The European Journal of Research and Development

journals.orclever.com/ejrnd

OR CLEVER Science & Research Group

Research Article

Secure Use of Artificial Intelligence with Artificial Intelligence Based Control

Fatih Mehmed Bilgin[1], Ali Aydin[2], Tugberk Zurnaci[3], Engin Bilici[4]

[1] Turkish Technology, Orcid ID: https://orcid.org/0009-0003-6244-5806, fatih.bilgin@thy.com
[2] Turkish Technology, Orcid ID: https://orcid.org/0009-0004-8973-8681, aydinali@ thy.com
[3] Turkish Technology, Orcid ID: https://orcid.org/0009-0006-8862-0934, t.zurnaci@ thy.com
[4] Turkish Technology, Orcid ID: https://orcid.org/0009-0003-4771-3024, enginbilici@ thy.com
Correspondence: fatih.bilgin@thy.com; Tel.: +90 531 961 7324

Received: 7 July 2025
Revised: 28 October 2025
2nd Revised: 26 November 2025
Accepted: 05 December 2025
Published: 12 December 2025

Reference: Bilgin, F. M., Aydin, A., Zurnaci, T., & Bilici, E. (2025). Secure use of artificial intelligence with artificial intelligence based control. *The European Journal of Research and Development, 5*(1), 465–468.

Abstract

Artificial intelligence applications have increased in recent years, providing benefits that increase the productivity of individuals and organizations. Individuals and organizations consult with AI tools in many areas, seek their assistance, and create value using these tools. However, the use of AI tools brings with it various security concerns. Open-source AIs have higher capabilities than those hosted on-premise environments. This encourages individuals and organizations to use open-source or paid versions. This study aims to identify and prevent unauthorized sharing of potentially sensitive data with third parties during paid or open-source use of AI tools using AI-assisted detection and prevention. The study, aims to use a combination of natural language processing, big data, and machine learning methods during detection processes, will also focus on customizing the models to be organizations or person-focused, in addition to general sensitive data, and increasing success in capturing sensitive data by fine-tuning the models. It will enable the implementation of blocking or masking processes after a successful detection process.

Keywords: Artificial Intelligence, Nature Language Processing, Named Entity Recognition, Sensitive Data Detection

1. Introduction

Large Language Models (LLMs) have become widely used across various domains, supporting tasks such as summarization, drafting, and automated reasoning. Despite their broad adoption, organizational governance structures regarding AI usage remain underdeveloped. According to Gartner, although AI is integrated into production workflows in many organizations, only a small proportion have established effective governance practices [1]. As a result, users may unknowingly transmit confidential or personal information when writing prompts to cloud-based AI systems.

Traditional Data Loss Prevention (DLP) mechanisms focus on structured communication channels, such as email or file transfers. These tools are not designed to handle conversational, unstructured natural language input. To address this gap, this study introduces an intermediary AI Data Loss Prevention Gateway (AI-DLP Gateway), which operates between the user and the AI platform. The gateway analyzes prompts in real time. Initially, rule-based detection methods are applied to identify structured sensitive data. If needed, a secondary contextual analysis is performed using a BERT-based NER model. Inputs identified as containing sensitive information are blocked, while safe prompts are allowed to proceed.

The system provides a practical means for organizations individuals to benefit from LLM capabilities without compromising data privacy.

2. Materials and Methods

2.1 Traditional DLP Approaches

Traditional DLP solutions operate by examining email traffic, endpoint activity, or network gateways to detect known patterns of sensitive data. However, these systems have difficulty capturing meaning when sensitive content is embedded in free-form natural language. In cloud-based AI interactions, the dynamic and user-generated nature of prompts makes these approaches insufficient [2].

2.2 Rule-Based and Regex Detection Systems

Rule-based methods, especially regular expressions, are effective for detecting structured data such as identity numbers or email addresses. Yet, slight formatting changes or paraphrasing can bypass these rules. Therefore, regex alone is not adequate for detecting sensitive data expressed in conversational contexts [3].

2.3 Contextual Representation Using BERT

BERT introduced a contextual representation framework that allows modeling the relationship between words based on surrounding text. This makes BERT-based models effective in identifying sensitive expressions that are not explicitly structured [4].

2.4 Named Entity Recognition

Named Entity Recognition (NER) models detect entities such as names, locations, organizations, and identifiers. In data protection contexts, NER enables identifying sensitive information that may not follow a fixed format. Studies show that transformer-based NER models are particularly effective for privacy-related tasks.

2.5 Machine Learning and Deep Learning Approaches

Machine learning models trained on large text datasets can identify variations of sensitive expressions that rule-based approaches miss. Hybrid approaches combining ML and rule-based detection have demonstrated strong performance in sensitive data obfuscation and anonymization tasks [5].

3. Results

The hybrid AI-DLP system developed in this study was evaluated using a dataset of 50,000 text samples. Of these, 60% contained personal or organizationally sensitive information, while 40% were neutral prompts.
Regex-only detection achieved high precision (93%) but lower recall (78%). The BERT-based NER model improved recall to 91% but caused a slight drop in precision due to contextual misclassification.
The combined hybrid model achieved:

Table 1: Performance Metrics of Regex, NER and Hybrid Detection Model

Method	Precision	Recall	F1 Score
Regex	93%	78%	84%
NER Model	79%	91%	82.5%
Hybrid Model	95%	93%	94%

Latency measurements averaged 172 ms, making the system suitable for real-time use. The hybrid architecture also demonstrated resilience against evasion attempts, such as spacing, symbolic substitutions, and text masking.

4. Discussion and Conclusion

The hybrid approach is effective because each layer addresses a limitation of the other. Rule-based detection provides strong performance for structured data, while the NER model captures context-dependent sensitive content. However, some challenges remain in detecting ambiguous categories such as workplace titles or informal address expressions. Future work may include ensemble models, threshold calibration, and model compression for performance enhancement [6].

This study presents a practical and adaptable gateway for detecting sensitive information in user prompts submitted to cloud-based AI systems. By integrating rule-based and contextual detection methods, the system reduces the risk of data leakage while maintaining usability. Future developments may expand the model to multilingual and multimodal inputs, including images and code segments, and extend detection to AI-generated outputs.

5. Acknowledge

We would like to express our appreciation to Turkish Technology for their support and valuable perspectives throughout the development of this study. The exchanges held with their technical teams, particularly in relation to data protection practices and implementation considerations, contributed to shaping the practical relevance of the work. All interpretations and conclusions presented in this paper remain solely those of the authors.

References

[1] Gartner. (2025). AI Governance Trends. Gartner Research.

[2] Gómez-Hidalgo, J. M., Martín-Abreu, J. M., Nieves, J., Santos, I., Brezo, F., & Bringas, P. G. (2010). *Data leak prevention through named entity recognition*. In 2010 IEEE Second International Conference on Social Computing(pp. 1129–1134).

[3] Mishra, K., Pagare, H., & Sharma, K. (2025). A hybrid rule-based NLP and machine learning approach for PII detection and anonymization in financial documents. *Scientific Reports, 15*, Article number: 4971.

[4] Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). *BERT: Pre-training of deep bidirectional transformers for language understanding*. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* (pp. 4171-4186). Association for Computational Linguistics.

[5] Singh, D., & Narayanan, S. (2025). *Unmasking the reality of PII masking models: Performance gaps and the call for accountability* (arXiv preprint).

[6] Velishetty, N. (2023). *Personal Identifiable Information (PII) Detection and Identification for Fintech with AI and Text Analytics* (Master's thesis, National College of Ireland, Dublin).