacy of an organization's data. It requires organizations to implement and document comprehensive controls over their systems and data. This includes the establishment of policies and procedures that govern data classification, access controls, data protection measures, and incident response. The policy must also ensure continuous monitoring and logging of data usage and access. Regular audits and reviews are conducted to ensure compliance and identify any potential security incidents. Additionally, employee training and awareness programs are essential components to ensure that all personnel understand and adhere to the established policies.

It is important to ensure that the SOC 2 Type 2 policy is fully understood and implemented correctly. This includes verifying that all data classification levels are clearly defined and that roles and responsibilities are assigned appropriately. It is also crucial to maintain an up-to-date data inventory and mapping to ensure accurate tracking and protection of all data assets.

Information classification policy

1. Definition and scope.

Establish a formal Information Classification Policy that defines the categories of data, including Confidential and Personal Customer Data.

Ensure the policy covers all data types, storage mediums, and transmission methods.

2. Classification levels.

Define classification levels such as Public, Internal, Confidential, and Personal Customer Data.

Provide clear criteria for classifying data into each category.

3. Roles and responsibilities.

Assign roles and responsibilities for data classification to specific personnel, including Data Owners, Data Stewards, and Data Custodians.

Ensure accountability for maintaining the classification of data throughout its lifecycle.

Data Inventory and mapping

1. Data inventory.

Maintain an up-to-date inventory of all data assets, including their classification.

Include details such as data type, location, owner, and access controls.

2. Data flow mapping.

Document and regularly update data flow diagrams that illustrate how Confidential and Personal Customer Data moves through the organization.

Identify all points of data entry, processing, storage, and exit.

Access controls

1. Access management.

Implement strict access controls to ensure only authorized personnel can access Confidential and Personal Customer Data.

Use role-based access control (RBAC) and the principle of least privilege.

2. Authentication and authorization.

Enforce multi-factor authentication (MFA) for accessing systems containing sensitive data.

Regularly review and update access permissions.

Data protection measures

1. Encryption.

Encrypt Confidential and Personal Customer Data both at rest and in transit using industry-standard encryption protocols.

Regularly update encryption keys and manage them securely.

2. Data masking and anonymization.

Implement data masking or anonymization techniques where appropriate to protect sensitive data in non-production environments.

3. Secure disposal.

Establish procedures for the secure disposal of Confidential and Personal Customer Data when it is no longer needed.

Ensure data is irretrievably deleted from all storage mediums.

Monitoring and logging

1. Activity monitoring.

Continuously monitor access to and usage of Confidential and Personal Customer Data.

Implement automated tools to detect and alert 31, on unauthorized access or anomalies.

2. Logging and auditing.

Maintain detailed logs of access and changes to sensitive data.

Regularly audit logs to ensure compliance with policies and identify potential security incidents.

Incident response and management

1. Incident response plan.

Develop and maintain an incident response plan specifically addressing breaches involving Confidential and Personal Customer Data.

Include procedures for containment, investigation, notification, and remediation.

2. Breach notification.

Establish protocols for timely notification to affected customers and regulatory bodies in the event of a data breach.

Ensure compliance with relevant data protection laws and regulations.

Training and awareness

1. Employee training.

Conduct regular training sessions for employees on the importance of data classification and protection.

Include specific modules on handling Confidential and Personal Customer Data.

2. Awareness programs.

Implement ongoing awareness programs to reinforce the importance of data security and compliance with the Information Classification Policy.

Review and continuous improvement

1. Policy review.

Regularly review and update the Information Classification Policy to reflect changes in the regulatory environment, business processes, and technology.

Conduct periodic risk assessments to identify new threats and vulnerabilities.

2. Continuous improvement.

Establish a feedback loop for continuous improvement of data classification and protection measures.

Encourage reporting of potential weaknesses and suggestions for enhancements [7, 14, 34].

3. Research results

3.1. Overview of experiment

LLMs detections

In today's digital age, the protection of personal and confidential customer data is paramount. With the increasing volume of data being generated and processed, organizations face significant challenges in ensuring that sensitive information is adequately protected. One of the critical steps in safeguarding this data is the detection and classification of personal and confidential information within text documents. This process involves identifying sensitive data, classifying it appropriately, and storing the results in a semi-structured format for further analysis and action.

This overview aims to provide a comprehensive understanding of the need for detecting and classifying personal and confidential customer data, the methodologies involved, and the criteria for evaluating different models based on cost, performance, and detection quality.

The Need for Detecting and Classifying Sensitive Data Regulatory Compliance: Various regulations such as GDPR, CCPA, and HIPAA mandate the protection of personal data. Organizations must ensure that they comply with these regulations to avoid hefty fines and legal repercussions.

Data security: Identifying and classifying sensitive data helps in implementing appropriate security measures to protect it from unauthorized access, breaches, and leaks.

Risk management: By detecting and classifying sensitive data, organizations can assess the risk associated with data handling and take proactive measures to mitigate potential threats.

Operational efficiency: Automating the detection and classification process reduces the manual effort required, thereby increasing operational efficiency and allowing resources to focus on more strategic tasks.

3.2. Methodologies for detecting and classifying sensitive data

Rule-Based Systems: These systems use predefined rules and patterns to identify sensitive data. For example, regular expressions can be used to detect patterns such as Social Security Numbers, credit card numbers, and email addresses. While rule-based systems are straightforward to implement, they may not be flexible enough to handle complex data patterns and variations.

Machine Learning Models: Machine learning models can be trained to recognize sensitive data based on labeled datasets. These models can learn from examples and improve their detection accuracy over time. By the way in [21] was described case studies where machine learning models significantly improved the accuracy of sensitive data detection compared to traditional methods. Common machine learning techniques include decision trees, support vector machines, and neural networks.

Natural Language Processing (NLP): NLP techniques can be used to analyze the context and semantics of the text to identify sensitive information. Named Entity Recognition (NER) is a popular NLP technique that can detect entities such as names, locations, and organizations within the text.

Hybrid Approaches: Combining rule-based systems with machine learning and NLP techniques can provide a more robust solution for detecting and classifying sensitive data. Hybrid approaches leverage the strengths of each methodology to improve detection accuracy and reduce false positives.

While methodologies for detecting and classifying sensitive data offer significant benefits, they also come with certain disadvantages. Rule-based systems, for instance, can be rigid and may not adapt well to new or evolving data patterns, leading to potential gaps in detection. They also require continuous updates and maintenance to remain effective, which can be resource intensive. Machine learning models, on the other hand, necessitate large, labeled datasets for training, which can be difficult to obtain and may raise privacy concerns. Additionally, these models can be complex to implement and require significant computational resources. NLP techniques, while powerful, can struggle with the nuances and ambiguities of human language, potentially leading to misclassification. Hybrid approaches, though more robust, can be challenging to integrate and may require expertise in multiple domains to optimize effectively. Overall, while these methodologies enhance the ability to detect and classify sensitive data, they also introduce complexities and resource demands that must be carefully managed [4–6, 34].

Evaluation criteria for models

Cost: The cost of implementing and maintaining a detection and classification system is a critical factor. This includes the initial setup cost, licensing fees, hardware and software requirements, and ongoing maintenance expenses. Organizations must evaluate the total cost of ownership and ensure that the chosen solution fits within their budget. It is essential to consider both short-term and long-term financial implications to avoid unexpected expenses that could strain resources.

Performance: The performance of the model is measured in terms of its speed and efficiency in processing large volumes of data. Key performance indicators include processing time, throughput, and resource utilization. A high-performance model should be able to handle real-time data processing with minimal latency. Additionally, the model should maintain consistent performance under varying workloads to ensure reliability and effectiveness in different scenarios.

Detection quality: The quality of detection is assessed based on the model's accuracy, precision, recall, and F1 score. Accuracy measures the overall correctness of the model, while precision and recall evaluate its ability to correctly identify true positives and minimize false positives and false negatives. The F1 score provides a balanced measure of precision and recall. High detection quality is crucial for maintaining the integrity of the data and ensuring that sensitive information is accurately identified and protected.

Scalability: The chosen solution should be scalable to accommodate the growing volume of data and the increasing complexity of data patterns. Scalability ensures that the system can handle future demands without significant performance degradation. This includes the ability to expand resources, such as processing power and storage, as needed to support larger datasets and more complex analysis.

Ease of integration: The detection and classification system should be easy to integrate with existing data processing workflows and systems. This includes compatibility with various data formats, APIs, and third-party tools. Seamless integration minimizes disruptions and allows organizations to leverage their current infrastructure while enhancing their data protection capabilities.

User-friendliness: A user-friendly interface and intuitive configuration options are essential for ensuring that the system can be easily managed and operated by non-technical users. This includes clear documentation, support resources, and training materials. A user-friendly system reduces the learning curve and empowers users to effectively utilize the system without extensive technical knowledge.

Saving results in JSON format

Once the sensitive data has been detected and classified, the results need to be saved in a structured format for further analysis and action. JSON (JavaScript Object Notation) is a widely used format for storing and exchanging data due to its simplicity and readability. The JSON format allows for easy integration with various data processing tools and systems. By using JSON, organizations can ensure that the detected data is organized and accessible for subsequent processing and reporting.

A typical JSON output for detected and classified data might include the following fields:

1. Document ID: A unique identifier for the document being processed.
2. Detected Entities: An array of detected entities, each with the following attributes.
3. Entity Type: The type of sensitive data detected (e.g., name, email, SSN).
4. Entity Value: The actual value of the detected entity.

This structured approach facilitates efficient data management and supports comprehensive analysis, enabling organizations to take appropriate actions based on the detected sensitive information.

This structured approach facilitates efficient data management and supports comprehensive analysis, enabling organizations to take appropriate actions based on the detected sensitive information. By organizing data in a clear and consistent format, it enhances data security and aids in regulatory compliance. The use of JSON format allows for seamless integration with machine learning models and analytical tools, fostering deeper insights and improved detection algorithms. JSON's flexibility supports customization to meet specific organizational needs, ensuring relevance as requirements evolve. Its readability makes it accessible to both technical and non-technical users, facilitating easier troubleshooting and data manipulation. JSON supports interoperability with a wide range of software applications, maximizing return on investment and promoting best practices in data management. Automation in data processing workflows is enhanced, increasing efficiency and reducing human error. The structured data can be easily shared across departments, fostering collaboration and informed decision-making. Overall, saving results in JSON format enhances the robustness and scalability of data management systems. This approach provides a solid foundation for advanced data protection strategies and effective response to emerging threats.

Confidence Score: A score indicating the confidence level of the detection.

Figure 1 illustrates several use cases of employing Large Language Models (LLMs). It was published and researched in publication [16].

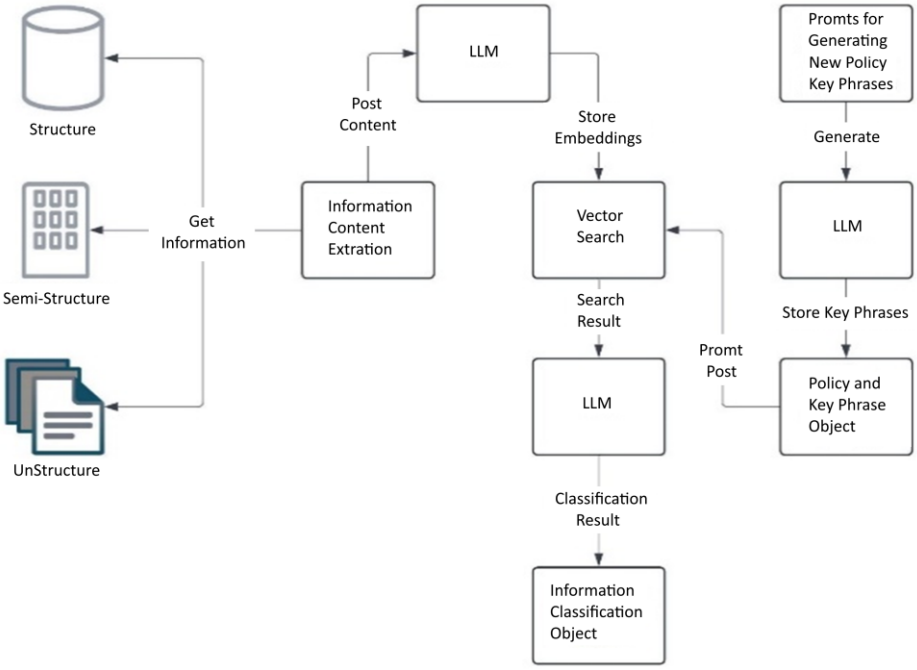


Fig. 1. Information classification

These use cases include:

Creating Embedders for Information Retrieval: One prominent application involves using LLMs, such as OpenAI's ADA model, to create embedders for efficient information retrieval. Embedders transform text into dense vector representations that capture semantic meaning, enabling more accurate and context-aware searches.

Generating Policies for Prompts: Another innovative idea within the framework is generating policies for prompts. This involves training LLMs to recognize different types of prompts and generate appropriate responses based on predefined guidelines or rules. This ensures that responses are consistent, accurate, and contextually appropriate.

Extracting Policies from Unstructured Data: The final application focuses on using LLMs to extract policies from unstructured data and convert them into a semi-structured format. This process involves identifying.

Confidence Score: A score indicating the confidence level of the detection.

Position: The position of the detected entity within the document (e.g., start and end indices) [3, 4, 10, 26].

When considering the use of the Varonis Data Security Platform versus creating your own data classification product, it's important to weigh the potential disadvantages of each approach.

One of the primary concerns with Varonis is the cost. For small to medium-sized enterprises, the licensing fees, maintenance costs, and potential hidden expenses can quickly add up, making it a significant financial commitment. Additionally, implementing and managing Varonis requires specialized knowledge and training. Organizations without dedicated IT security teams might find this complexity to be a barrier, as it necessitates investing in training or hiring skilled personnel.

Customization limitations also pose a challenge. While Varonis offers robust features, it may not meet all the specific needs of an organization. Customizing the platform to fit unique requirements can be difficult and constrained by the software's inherent design. This can lead to a reliance on the vendor for updates and support, creating a situation of vendors lock-in. Such dependency can be risky if the vendor changes its business model, increases prices, or discontinues the product.

Integration issues are another potential drawback. Integrating Varonis with existing systems and workflows can be complex and time-consuming, potentially causing disruptions in daily operations. Moreover, while Varonis is designed to scale, organizations with rapidly growing data needs might find it challenging to scale efficiently without incurring significant additional costs.

Data privacy is a critical concern when using third-party platforms. Storing sensitive data on Varonis can raise issues about data privacy and compliance, especially in regulated industries where stringent data protection laws apply.

On the other hand, creating your own data classification product comes with its own set of challenges. The development process requires significant time, resources, and expertise. This can divert focus from core business activities and delay the implementation of a functional solution. Once developed,

an in-house solution demands ongoing maintenance, updates, and security patches, which can be resource-intensive and require continuous investment.

Moreover, there is a risk of inadequate security with an in-house solution. If not properly designed and tested, the product might have security vulnerabilities, potentially exposing the organization to data breaches and compliance issues.

In conclusion, each approach has its trade-offs. The decision to use Varonis or develop an in-house solution should be based on the organization's specific needs, available resources, and long-term strategic goals. Careful consideration of these factors will help determine the most suitable path for effective data classification and management.

3.3. LLM prompt

A prompt in the context of Large Language Models (LLMs) like GPT-3, GPT-4, and others, is the initial input or query provided to the model to generate a response. It serves as the starting point for the model to understand the context and produce relevant and coherent text. Prompts can range from simple questions to complex instructions, depending on the desired output.

Prompts are crucial for several reasons:

Guidance: They direct the model towards the specific task or information you need. Without a clear prompt, the model may generate irrelevant or off-topic responses.

Context: Prompts provide context, helping the model understand the background and nuances of the query.

Control: They allow users to control the tone, style, and specificity of the generated text.

Efficiency: Well-crafted prompts can lead to more accurate and useful outputs, saving time and computational resources.

Creating effective prompts is both an art and a science. Here are some best practices to consider:

Be clear and specific: Ambiguity can lead to unpredictable results. Clearly state what you want the model to do.

Example: Instead of "Tell me about dogs," use "Provide a brief overview of the different breeds of dogs and their characteristics."

Provide context: The more context you provide, the better the model can understand and generate relevant responses.

Example: "In the context of urban living, what are the best dog breeds for small apartments?"

Use examples: If possible, provide examples of the desired output.

Example: "Write a product description for a new smartphone.
Example: 'The XYZ Phone features a 6.5-inch display, 128GB storage, and a 48MP camera.'"

Ask direct questions: Direct questions can lead to more focused and concise answers.

Example: "What are the health benefits of a Mediterranean diet?"

Iterate and Refine: Experiment with different prompts and refine them based on the quality of the responses.

Example: If the initial prompt "Explain climate change" yields too broad an answer, refine it to "Explain the impact of climate change on coastal cities."

While crafting prompts, it's essential to avoid certain pitfalls that can lead to suboptimal results:

Being too vague: Vague prompts can confuse the model and result in irrelevant or generic responses.

Example to Avoid: "Tell me something interesting."

Overloading with information: Providing too much information can overwhelm the model and dilute the focus.

Example to Avoid: "Explain the history of the Roman Empire, including its founding, major battles, key emperors, cultural achievements, and eventual decline."

Using ambiguous language: Ambiguity can lead to misinterpretation.

Example to Avoid: "Describe a good book." (What constitutes "good" can vary widely.)

Ignoring model limitations: Be aware of the model's limitations and avoid asking for highly specialized or sensitive information.

Example to Avoid: "Provide a detailed medical diagnosis based on these symptoms."

Neglecting ethical considerations: Ensure that your prompts do not encourage harmful, biased, or unethical responses.

Example to Avoid: "Generate a controversial opinion on a sensitive topic." [3, 25].

This instruction generates rules for detecting confidential information based on specific key terms that are needed for further analysis.

Prompt for experiment look like that.

The first section its generated text with confidential data:

(Once upon a time, in the beautiful country of Ukraine, there lived a talented and ambitious man named Oleg Deineka (email:deinekaoleg@gmail.com). Oleg was known for his exceptional skills and expertise in his field, which had earned him a stellar reputation among his peers. He had always dreamed of working on international projects and making a significant impact on a global scale. One day, Oleg received an exciting offer from a prestigious company named Zuma, located in the sunny state of California, USA (email: zuma@gmail.com). Zuma was renowned for its innovative approach and cutting-edge technology, and Oleg was thrilled at the prospect of joining their team. The contract was valued at a substantial 100,000 y.o., and Oleg knew that this opportunity would be a turning point in his career. With his Ukrainian passport number FF0899 in hand and Credit Card Number is 8888-9999-0000-11111-44444. Oleg eagerly signed the contract and began preparing for his new adventure. He communicated with the company through email, exchanging ideas and plans with his future colleagues. The anticipation of starting this new chapter in his life filled him with excitement and determination. As the days passed, Oleg's excitement grew. He began to envision the possibilities that lay ahead – the chance to work with some of the brightest minds in the industry, to contribute his expertise to groundbreaking projects, and to experience the vibrant culture of California. He knew that this opportunity would not only enhance his professional growth but also enrich his personal life. Oleg's journey to California was filled with anticipation and wonder. As he boarded the plane, he couldn't help but feel a sense of pride and accomplishment. He was about to embark on a journey that would take him to new heights, both professionally and personally. Upon arriving in California, Oleg was warmly welcomed by his new colleagues at Zuma. They were impressed by his skills and enthusiasm, and they quickly integrated him into their team. Oleg's expertise proved to be invaluable, and he soon became an integral part of the company's success. As Oleg settled into his new life in California, he embraced the opportunities that came his way. He explored the beautiful landscapes, enjoyed the diverse culture, and formed lasting friendships with his colleagues. His work at Zuma was both challenging and rewarding, and he relished every moment of it. Oleg's journey was a testament to the power of dreams and determination. He had taken a leap of faith, leaving behind the family to pursue his aspirations on a global stage. Through hard work and perseverance, he had achieved his goals and created a life filled with purpose and fulfillment. And so, Oleg Deineka's story became an inspiration to many – a reminder that with passion, dedication, and a willingness to embrace new opportunities, one can achieve greatness and make a lasting impact on the world. User: odein Cred: 1231)) WEW, kemqwe)

The second section is about what to do:

(Action: Please detect PII and Customer Confidential as in the result section and add a new entity if it exists in section one.)

The last one results section:

```
({
  "PII_and_Confidential_Data”:
    "Full_Name": "",
    "Passport_Number": "",
    "Nationality": "",
```

```
    "Email_Addresses": [''],
    "Login_Information”:
      "Username": "",
      "Password": ""
    },
    "Credit_Card_Number": "",
    "Client_Information”:
      "Client_Name": "",
      "Location": "",
      "Contract_Amount": "",
      "Username": "",
      "Password": "",
      "Email”: "
  }
  },
  "Request_Metadata”:
    "Token_Count": "(count tokens of Description)",
    "Cost": "(count costs of Description)",
    "Time_to_Response": "(count response time in milliseconds of
Description)"
  }
      })
```

3.4. LLMs request result

The expected result is to detect all confidential data in semi-structure format. We are going to test in on the next models GPT 3.5 Turbo, GPT 4, GPT 4o, Claude 3 Sonnet, Gemini 1.5 Pro.

Result:

```
(
{
  "PII_and_Confidential_Data”:
    "Full_Name": "Oleg Deineka",
    "Passport_Number": "FF0899",
    "Nationality": "Ukrainian",
    "Email_Addresses": [
      "deinekaoleg@gmail.com",
      "zuma@gmail.com"
    ],
    "Login_Information”:
      "Username": "odein",
      "Password": "1231)) WEW, kemqwe"
    },
    "Credit_Card_Number": "8888-9999-0000-11111-44444",
    "Client_Information”:
      "Client_Name": "Zuma",
      "Location": "California, USA",
      "Contract_Amount": "100,000 y.o.",
      "Username": "",
      "Password": "",
      "Email": "zuma@gmail.com"
    }
  },
  "Request_Metadata”:
    "Token_Count": 382,
    "Cost": "(cost calculation not provided)",
    "Time_to_Response": "(response time not provided)"
  }
}
)
```

The JSON response effectively identifies and categorizes all the personal identifiable information (PII) and confidential data present in the provided description. This structured approach ensures that sensitive information is clearly identified and can be appropriately handled according to privacy and security guidelines.

Key Points:

Identification and categorization: The response accurately identifies various forms of PII and confidential data, categorizing them into distinct fields such as full name, passport number, nationality, email addresses, login information, credit card number, and client information.

Sensitive information: The detection of sensitive information like passport numbers, credit card numbers, and login credentials highlights the importance of securing such data to prevent misuse and ensure privacy.

Client confidentiality: The inclusion of client-specific information such as the client's name, location, contract amount, and email address underscores the need to maintain confidentiality in business dealings.

Metadata insights: While the token count provides a quantitative measure of the text's length, the placeholders for cost and response time indicate areas where further details could enhance the understanding of processing metrics.

This comprehensive identification and categorization of sensitive information ensure that appropriate measures can be taken to protect and manage this data in compliance with privacy and security standards.

3.5. LLMs analytics results

The steps involved in this experiment include:

Dataset Preparation: Collect and annotate a diverse dataset of unstructured text data containing policy-related information. This dataset should include various document types, such as legal contracts, company policies, and regulatory guidelines.

However, LLMs have several limitations that need to be considered. One significant issue is the potential for generating biased or inaccurate information, especially if the training data contains inherent biases. Additionally, LLMs can struggle with understanding and maintaining context over long documents, which is particularly challenging for policy-related texts that often require nuanced comprehension. The computational resources required to train and deploy LLMs are substantial, making it difficult for smaller organizations to leverage these technologies effectively.

According to the results presented in this table 1, two models show the best performance in terms of detecting sensitive data and converting it into semi-structured format. Regarding speed, they are not the best but are acceptable for solving the task.

Lastly, ensuring data privacy and security is paramount, as handling sensitive policy-related prevents data breaches and misuse. Model Selection: Choose a range of models for evaluation, including:

Pre-trained LLMs like GPT-3 and its variants (e.g., ADA).

Named Entity Recognition (NER) models such Gemini or Claude models.

Pattern recognition models, including rule-based systems and deep learning models specialized in identifying policy-related patterns.

Performance Metrics: Use various metrics to evaluate the models, including:

Precision: The percentage of correctly identified policies out of all identified policies.

Recall: The percentage of correctly identified policies out of all actual policies in the text. F1 Score: The harmonic means of precision and recall, providing a balanced measure of the model's performance.

Accuracy: The overall percentage of correctly identified policies.

Extraction Quality: The quality of the extracted policies in terms of completeness and correctness.

During our data detection experiment, we achieved good results Two models were provided high detection quality with 96 %. We noticed that different vendors, like OpenAI and AWS, offer high-quality models. However, limitations and costs play a significant role in our decision-making process. When developing a solution, it's important to consider the constraints and limitations provided by our stakeholders, as well as have a clear understanding of the budget. This helps us choose the most suitable option for our needs. Additionally, we found that all models have similar performance and token usage (prompt and completion) for the same input.

Table 1. Model comparison

Vendor	Model	Input ($) (1000 tokens)	Output ($) (1000 tokens)	Prompt limits	Quality resp. (%)	Costs ($)	Performance (sec)	Tokens quantity
OpenAI	GPT 3.5 Turbo	0.0031	0.0041	16K	80.77	0.0034	24	1005
OpenAI	GPT 4	0.0618	0.0613	32K	80.77	0.0771	17	1007
OpenAI	GPT 4o	0.005	0.015	128K	96.15	0.0067	20	981
AWS	Claude 3 Sonnet	0.003	0.015	200K	96.15	0.0135	21	846
Google	Gemini 1.5 Pro	0.0000075	0.000023	128K	38.46	0.0023	16	1000

As for the models' limitations on the number of limits, they have now been increased and do not affect this task. The disadvantage of these models is hallucinations due to processing a large amount of unreliable data. As a result, all automation needs to be checked using data quality tools.

4. Conclusions

In the current digital era, safeguarding personal and confidential customer data is of utmost importance. With the exponential growth of data generation and processing, organizations face significant challenges in ensuring the protection of sensitive information. Detecting and classifying personal and confidential information within text documents is a crucial step in this process. This involves identifying sensitive data, classifying it accurately, and storing the results in a structured format like JSON for further analysis and action.

The necessity for detecting and classifying sensitive data is driven by regulatory compliance, data security, risk management, and operational efficiency. Various methodologies, including rule-based systems, machine learning models, natural language processing (NLP), and hybrid approaches, are employed to achieve this. Large Language Models (LLMs) such as GPT-3 and BERT, trained on extensive text data, are revolutionizing data management and governance, which are essential for SOC 2 Type 2 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 support a robust data classification policy, a key requirement for SOC 2 Type 2.

The evaluation criteria for these models include cost, performance, detection quality, scalability, ease of integration, and user-friendliness. Once sensitive data is detected and classified, the results need to be saved in a structured format for further analysis and action. JSON (JavaScript Object Notation) is a widely used format for storing and exchanging data due to its simplicity and readability. The JSON format allows for easy integration with various data processing tools and systems.

Large Language Models represent a significant advancement in artificial intelligence, offering powerful tools for understanding and generating human language. Their development and deployment, however, require careful consideration of computational resources, ethical implications, and potential biases. The versatility of these models allows them to be integrated into numerous applications, enhancing productivity and enabling new possibilities in fields such as customer support, healthcare, education, and more. However, the deployment of LLMs must be carefully managed to address potential ethical concerns, such as bias and misuse, ensuring that their benefits are realized responsibly.

References

[1] Amaratunga T.: Understanding Large Language Models. Apress, 2023.
[2] Berryman J., Ziegler A.: Prompt Engineering for LLMs. O'Reilly, 2024.
[3] Bezzi M.: Large Language Models and Security. IEEE Security & Privacy 22(2), 2024, 60–68 [https://doi.org/10.1109/MSEC.2023.3345568].
[4] Calder A., Watkins S.: IT Governance: An International Guide to Data Security and ISO27001/ISO27002 (6 edition). CoganPage, 2015.
[5] Jurafsky D., Martin J. H.: Speech and Language Processing (3 edition). Prentice-Hall, Inc., 2024.
[6] Deineka O., et. al.: Designing Data Classification and Secure Store Policy According to SOC 2 Type II. CEUR Workshop Proceedings 3654, 2024, 398–409 [https://ceur-ws.org/Vol-3654/short7.pdf].
[7] Dreis Y., et al.: Model to Formation Data Base of Internal Parameters for Assessing the Status of the State Secret Protection. Cybersecurity Providing in Information and Telecommunication Systems 3654, 2024, 277–289 [https://ceur-ws.org/Vol-3654/paper23.pdf].
[8] Falchenko S., et al.: Method of Fuzzy Classification of Information with Limited Access. IEEE 2nd International Conference on Advanced Trends in Information Theory (IEEE ATIT 2020) 2020, Kyiv, Ukraine, 255–259 [https://doi.org/10.1109/ATIT50783.2020.9349358].
[9] Giulio C. D., et. al.: IT Security and Privacy Standards in Comparison: Improving FedRAMP Authorization for Cloud Service Providers. 17th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing (CCGRID) 2017, Madrid, Spain, 1090–1099 [https://doi.org/10.1109/CCGRID.2017.137].
[10] Gupta B. B., Sheng Q. Z.: Machine Learning for Computer and Cyber Security. Boca Raton, 2019.
[11] Goldberg Y.: Neural Network Methods for Natural Language Processing. Springer, 2017.
[12] Manning C. D., Raghavan P., Schütze H.: Introduction to Information Retrieval. Cambridge University Press, 2008.
[13] Martseniuk Y., et. al.: Research of the Centralized Configuration Repository Efficiency for Secure Cloud Service Infrastructure Management. CEUR Workshop Proceedings 3991, 2025, 260–274 [https://ceur-ws.org/Vol-3991/paper19.pdf].
[14] Mitchell M.: Artificial Intelligence: A Guide for Thinking Humans. Penguin, 2019.
[15] Radford A., et. al.: Improving Language Understanding by Generative Pre-Training. 2018 [https://doi.org/10.48550/arXiv.1801.06146].
[16] Raiaan M. A. K.: A Review on Large Language Models: Architectures, Applications, Taxonomies, Open Issues and Challenges. IEEE Access 12, 2024, 26839–26874 [https://doi.org/10.1109/ACCESS.2024.3365742].
[17] Rothman D.: Transformers for Natural Language Processing: Build and Train State-of-the-Art NLP Models Using Transformers Architecture. Packt Publishing, 2021.
[18] Routray S. K., et. al.: Large Language Models (LLMs): Hypes and Realities. International Conference on Computer Science and Emerging Technologies (CSET) 2023, Bangalore, India, 1–6 [https://doi.org/10.1109/CSET58993.2023.10346621].
[19] Rzaieva S., et al.: Methods of Personal Data Protection in Retail: Practical Solutions. Cybersecurity Providing in Information and Telecommunication Systems 3991, 2025, 492–506 [https://ceur-ws.org/Vol-3991/paper35.pdf].
[20] Sabbatella A., et al.: Prompt Optimization in Large Language Models. Mathematics 12(6), 2024, 929 [https://doi.org/10.3390/math12060929].
[21] Shevchenko S., et al.: Protection of Information in Telecommunication Medical Systems based on a Risk-Oriented Approach. Cybersecurity Providing in Information and Telecommunication Systems 3421, 2023, 158–167 [https://ceur-ws.org/Vol-3421/paper16.pdf].
[22] Shevchuk D., et. al.: Designing Secured Services for Authentication, Authorization, and Accounting of Users. Cybersecurity Providing in Information and Telecommunication Systems 3550, 2023, 259–274 [https://ceur-ws.org/Vol-3550/short4.pdf].
[23] Vaswani A., et. al.: Attention is All You Need. 2017 [https://doi.org/10.48550/arXiv.1706.03762].
[24] Wolf T., et. al.: Transformers: State-of-the-Art Natural Language Processing. Association for Computational Linguistics, 2020, 38–45 [https://doi.org/10.18653/v1/2020.emnlp-demos.6].
[25] Yang X., et. al.: Exploring the Application of Large Language Models in Detecting and Protecting Personally Identifiable Information in Archival Data: A Comprehensive Study. IEEE International Conference on Big Data (BigData) 2023, Sorrento, Italy, 2116–2123 [https://doi.org/10.1109/BigData59044.2023.10386949].
[26] Advancing AI Through Fundamental and Applied Research [https://ai.meta.com/research].
[27] AICPA "SOC 2 – SOC for Service Organizations: Trust Services Criteria". [https://us.aicpa.org/interestareas/frc/assuranceadvisoryservices/soc-for-service-organizations].
[28] Amazon Bedrock – Automating Large-Scale, Fault-Tolerant Distributed Training in the Deep Learning Compiler Stack [https://aws.amazon.com/blogs/aws/amazon-bedrock-automating-large-scale-fault-tolerant-distributed-training-in-the-deep-learning-compiler-stack].
[29] Anthropic. Researching at the Frontier [https://www.anthropic.com/research].
[30] BERT by Google [https://ai.googleblog.com/2018/11/open-sourcing-bert-state-of-art-pre.html].
[31] Gelbstein E.: Is Audit Basics: The Domains of Data and Information Audits, 2016 [https://www.isaca.org/resources/isaca-journal/issues/2016/volume-6/is-audit-basics-the-domains-of-data-and-information-audits].
[32] GPT- by OpenAI [https://platform.openai.com/docs/models/gpt-3.5-turbo?utm_source=chatgpt.com].
[33] Mattsson U.: Practical Data Security and Privacy for GDPR and CCPA, 2020. [https://www.isaca.org/resources/isaca-journal/issues/2020/volume-3/practical-data-security-and-privacy-for-gdpr-and-ccpa].
[34] Open AI [https://openai.com/index/teaching-with-ai].

M.Sc. Oleh Deineka
e-mail: deinekaoleg.86@gmail.com

Postgraduate of Cyber Security Department of Information Protection, Lviv Polytechnic National University, Lviv, Ukraine.

Research interests: big data, data governance, artificial intelligence, AI agents, large language models, security standards, cybersecurity.

https://orcid.org/0009-0005-9156-3339

Ph.D. Oleh Harasymchuk
e-mail: oleh.i.harasymchuk@lpnu.ua

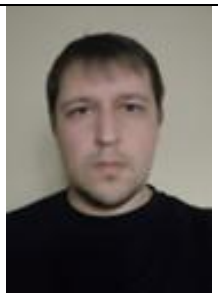
Ph.D., associate professor at the Department of Information Protection, Lviv Polytechnic National University, Lviv, Ukraine.
Research interests: cybersecurity, pseudo-random number generators, large language models, security standards, information protection systems, authentication and authorized access systems, database and knowledge systems.

https://orcid.org/0000-0002-8742-8872

Ph.D. Andrii Partyka
e-mail: andrii.i.partyka@lpnu.ua

Ph.D., associate professor at the Department of Information Protection, Lviv Polytechnic National University, Lviv, Ukraine.
Research interests: cybersecurity, cloud technology, ethical hacking, security isolation, DevSecOps, SDLC, AI security, risk management.

https://orcid.org/0000-0003-3037-8373



Ph.D. Yurii Dreis
e-mail: y.dreis@mu.edu.ua

Ph.D., associate professor at the Department of Analytics System and Information Technology, Mariupol State University, Kyiv, Ukraine.
Research interests: information security, protection information with limited accesses, consequence assessment, risk analysis.

https://orcid.org/0000-0003-2699-1597

Ph.D. Yuliia Khokhlachova
e-mail: y.khokhlachova@knute.edu.ua

Ph.D., professor of the Department of Software Engineering and Cybersecurity, State University of Trade and Economics, Kyiv, Ukraine.
Research interests: information security, vulnerability assessment, assessment of the cybersecurity status of information systems, security of cyber-physical systems.

https://orcid.org/0000-0002-0787-5112

Ph.D. Yuriy Pepa
e-mail: yurka14@ukr.net

Ph.D., professor of the Department of Technical Systems of Cybersecurity, State University of Information and Communication Technologies, Kyiv, Ukraine.
Research interests: cybersecurity, telecommunication systems, intelligent robotic systems, decision-making systems, radio electronic devices, antenna technology.

https://orcid.org/0000-0003-2073-1364