УДК 004.056

Л. М. Куперштейн

Б. А. Горенський

ВЕЛИКІ МОВНІ МОДЕЛІ ЯК ОСНОВА DLP-СИСТЕМИ

Вінницький національний технічний університет

Анотація

У роботі розглядається застосування великих мовних моделей (LLM) як аналітичного ядра систем запобігання витоку даних (DLP). Охарактеризовано класичні DLP-системи та розкрито концепцію LLM-орієнтованого підходу до захисту інформації. Розглянуто типологію LLM-орієнтованих DLP-систем за чотирма напрямками, визначено ключові переваги LLM-орієнтованих DLP порівняно з класичними системами, а також систематизовано їхні недоліки.

Ключові слова: DLP, LLM, захист даних, інформаційна безпека, семантичний аналіз, витік інформації, кібербезпека, промпт-ін'єкція, галюцинації, інсайдерські загрози.

Abstract

The paper considers the use of large language models (LLM) as the analytical core of data loss prevention (DLP) systems. Classical DLP systems are characterized and the concept of an LLM-oriented approach to information protection is revealed. The typology of LLM-oriented DLP systems in four areas is considered, the key advantages of LLM-oriented DLP compared to classical systems are identified, and their shortcomings are systematized.

Keywords: DLP, LLM, data protection, information security, semantic analysis, information leakage, cybersecurity, prompt injection, hallucination, insider threats.

Вступ

Проблема витоку конфіденційної інформації залишається одним із найжорсткіших викликів сучасної кібербезпеки. За даними IBM, середній глобальний збиток від одного витоку даних у 2025 році склав 4,44 млн дол. США, що свідчить про високу актуальність проблеми захисту інформації від витоків [1].

Традиційні засоби запобігання витоку даних функціонують за принципом зіставлення контенту з попередньо визначеними шаблонами: регулярними виразами, словниками, хеш-сумами документів тощо, що суттєво обмежує їх ефективність у роботі з неструктурованими та контекстуально-насиченими даними. Стрімкий розвиток великих мовних моделей (LLM – large language model) відкриває новий підхід до аналізу інформації. LLM – це нейронна мережа великого масштабу, навчена на об'ємних текстових даних за допомогою методів самоконтрольованого навчання, що демонструє здатності до розуміння природної мови, генерації тексту, класифікації та семантичного аналізу без необхідності явного програмування правил [2]. Інтеграція LLM у архітектуру DLP дозволяє перейти від методів на основі заздалегідь прописаних правил до семантичного аналізу контенту, що є значним поштовхом у розвитку засобів захисту інформації.

Метою даної роботи є аналіз концепції LLM-орієнтованої DLP, класифікація таких систем та визначення їх ключових переваг і недоліків порівняно з класичними підходами.

Основна частина

Класична DLP-система – це комплекс програмних і/або апаратних засобів, призначених для виявлення, моніторингу та блокування несанкціонованого копіювання або передавання конфіденційної інформації за межі інформаційної системи організації [3]. DLP-система на основі LLM – це програмне рішення з аналогічною метою по відношенню до класичної DLP-системи, проте основна відмінність полягає в тому, що для аналізу та ідентифікації чутливого контенту використовується мовна модель,

яка враховує контекст повідомлень або документів, що копіюються чи передаються користувачем [4].

За типом DLP-системи можна розглянути за чотирма напрямками, а саме за: рівнем розгортання, станом даних, що захищаються, методом аналізу даних та способом реагування системи на витоки чутливої інформації [5]. Рівень розгортання вказує на місце розгортання мовної моделі. Можна виокремити чотири рівні, а саме: рівень кінцевих точок, мережевий рівень, рівень хмарного провайдера та рівень приватного розгортання. На рівні кінцевих точок DLP-система розгортається безпосередньо на пристрої користувача у вигляді програмного забезпечення або розширення браузера. Зазвичай у цьому випадку використовуються локальні легкі мовні моделі, оскільки система призначена для одного користувача, а налаштування політик виконується самим користувачем. DLP-система на мережевому рівні розгортається на корпоративному проксі-сервері, який перехоплює весь вихідний трафік мережі. Для аналізу даних використовується мовна модель, яка може працювати безпосередньо на сервері або до якої сервер надсилає запити для перевірки перехоплених даних. DLP-система на рівні хмарного провайдера інтегрується з API хмарних сервісів через корпоративні акаунти, що дозволяє перевіряти дані, які користувачі завантажують або вивантажують, та надавати або обмежувати дозвіл на операції залежно від повноважень користувача та виду інформації. Останнім рівнем роботи DLP-системи є рівень приватного розгортання, в якому організація створює власну мовну модель всередині захищеного мережевого периметра, використовуючи власну інфраструктуру. Даний рівень характеризується тим, що дані не покидають межі мережі компанії, що зводить до мінімуму ризики витоку інформації.

LLM орієнтовані системи як і класичні мають здатність запобігати або повідомляти про витоки даних, що перебувають у спокої, у русі та у роботі. Дані у стані спокою зберігаються на фізичних носіях. У даному випадку DLP-система забезпечує безперервний або періодичний моніторинг збереженої інформації на фізичних та хмарних носіях із виконанням глибинного аналізу контенту та контекстну класифікацію файлів у сховищах. Мета системи полягає у виявленні конфіденційних даних у несанкціонованих локаціях, порушень прав доступу та автоматичне вжиття заходів (шифрування, видалення або переміщення в карантин) для мінімізації ризиків до моменту потенційного витоку. Крим того, DLP-система може здійснювати перевірку даних у русі, тобто тих, що передаються мережею в реальному часі. При перевірці даних у русі аналізується їх вміст і у разі виявлення чутливої інформації дані блокуються та/або повідомляється адміністратор про витік. Система може бути інтегрована у програмне забезпечення для контролю трафіку конкретного застосунку або здійснювати перехоплення всього трафіку через проксі-сервер, аналізуючи веб-трафік [6]. DLP-система також може контролювати дані, з якими користувач безпосередньо взаємодіє, наприклад з буфером обміну, скріншотами, файлом, що є відкритим в певний момент часу. В такому випадку LLM перевіряє, чи відповідає контекст використання даних наданим правам користувача.

DLP-системи можуть проводити контекстний та контентний аналіз. Контекстний аналіз класифікує за атрибутами файлу (автор, тип, розширення тощо) та контекстом передачі інформації, що визначає, хто і кому передає файли. LLM оцінює відповідність ситуації встановленим політикам безпеки [7]. Контентний аналіз розпізнає текстову інформацію, текст на зображеннях, тощо. Тут LLM демонструє ключову перевагу, якою є здатність до семантичного розуміння контенту [8].

DLP-системи можуть по різному реагувати на ініціювання витоків інформації, вони можуть використовувати як активні методи для усунення витоків так і пасивні. Якщо система запобігання витокам інформації використовує активні методи, вона блокуватиме несанкціоновані дії щодо передавання, поширення або використання файлів із конфіденційною інформацією. При пасивних методах LLM у складі DLP лише фіксує інцидент несанкціонованих чи підозрілих маніпуляцій щодо чутливої інформації шляхом реєстрації факту в системі журналювання із одночасним повідомленням адміністратора безпеки, не перериваючи бізнес-процеси.

Переваги LLM-орієнтованих DLP-систем порівняно з класичними системами полягають у:

– семантичному розумінні контексту, у якому LLM здатна визначити чутливість інформації не лише за ключовими словами як у класичних системах, а й за змістом повідомлення в цілому, виявляючи завуальований витік даних у тексті, що зовні виглядає нейтрально або структура якого є перебудованою для плутання DLP;

– зниженні рівня хибно-позитивних спрацьовувань, оскільки класичні DLP нерідко блокують легітимні документи через можливий збіг із заздалегідь прописаними шаблонами. Оцінка повного контексту дозволяє підвищити точність класифікації та зменшити операційне навантаження на аналітиків безпеки;

– підтримці багатомовного аналізу, тому що LLM опрацьовує текст різними мовами без необхідності розробляти окремі набори правил для кожної з них, що суттєво спрощує захист у міжнародних організаціях;

– адаптивності щодо нових загроз, так як LLM замість ручного оновлення правил достатньо перенавчити на нових прикладах інцидентів, що суттєво скорочує час реакції на нові вектори витоку даних [8];

– здатності одночасно обробляти контент різного типу, що переважає в гнучкості аналізу у порівнянні з класичними методами пошуку конфіденційної інформації за заздалегідь підготовленими патернами та правилами. Здатність працювати з різним контентом також висвітлює універсальність даного підходу;

– зниженні навантаження на аналітиків безпеки через можливість LLM виявляти інцидент та автоматично формувати його короткий опис, що саме було передано, яка категорія даних, який ступінь ризику, що суттєво пришвидшує розслідування інцидентів;

– виявленні інсайдерських загроз через поведінковий аналіз. LLM здатна аналізувати не лише окремий файл чи повідомлення, а й виявляти підозрілі патерни від потенційного порушника, які по одному виглядають легітимно та в сукупності свідчать про підготовку до витоку;

– ефективності роботи з неструктурованим контентом, через здатність LLM до опрацьовування текстів, транскриптів аудіо, зображень, де класичні підходи малоефективні.

Попри суттєві переваги над класичними, DLP-системам з LLM- ядром притаманні також і недоліки, а саме:

– висока обчислювальна вартість при розгортанні великих моделей вимагає значних апаратних ресурсів, особливо для локальних рішень рівня приватного розгортання;

– потреба у навчанні моделі у разі використання приватного рішення, на що будуть витрачені час та ресурси компанії;

– залежність від якості навчальних даних, яка виникає якщо модель навчена на недостатньо репрезентативних прикладах або застарілих даних, через що вона може систематично помилятися у класифікації певних типів конфіденційної інформації, специфічних для конкретної галузі чи організації;

– складність інтеграції із застарілими системами, через те, що більшість організацій використовують інфраструктуру, що формувалася роками впровадження процесів DLP на основі мовної моделі у таке середовище може потребувати значних змін в архітектурі мережі та додаткових витрат на інтеграцію;

– вразливість до промпт-ін’єкцій з допомогою яких зловмисник може вбудувати у текст документа або повідомлення спеціальні інструкції, які змусять LLM проігнорувати правила безпеки та сформувати рішення щодо дозволу витоку чутливого контенту [10];

– схильність до галюцинацій, яка характеризується тим, що LLM може ідентифікувати нейтральну інформацію як конфіденційну або навпаки пропускати реальний витік, що знижує надійність системи та вимагає обов’язкового людського контролю над рішеннями LLM [11];

– затримка при аналізі, при якій обробка одного запиту моделі може ускладнювати роботу в режимі реального часу для високонавантажених мережевих потоків на мережевому рівні;

– низький рівень пояснення рішень LLM, тому що не завжди рішення мовної моделі піддається інтерпретації, що ускладнює аудит та підтвердження відповідності регуляторним вимогам.

Висновки

Проведений аналіз свідчить, що інтеграція LLM у архітектуру DLP-систем є новим і перспективним напрямом розвитку засобів захисту інформації, який принципово змінює підхід до виявлення загроз витоку даних. Розглянута типологія DLP-систем дозволяє вибрати оптимальну конфігурацію системи залежно від масштабу організації, галузі та регуляторних вимог.

Ключова перевага поєднання LLM з DLP полягає у переході від лексичного до семантичного рівня аналізу контенту. Класичні DLP-системи здатні виявляти лише те, що явно визначено в правилах, а LLM натомість розуміють зміст повідомлення як ціле, що дозволяє виявляти приховані, завуальовані або нетипово сформульовані спроби передати конфіденційну інформацію. Суттєвою перевагою є також універсальність підходу, оскільки одна LLM здатна одночасно обробляти різні типи даних, тоді як класичні системи потребують окремих модулів і наборів правил для кожного типу контенту. Разом

із тим, впровадження LLM-орієнтованих DLP-систем пов'язане з низкою суттєвих обмежень, які потрібно враховувати при їх інтеграції в інфраструктуру компанії або на пристрій користувача.

Таким чином, LLM-орієнтовані DLP не є простим доповненням до існуючих рішень, а представляють іншу парадигму захисту даних, хоча вони і мають недоліки, проте це є новим напрямком розвитку кібербезпеки, який потрібно активно досліджувати і його практична цінність лише зростатиме разом із розвитком мовних моделей.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

1. IBM. Cost of a Data Breach Report 2025. – URL: https://www.ibm.com/reports/data-breach (дата звернення: 17.03.2026).
2. SAP. What is a Large Language Model (LLM)? – 2024. – URL: https://www.sap.com/ukraine/resources/what-is-large-language-model (дата звернення: 17.03.2026).
3. Що таке DLP-системи: види, можливості, порівняння рішень. – URL: https://netwave.ua/blog/shho-take-dlp-systemy-vydy-mozhlyvosti-porivnyannya-rishen/ (дата звернення: 17.03.2026).
4. Fried A. How LLMs are Revolutionizing Data Loss Prevention 2024. – URL: https://securityboulevard.com/2024/08/how-llms-are-revolutionizing-data-loss-prevention/ (дата звернення: 17.03.2026).
5. Süzen A. A., Ceylan O. Robustness Assessment of Data Loss Prevention (DLP) Software for Data Leakage against Different Data Types and Sources // International Journal of Education and Management Engineering. – 2025. – Vol. 15. – No. 2. – P. 1–10.
6. Куперштейн Л. Інформаційна технологія моніторингу безпеки даних програмного забезпечення/Л.Куперштейн, Г.Луцишин, М.Кренцін// Кібербезпека: освіта, наука, техніка. –2024. –No 3 (23). –С. 71–84
7. Syarova S., Toleva-Stoimenova S., Kirkov A., Petkov S., Traykov K. Data leakage prevention and detection in digital configurations: a survey // Environment. Technology. Resources. Proceedings of the 15th International Scientific and Practical Conference. – Rezekne Academy of Technologies, 2024. – Vol. 2. – P. 253–258.
8. Gupta K., Kush A. A Learning oriented DLP System based on Classification Model // arXiv preprint. – 2023.
9. Content Analysis Using Specific Natural Language Processing Methods for Big Data. – URL: https://www.mdpi.com/2079-9292/13/3/584 (дата звернення: 17.03.2026).
10. Клиш В. М., Куперштейн Л. М. Аналіз атак типу Prompt Injection на великі мовні моделі // Матеріали науково-технічної конференції підрозділів ВНТУ. – Вінниця, 2025. – URL: https://ir.lib.vntu.edu.ua/bitstream/handle/123456789/48606/24458.pdf (дата звернення: 17.03.2026).
11. Survey and analysis of hallucinations in large language models: attribution to prompting strategies or model behavior. – URL: https://www.frontiersin.org/journals/artificial-intelligence/articles/10.3389/frai.2025.1622292/full (дата звернення 19.03.2026)

Куперштейн Леонід Михайлович – к.т.н., доцент кафедри захисту інформації, Вінницький національний технічний університет, м. Вінниця email: kupershtein@vntu.edu.ua.

Kupershtein Leonid M. – PhD, Associated Professor of Information Protection Chair, Vinnytsia National Technical University, Vinnytsia, email: kupershtein@vntu.edu.ua

Горенський Богдан Анатолійович – студент групи 1БКС-22б, Факультет інформаційних технологій та комп'ютерної інженерії, Вінницький національний технічний університет, м. Вінниця, Україна, e-mail: h.b.a.94855@gmail.com

Horenskyi Bohdan Anatoliyovych – student of group 1BKS-22b, Faculty of Information Technologies and Computer Engineering, Vinnytsia National Technical University, Vinnytsia, Ukraine, e-mail: h.b.a.94855@gmail.com