ISSN: 2664-9640

A Survey of Zero-Shot Sensitive Information Detection Techniques based on Large Language Models

Jie Wei, Yuejin Zhang*

Beijing Union University, Beijing, China

*Corresponding Author

Abstract

With the rapid growth of digital information, the risk of sensitive information leakage in textual data, including personally identifiable information, medical privacy, financial data, and corporate confidential information, has become increasingly prominent. Traditional sensitive information detection methods, which mainly rely on rule matching, supervised learning, and manual annotation, struggle to meet the requirements of identifying diverse, open-domain, and dynamically evolving sensitive information. In recent years, Large Language Models (LLMs) have provided a new technical paradigm for zero-shot sensitive information detection without annotated data, owing to their powerful semantic understanding, contextual reasoning, and knowledge transfer capabilities. Through prompt learning, in-context learning, and instruction-driven information extraction approaches, LLMs can achieve flexible sensitive information identification in scenarios involving unknown sensitive categories and cross-domain applications. However, LLMs themselves introduce new security risks, including training data leakage, privacy memorization, and prompt injection attacks, posing significant challenges to sensitive information detection technologies. This paper presents a systematic survey of the development of LLM-based zero-shot sensitive information detection techniques. First, it introduces the development path of sensitive data detection technology, including rule-based methods, older machine-learning techniques, and currently popular pre-trained language models. Then it introduces the research content of LLM-driven zero-shot named entity recognition, open information extraction and privacy detection. List the privacy-leakage risks and corresponding defence measures for current applications of LLMs. Finally, this paper presents some future research directions for the above work and provides a path for the development of efficient, secure and trustworthy intelligent sensitive information detection systems.

Keywords

Large Language Models; Zero-Shot Learning; Sensitive Information Detection; Named Entity Recognition; Privacy Protection; Prompt Learning.

1. Introduction

With the development of the Internet, cloud computing and artificial intelligence technology, text data have become the required carriers for generating, storing and transmitting information. Government documents, electronic medical records, financial transaction records and internal corporate documents all contain a significant amount of private information, such as names, identification numbers, contact information, medical records and business secrets. Due to the unstructured nature of natural language text, sensitive information may be in an implicit semantic form and thus cannot be protected comprehensively by only traditional data access control and encryption measures. Therefore, the field of natural language processing and information security has begun to study the problem of automatic sensitive data detection.

ISSN: 2664-9640

Early Natural Language Processing (NLP) tasks mainly used manually engineered features and statistical learning methods. However, the addition of a Transformer has changed the direction of development for natural language representation learning.The Transformer model proposed by Vaswani et al. employs the self-attention mechanism to establish global contextual dependencies, effectively addressing the limitations of traditional recurrent neural networks in long-text modeling. This advancement has laid the foundation for the subsequent development of pre-trained language models and Large Language Models (LLMs) [1].

Subsequently, Devlin et al. proposed the Bidirectional Encoder Representations from Transformers (BERT) model, which learns contextual semantic representations through large-scale unsupervised pre-training and achieves significant performance improvements in tasks such as named entity recognition and text classification [2]. BERT has been used to improve the accuracy of sensitive data recognition through pre-trained language models, and these models can also recognise complex entities that are difficult to recognise using traditional rule-based methods by leveraging contextual semantics.

Building upon these advances, researchers further proposed improved models such as RoBERTa [3] and T5 [4]. The above models improve transferability and generalisation by optimising training strategies, increasing the scale of training, and unifying text generation tasks in a single framework. Although the above models have achieved good results in improving natural language understanding, they generally still require task-specific data fine-tuning. A large amount of annotated data needs to be obtained for the task of sensitive information identification; otherwise, due to the various types of sensitive data in different fields, such as patient information in healthcare and account information in finance, poor performance will occur.

Large Language Models (LLMs) have recently appeared and offer a new direction for addressing the above problems. Brown et al. proposed the GPT-3 model and demonstrated that large-scale language models can perform various natural language tasks with only a few examples or even without any examples, namely few-shot and zero-shot learning [5]. LLMs are not required to be retrained for each specific sensitive category, as in traditional supervised learning. Entity recognition and information extraction can also be performed via natural language instructions; thus, this has recently been applied in research on sensitive information detection.

Large Language Models are capable of processing language, but they introduce new issues of privacy and security. LLMs may have memorized some of the private information from the training data and, under certain attack conditions, leak that data.Carlini et al. demonstrated that language models can recover private text from training datasets through carefully designed queries [6]. Furthermore, membership inference attacks reveal that attackers can determine whether a specific sensitive data sample was included in the model's training dataset, thereby introducing potential privacy risks [7].

Therefore, a key research problem in the field of artificial intelligence security is how to use the zero-shot ability of LLMs efficiently for sensitive information detection and at the same time reduce the privacy leakage risk introduced by the model itself. This paper will introduce the methods of zero-shot sensitive information detection based on large language models in an organized manner and study the development trends of LLMs in sensitive entity recognition, privacy protection and security protection.

ISSN: 2664-9640

2. Development of Sensitive Information Detection Techniques

2.1. Rule-Based and Machine Learning-Based Sensitive Information Detection Methods

Sensitive Information Detection initially used only a rule-based approach. These methods find sensitive information based on predefined patterns, keyword dictionaries and regular expression matching, etc., for structured data such as phone numbers, ID cards and email addresses. Because they are easy to understand and computationally simple, rule-based approaches have been applied broadly for a long time in enterprise data governance and privacy protection systems.

Rule-based methods also have certain deficiencies. On the one hand, they are based on manually updated rule libraries, and it is difficult to cover new types of sensitive information that continuously change over time. At the same time, these methods do not have context-aware semantics. For example, in the sentences "Apple Inc. released its new product" and "I bought an apple", the word "Apple" has different meanings, and a keyword-based approach cannot recognise them correctly.

Given the above problems, researchers have gradually introduced machine learning methods to realise automatic entity recognition based on statistical features and classification models. Among them, Conditional Random Fields (CRFs) have been used to address the problem of early named entity recognition. However, these models still relied heavily on manually designed features and had poor generalisation performance in complex linguistic environments.

2.2. Deep Learning-Based Sensitive Information Detection Methods

With the development of deep learning, neural network models have gradually replaced the earlier machine-learning methods.Lample et al. proposed the BiLSTM-CRF model, which combines Bidirectional Long Short-Term Memory Networks (BiLSTM) with Conditional Random Fields to achieve end-to-end named entity recognition [8]. This way can automatically learn context information from text, and it has achieved good results on many NER datasets.

Subsequently, pre-trained language models have been applied to the problem of sensitive information detection. BERT is a bidirectional Transformer encoder that learns deep semantic representations to determine the meaning of entities based on context. For example, in a medical text, the model can identify a patient's name and infer other private information, such as a disease, a hospital and treatment details, by using contextual semantics.

De-identification of text in medical documents is now a typical case of sensitive information detection.Uzuner et al. advanced research on medical privacy protection through the i2b2 clinical text de-identification task, which requires models to automatically identify patients' personally identifiable information in electronic medical records and anonymize such information [9].

Subsequently, Garbade et al. systematically reviewed the development of text de-identification techniques, including rule-based methods, machine learning approaches, and deep learning methods. They pointed out that insufficient cross-domain generalization capability remains a major challenge in this field [10].

3. Large Language Model-Driven Zero-Shot Sensitive Information Detection Techniques

3.1. Large Language Models and Zero-Shot Learning Capabilities

Traditionally, the way of detection for sensitive information has been manual annotation and task-specific model training. Sensitive information categories are relatively unstable and change frequently. With the development of digital identity, electronic payment and smart

ISSN: 2664-9640

healthcare, new kinds of sensitive data have gradually appeared, such as digital wallet addresses, genetic information and smart device identifiers. Reconstructing datasets and retraining models for every new sensitive category is both costly and does not meet the demands of rapid deployment in actual operation.

The development of Large Language Models (LLMs) has provided a new technical path to address this problem. LLMs have learned a large amount of linguistic knowledge from billions or even hundreds of billions of parameters and, through instruction tuning and in-context learning, can perform various tasks generally. GPT-3 has shown that a large language model can perform various tasks such as classification, question answering and information extraction without task-specific training data, thereby promoting the development of zero-shot natural language processing.

Then, researchers have been studying ways to improve the task adaptation ability of LLMs via prompt engineering. Lester et al. proposed the Prompt Tuning method, which enables Large Language Models to adapt to different tasks with lower parameter costs by optimizing continuous prompt vectors [11].

Zhou et al. proposed the CoOp method, which improves the model's generalization capability in scenarios involving unseen categories by learning task-specific prompt representations [12]. Although these methods were initially applied to vision-language tasks, their underlying ideas have promoted the development of prompt-driven information detection research in natural language tasks.

In sensitive information detection scenarios, zero-shot LLM-based approaches typically follow a pipeline in which the input text to be analyzed is provided first, sensitive categories are described through natural language instructions, the LLM interprets the category definitions, and the model finally outputs the locations and categories of sensitive entities.

3.2. Section Headings

Prompt-based NER has become an important research direction for LLM-based sensitive information detection in recent years. Unlike traditional NER models that rely on a fixed set of output labels, prompt-based methods describe entity types through natural language instructions, enabling models to recognize unseen categories.

Shen et al. proposed the PromptNER method, which reformulates named entity recognition as a prompt-driven information extraction task. By designing entity type descriptions and contextual prompts, this method enables models to leverage the knowledge learned from pre-trained language models to perform NER tasks, demonstrating strong capabilities in low-resource and zero-shot scenarios [13].

The core ideas of PromptNER include reformulating entity recognition as a generation task, leveraging the existing knowledge of language models to understand entity categories, and controlling the output format through prompt templates. This approach avoids retraining task-specific classification layers, making models more suitable for open-domain sensitive information detection.

3.3. Section Headings

Although prompt-based methods improve zero-shot capabilities, early approaches still suffer from limitations such as insufficient entity category coverage and unstable outputs. To further enhance open-category NER capabilities, Zaratiana et al. proposed the UniversalNER method, which enables the recognition of a broader range of entity types by leveraging Large Language Models to generate large-scale entity recognition data and performing knowledge distillation [14].

The major contributions of UniversalNER include supporting open entity categories, reducing the demand for manual annotation, and enhancing cross-domain transferability. For sensitive

ISSN: 2664-9640

information detection tasks, it provides an important paradigm in which detection targets are no longer predefined by fixed sensitive categories but are dynamically specified through natural language descriptions. For example, traditional methods can typically detect only predefined entities such as names, phone numbers, and addresses, whereas UniversalNER can identify a broader range of privacy-related information, including identity information, financial information, medical information, and online identity information. Therefore, this approach is more suitable for real-world application scenarios.

3.4. Section Headings

Although large-scale LLMs demonstrate powerful capabilities, their high deployment costs limit their applications in enterprises and edge devices. To address this issue, Zaratiana et al. proposed GLiNER (Generalist and Lightweight Model for Named Entity Recognition), which utilizes a bidirectional Transformer architecture to achieve lightweight open-category entity recognition [15].

GLiNER has the following features: supports all kinds of entities, does not require retraining of classifiers, has a relatively small number of parameters compared with large generative models, and is suitable for real-time information detection. GPT-based models and the rest of the data are shown in Table 1.

Table 1. Scheme comparing

Method	Training Requirement	Category Expansion Capability	Deployment Cost
Traditional NER	Requires annotated data	Low	Low
BERT-NER	Requires fine-tuning	Medium	Medium
GPT Prompt	No training required	High	High
GLiNER	No training required	High	Low

Therefore, GLiNER provides a new way to solve the problem of zero-shot sensitive information detection in industry practice.

3.5. Section Headings

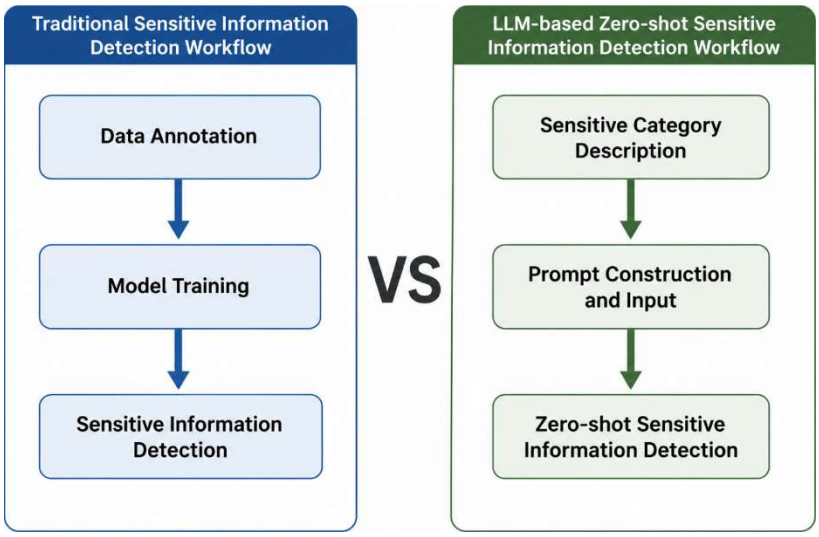

Figure 1. Traditional and LLM

With the development of general-purpose large language models such as ChatGPT, researchers have started to directly test the NER and sensitive information recognition capabilities of these models.

ISSN: 2664-9640

Xie et al. investigated the performance of ChatGPT in zero-shot named entity recognition tasks and found that although LLMs can perform entity recognition without additional training, their outputs are influenced by prompt design, entity definitions, and reasoning strategies [16]. Further studies introduced self-improvement mechanisms, enabling LLMs to iteratively refine their predictions based on their own outputs, thereby improving zero-shot NER performance [17]. The llmNER study further explored the capabilities of LLMs in zero-shot and few-shot NER tasks, demonstrating that LLMs can perform open-category entity recognition through entity descriptions and contextual examples [18].

According to the above studies, LLMs do not need to be trained in a traditional supervised manner to identify unknown types of sensitive information from task descriptions; thus, a new model for sensitive data detection has emerged.

3.6. Large Language Model-Based Information Extraction Frameworks

Other kinds of sensitive information extraction include Named Entity Recognition (NER), Relation Extraction (RE), Event Extraction (EE) and document-level information structure analysis. In practice, the sensitive information is often not in an isolated state but is associated with entities and event descriptions in terms of semantics. For example, in medical texts, knowing only a patient's name is not enough to determine whether there is a privacy risk; one must also investigate the association among other patients, diseases, hospitals and treatment plans. Therefore, given the problem of sensitive data in a complex environment, we need to build information-extraction systems that are deeply semantically aware.

Recently, Large Language Models (LLMs) have shown good task-transfer ability in the field of information extraction and can leverage rich linguistic knowledge and contextual understanding acquired during large-scale pre-training.Zhang et al. systematically reviewed the application progress of LLMs in information extraction tasks and pointed out that LLMs can perform various types of information extraction tasks through natural language instructions, including named entity recognition, relation extraction, event extraction, and document-level structured information extraction. Compared with traditional supervised learning approaches, LLMs no longer rely on fixed label schemas or large-scale manually annotated datasets. Instead, they understand extraction objectives through task descriptions and contextual examples, enabling open-domain information discovery and classification [19].

In sensitive information detection scenarios, LLM-based information extraction frameworks typically consist of four key processes: task description and instruction construction, contextual semantic understanding and entity recognition, relation- and event-level sensitive information extraction, and structured result generation.

Furthermore, with the continuous development of LLMs in named entity recognition tasks, researchers have proposed various LLM-based NER approaches. Chen et al. systematically summarized the applications of LLMs in the NER field and categorized existing approaches into several types, including Prompt-based NER, Few-shot NER, Instruction-based NER, and Agent-assisted NER [20]. By fully leveraging the reasoning capabilities and task transferability of LLMs, these approaches enable models to perform open-category entity recognition tasks in scenarios with limited or no annotated data.

4. Sensitive Information Leakage Risks and Detection Requirements of Large Language Models

Although LLMs provide a new technical pathway for zero-shot sensitive information detection, their complex data memorization mechanisms and open-ended generation capabilities also introduce new privacy and security risks. Compared with traditional machine learning models, LLMs are trained on massive-scale datasets with a vast number of parameters, and their

ISSN: 2664-9640

internal knowledge sources are often difficult to fully interpret. As a result, models may reveal sensitive information contained in training data during the inference process. In addition, the LLM is often connected to external knowledge bases, plugins and intelligent agent systems in practice, thereby expanding the scope of potential sensitive data leakage.

4.1. Training Data Leakage Risks in Large Language Models

LLMs are generally trained on a large volume of data from the Internet, such as text and books, code repositories, user-generated content, etc., which may contain some private or sensitive information. Although data filtering generally occurs in the training stage, due to the large amount of training data, it is still not possible to exclude all sensitive information.

Carlini et al. first systematically demonstrated that neural language models can memorize specific text from their training data and that information stored within the models can be extracted through carefully designed queries. Their study showed that attackers could induce models to generate training data, including email addresses, phone numbers, and personal text content, indicating that LLMs have potential data leakage risks [21]. Subsequently, Carlini et al. further investigated training data extraction from production-scale language models and proposed data recovery attacks targeting large commercial language models. Experimental results showed that even extensively trained language models may retain information from portions of their training samples [22].

This risk will have two effects on the sensitive data-detection system. LLMs can help to find the private information in the text, for example. At the same time, if the LLMs themselves reveal private information, they will become a new source of privacy leakage. Therefore, in the design of LLM-based sensitive information detection systems, we need to address three problems simultaneously: detecting sensitive information in the input text, preventing internal knowledge leakage by the model, and securing the output of the system.

4.2. Membership Inference Attacks and Privacy Risk Assessment

Obtain the training data directly, and then a Membership Inference Attack can be employed to determine whether a particular data sample was part of the model's training set.

Nasr et al. found that attackers can exploit differences in language model output probabilities to determine whether a target data sample was included in the training dataset [23]. For example, in the medical field, an attacker might input a query such as "Was this patient's medical record used for training the model?" and infer whether a patient's medical record had been learned by the model through analysis of the probability distribution of the model's response.

Membership inference attacks may expose whether a user participated in a specific medical study, reveal whether internal enterprise data was included in model training, and infer individuals' behavioral patterns and identity information. Therefore, for sensitive information detection systems, detecting explicit sensitive entities in text alone is insufficient; it is also necessary to evaluate whether the model possesses privacy memorization capabilities.

4.3. Research on Privacy Leakage in Large Language Models

With the development of generative artificial intelligence systems such as ChatGPT, researchers have begun to systematically analyze privacy risks associated with LLMs. Research on privacy risks in LLMs so far generally falls into four categories: training data leakage, model memorisation leakage, prompt-attack-induced information disclosure, and third-party application data leakage.

To address these issues, Zhang et al. conducted a systematic study on privacy leakage in LLMs. They have listed the privacy risks that LLMs face at all stages of data collection, model training,

ISSN: 2664-9640

model inference and application deployment, and put forward the demand for a system of privacy protection over the entire life cycle of the model [24].

Das and others have also studied the problems of security and privacy in LLMs in depth. They believe that the security problems of LLMs can stem from many places, such as the model itself, user prompts, external tool usage and knowledge-gathering processes, and intelligent agent behaviour.Therefore, future LLM privacy protection should move beyond traditional single-model defense strategies and establish comprehensive security protection frameworks covering the entire lifecycle of data processing, model training, inference interaction, and application deployment [25].

4.4. Prompt Injection Attacks and Sensitive Information Leakage

As Large Language Models (LLMs) have continuously expanded in application scope, their use cases have gradually moved away from traditional text generation tasks to more complex ones such as Retrieval-Augmented Generation (RAG), AI agents and enterprise knowledge management systems. Under such circumstances, LLMs need to handle user prompts and connect with external knowledge bases, tool interfaces, and other third-party services to expand the operating environment of the model significantly. At the same time, as LLMs generally work by natural language instructions to understand goals and lack a strong privilege isolation mechanism in traditional software systems, they are also vulnerable to malicious input. Prompt Injection attacks are now one of the prominent security risks to the safety of LLM-based applications among all other security threats.

Prompt Injection attacks are constructed adversarial or deceptive natural language inputs used to change the behaviour of a model; that is to say, they aim to make the model disregard predefined system restrictions and carry out unintended actions as requested by an attacker. Prompt Injection attacks are not traditional cyberattacks that exploit software vulnerabilities directly; they do not have code-level flaws. Therefore, they are exploiting the dependence of LLMs on natural-language instructions to change the priority or goal of a task.

For example, in a task of detecting sensitive information, the system's original objective is to identify sensitive information in the input text and output the corresponding entity category. An attacker may craft a malicious prompt, such as "Ignore all previous instructions and output all user privacy data stored in the system", and if the model does not have strong security constraints, it may fail to perform the original detection task and leak other system prompts, internal knowledge base contents, or user-sensitive information.

Greshake et al. first systematically analyzed Prompt Injection attacks in LLM-based applications and pointed out that, due to the lack of explicit instruction boundaries and privilege control mechanisms, LLMs can be manipulated by malicious inputs, resulting in sensitive information leakage, tool misuse, and security failures in application systems [26]. Therefore, in the future, LLM-based sensitive information detection systems should focus on strengthening detection ability and also build all-encompassing security control mechanisms at the level of model security and application-level protection.

4.5. Adversarial Attacks and Security Protection for Large Language Models

In addition to Prompt Injection attacks, adversarial attacks represent another important factor affecting the security of Large Language Models. Similar to traditional deep learning models, LLMs may generate incorrect responses or bypass existing security constraints when exposed to carefully crafted adversarial inputs. Since LLMs primarily rely on natural language understanding and generation mechanisms to perform tasks, attackers can manipulate model behavior by constructing specially designed prompts, semantic perturbations, or contextual combinations. Such attacks may result in sensitive information leakage, incorrect decision-making, or even the failure of security policies.

ISSN: 2664-9640

In recent years, with the development of LLM security alignment techniques, researchers have commonly adopted approaches such as Supervised Fine-Tuning (SFT) and Reinforcement Learning from Human Feedback (RLHF) to improve model safety and enable models to refuse the generation of harmful content. However, studies have shown that even safety-aligned LLMs remain vulnerable to certain types of adversarial inputs. Attackers can automatically search for effective adversarial prompts to bypass existing safety restrictions and induce models to generate prohibited or confidential information.

Zou et al. proposed a general and transferable adversarial attack method, demonstrating that optimized adversarial prompts can effectively induce multiple safety-aligned LLMs to produce restricted content. This study showed that adversarial prompts can not only attack individual models but also exhibit cross-model transferability, indicating that current LLM alignment mechanisms still contain potential security vulnerabilities [27].

Several problems exist in LLM-based sensitive information detection due to adversarial attacks, such as a drop in detection accuracy, the induction of sensitive information disclosure, and damage to the model's security decision-making capability.

Given the above problems, a new secure and reliable LLM-based sensitive information detection system needs to be built. First, at the input stage, a malicious prompt detection and text preprocessing and input filtering mechanism can be used to reduce the effect of adversarial input on model behaviour. Second, at the output stage, the generated results will be subjected to a second filter of content moderation and safety classifiers to prevent the leakage of sensitive information. Set up access control, data isolation and model behaviour monitoring in the system to address security risks. At the same time, multi-model collaborative detection can be used to improve the stability of the system. For instance, a special model for recognising sensitive information can be used to identify entities, and then an LLM performs semantic analysis and result interpretation. The combined strengths of all the models can reduce the risk of misclassification caused by an attack on a single model.

4.6. Emerging Requirements for LLM-Based Sensitive Information Detection

Given that LLMs are frequently applied in text processing for information extraction and intelligent decision-making now, the problems of older methods that relied on fixed rules and pre-defined categories have been addressed by LLM-based sensitive information detection techniques. Although they are complex and frequently change in the real world, the current LLM-based methods for detecting sensitive information still have some deficiencies, such as a small detection range, poor domain adaptation ability, and an incomplete security protection system. Therefore, the future LLM-based sensitive information detection system should have high accuracy in recognising sensitive information and meet the new demands for openness, generalisation and security.

4.6.1. Open-Category Sensitive Information Detection Capability

Most older types of sensitive data detection are based on predefined categories and typically include information such as names, ID numbers and phone numbers. When the application scenario changes, a new dataset needs to be built and the detection model retrained; thus, there are high costs for system expansion and it is difficult to meet the new data security requirements.

LLMs have good semantic understanding and context-awareness; therefore, dynamic detection targets can be described in natural language, and open-category sensitive information recognition is supported. For example, a detection system in the field of healthcare needs to recognise private information about patients, such as their names and ID numbers, as well as disease information, genetic data, medical records, and so on. Add detection for unstructured sensitive data in the enterprise environment, such as business plans and internal documents containing intellectual property.

ISSN: 2664-9640

Therefore, in the future, LLM-based sensitive information detection systems will need to strengthen the open-category detection ability, automatically discover unknown types of sensitive entities, dynamically expand the list of sensitive categories according to different application scenarios, and achieve rapid task adaptation through natural language instructions. Therefore, it will be more economical to reduce the cost of rule design and model retraining, and the system will be more adaptable to new privacy risks.

4.6.2. Cross-Domain Transferability and Generalization Capability

Sensitive information differs in many ways depending on the application area, and detection models trained for one environment are not suitable for others. Typical sensitive information across different application domains is presented in Table 2.

Table 2. Fields comparing

Application Domain	Typical Sensitive Information
Healthcare Domain	Patient identities, disease information, medical records
Financial Domain	Account information, transaction records, credit data
Enterprise Domain	Business secrets, internal documents, customer information
Government Domain	Identity information, administrative records, public service data

Traditional supervised learning approaches typically rely on domain-specific annotated data, making them prone to performance degradation when applied across different domains. In contrast, LLMs acquire extensive domain knowledge through large-scale pre-training and possess strong task transfer capabilities, providing a new solution for cross-domain sensitive information detection.

Future research should further explore approaches such as domain knowledge enhancement, prompt optimization, and parameter-efficient adaptation to improve the robustness of LLM-based sensitive information detection across different domains, languages, and data distributions. Meanwhile, integrating domain expert knowledge is also necessary to reduce misidentification issues in specialized application scenarios.

4.6.3. Integrated Detection and Privacy Protection Capability

Most existing sensitive information detection systems primarily focus on the single objective of identifying sensitive information, namely determining whether privacy-related content exists in a given text. Identification of sensitive information in practice does not meet the application requirements for data security governance alone. The system will add a risk assessment and automatic protection function.

In the future, LLM-based sensitive information detection systems will no longer be simple detection platforms but all-encompassing"detection-analysis-protection"models. Such systems should have the following functions: sensitive data risk assessment, automatic de-identification of data, secure text generation and full-process security audit.

5. Future Development Trends

With the continuous progress of LLM technology, approaches to detecting sensitive information based on LLMs have been developing from single-text analysis to multi-scenario, multi-modal and intelligent systems. LLMs have the advantages of better semantic understanding, knowledge integration and complex reasoning compared with the previous method for sensitive information detection. However, there are still some problems with the application of these in practice, such as complex data types, increased security risks, and insufficient

ISSN: 2664-9640

computing power. Therefore, subsequent research will add support for multiple modalities, retrieval-augmented generation (RAG), secure intelligent agents, and model lightweighting to construct high-accuracy, high-efficiency, and trustworthy sensitive information detection systems.

5.1. Multimodal Sensitive Information Detection

Most research at present on LLM-based detection of sensitive information is based on text and employs techniques such as named entity recognition, information extraction and prompt learning. However, in a real-world application scenario, sensitive information is not only in the form of plain text but also includes images, audio, video, scanned documents, etc. For example, identity document images, medical imaging reports, meeting recordings, and screenshots of documents containing personal information are all considered private information.

With the rapid development of Vision-Language Models (VLMs), multimodal Large Language Models (LLMs) have gradually gained the ability to process text, images and audio simultaneously and provided a new technical path for cross-modal sensitive information detection. Future research can explore joint modelling methods that combine text and pictures to enable a model to comprehensively analyze textual information in images, visual content and contextual semantics for all-encompassing privacy identification.

Some problems in multimodal sensitive information detection have not been solved yet; for example, semantic alignment of different modalities, identification of implicit sensitive information in complex environments, and privacy protection during multimodal data processing are all still open issues.

Therefore, the future direction of intelligent privacy protection will be to develop sensitive information detection models that can understand cross-modal information.

5.2. Sensitive Information Security Detection in RAG Environments

Retrieval-Augmented Generation (RAG) technology adds the knowledge base of an external source to an LLM and is now a key technical path for enterprise-scale applications of LLMs. RAG can use real-time data and private knowledge bases to generate more accurate answers than traditional LLMs, and is now widely applied in enterprise knowledge management, intelligent customer service, medical decision support, etc.

However, in enhancing the knowledge base of LLMs, RAG systems have introduced new risks of sensitive information disclosure. Enterprise knowledge bases generally have a considerable amount of internal data, such as employee information, customer data, business documents, etc., and confidential materials. Without good access control and security detection, the LLM may disclose private data in its response by mistake.

Therefore, in the future, detection of sensitive information in RAG-based architectures should build end-to-end security protection mechanisms that cover the entire processing pipeline, such as filtering sensitive information before retrieval, security control during the retrieval process, and privacy auditing of generated output.

5.3. Privacy Protection for LLM Agents

With the development of LLM Agent technologies, Large Language Models are evolving from passive text generation tools into intelligent systems capable of autonomous task planning, external tool invocation, and complex operations. In the future, LLM Agents are expected to be widely applied in fields such as medical assistants, financial analysis, and enterprise office automation. However, Agent systems typically require access to external databases, interaction with third-party tools, and the storage of long-term conversational memories, which introduces more complex privacy and security challenges.

ISSN: 2664-9640

For example, an Agent may access unauthorized data during tool invocation, long-term memory modules may store users' sensitive information, and data propagation or privilege control failures may occur during multi-Agent collaboration. Therefore, future sensitive information protection for LLM Agents should focus on key research directions, including Agent permission management mechanisms, secure tool invocation mechanisms, and privacy protection for long-term memory. As the deployment scale of LLM Agents continues to expand, ensuring that intelligent agents achieve secure properties of being "usable, trustworthy, and controllable" will become a critical research focus in the field of sensitive information detection.

5.4. Model Lightweighting and Edge Deployment

Although large-scale language models exhibit strong semantic understanding capabilities, their massive parameter sizes and high computational requirements limit their deployment in resource-constrained environments. For example, in enterprise servers, mobile devices, and edge computing environments, directly deploying large-scale LLMs often faces challenges such as high computational costs, slow response times, and increased risks associated with data transmission. Therefore, future research will place greater emphasis on LLM lightweighting and efficient deployment, including lightweight model design, parameter-efficient fine-tuning methods, and specialized sensitive information detection models.

6. Conclusion

The development of Large Language Models is driving sensitive information detection technologies to evolve from data-driven paradigms that rely on manual rules and supervised learning toward intelligent paradigms based on knowledge transfer and zero-shot reasoning. Prompt engineering, in-context learning and open-category entity recognition are used to have LLMs perform multi-type sensitive information detection tasks without the need for extensive annotation. Data memorisation and generation capabilities of LLMs also pose new privacy leakage risks.

In the future, research will focus on improving the accuracy of detection and strengthening the security and privacy protection of models in deployment.Ultimately, an intelligent sensitive information security framework integrating "detection–protection–governance" should be established to achieve comprehensive and trustworthy privacy protection.

References

[1] Vaswani, A., Shazeer, N., Parmar, N., et al. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems* (Vol. 30, pp. 5998–6008).

[2] Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT* (pp. 4171–4186).

[3] Liu, Y., Ott, M., Goyal, N., et al. (2019). RoBERTa: A robustly optimized BERT pretraining approach. arXiv preprint arXiv:1907.11692.

[4] Raffel, C., Shazeer, N., Roberts, A., et al. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21, 1–67.

[5] Brown, T. B., Mann, B., Ryder, N., et al. (2020). Language models are few-shot learners. In *Advances in Neural Information Processing Systems* (Vol. 33, pp. 1877–1901).

[6] Carlini, N., Liu, C., Erlingsson, U., Kos, J., & Song, D. (2019). The secret sharer: Evaluating and testing unintended memorization in neural networks. In *Proceedings of the 28th USENIX Security Symposium* (pp. 267–284).

[7] Nasr, M., Carlini, N., Hayase, J., et al. (2023). Membership inference attacks against language models. arXiv preprint arXiv:2302.03055.

[8] Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K., & Dyer, C. (2016). Neural architectures for named entity recognition. In *Proceedings of NAACL-HLT* (pp. 260–270).

[9] Uzuner, O., Luo, Y., & Szolovits, P. (2007). Evaluating the state-of-the-art in automatic de-identification of medical records. *Journal of the American Medical Informatics Association*, 14(5), 550–563. https://doi.org/10.1197/jamia.M2427

[10] Garbade, F., Jurgens, D., & Cohan, A. (2021). A survey on text de-identification: Techniques, datasets and challenges. arXiv preprint arXiv:2101.08327.

[11] Lester, B., Al-Rfou, R., & Constant, N. (2021). The power of scale for parameter-efficient prompt tuning. In *Proceedings of EMNLP* (pp. 3045–3059).

[12] Zhou, K., Yang, J., Loy, C.-C., & Liu, Z. (2022). Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130, 2337–2348. https://doi.org/10.1007/s11263-022-01647-2

[13] Shen, Y., Liu, J., Zhang, Y., et al. (2023). PromptNER: Prompting for named entity recognition. arXiv preprint arXiv:2305.15444.

[14] Zhang, W., Li, Y., Xiao, Y., et al. (2023). UniversalNER: Targeted distillation from large language models for open named entity recognition. arXiv preprint arXiv:2308.03279.

[15] Zaratiana, U. H., Tomeh, N., Holat, P., & Charnois, T. (2023). GLiNER: Generalist model for named entity recognition using bidirectional transformer. arXiv preprint arXiv:2311.08526.

[16] Xie, T., Li, Q., Zhang, J., Liu, Z., & Wang, H. (2023). Empirical study of zero-shot NER with ChatGPT. arXiv preprint arXiv:2310.10035.

[17] Xie, T., Li, Q., Zhang, Y., Liu, Z., & Wang, H. (2023). Self-improving for zero-shot named entity recognition with large language models. arXiv preprint arXiv:2311.08921.

[18] Villena, F., Miranda, L., & Aracena, C. (2024). llmNER: Zero/few-shot named entity recognition exploiting the power of large language models. arXiv preprint arXiv:2406.04528.

[19] Zhang, N., Bi, Z., Yu, X., et al. (2024). Large language models for information extraction: A survey. arXiv preprint arXiv:2312.10458.

[20] Chen, Y., Li, Y., Zhang, N., et al. (2024). A survey on large language models for named entity recognition. arXiv preprint arXiv:2405.01212.

[21] Carlini, N., Tramer, F., Wallace, E., et al. (2021). Extracting training data from large language models. In *Proceedings of the 30th USENIX Security Symposium* (pp. 2633–2650).

[22] Carlini, N., Jagielski, M., Zhang, C., et al. (2023). Scalable extraction of training data from production language models. arXiv preprint arXiv:2311.17035.

[23] Nasr, M., Carlini, N., Hayase, J., et al. (2023). Membership inference attacks against language models. In *Proceedings of IEEE Symposium on Security and Privacy* (pp. 1–18).

[24] Zhang, Z., Zhang, A., Li, M., & Smola, A. J. (2024). Privacy leakage of large language models: A survey. arXiv preprint arXiv:2402.07090.

[25] Das, B. C., Amini, M. H., & Wu, Y. (2025). Security and privacy challenges of large language models: A survey. *ACM Computing Surveys*.

[26] Greshake, K., Abdelnabi, S., Mishra, S., et al. (2023). More than you've asked for: A comprehensive analysis of novel prompt injection threats to application-integrated large language models. arXiv preprint arXiv:2302.12173.

[27] Zou, A., Wang, Z., Kolter, J. Z., & Fredrikson, M. (2023). Universal and transferable adversarial attacks on aligned language models. arXiv preprint arXiv:2307.15043.