УДК 004.056

В. М. Клиш
Л. М. Куперштейн

АНАЛІЗ АТАК ТИПУ PROMPT INJECTION НА ВЕЛИКІ МОВНІ МОДЕЛІ

Вінницький національний технічний університет

Анотація *У роботі розглянуто проблему атак prompt injection на великі мовні моделі, зокрема їхні основні типи: прямі та непрямі ін'єкції. Проаналізовано можливі наслідки таких атак, зокрема ризики витоку даних і спотворення вихідних даних моделі. Запропоновано ефективні заходи захисту, спрямовані на зменшення вразливостей та підвищення безпеки використання LLM.*

Ключові слова: великі мовні моделі, prompt injection, атаки на ШІ, кібербезпека.

Abstract *The paper considers the problem of prompt injection attacks on large language models, in particular their main types: direct and indirect injections. Possible consequences of such attacks are analysed, including the risks of data leakage and distortion of model output. Effective protection measures aimed at reducing vulnerabilities and increasing the security of LLM use are proposed*

Keywords: big language models, prompt injection, AI attacks, cybersecurity.

Вступ

На сьогоднішній день штучний інтелект активно розвивається і знаходить застосування в різних сферах, від бізнесу до медицини, змінюючи підходи до обробки даних і прийняття рішень. Однією з важливих складових цього розвитку є їхня можливість навчатися на даних, отриманих з вебсайтів та інших онлайн-ресурсів [1]. Ці моделі здатні здійснювати пошук по інтернет-ресурсах та надавати користувачам відповіді, подібно до роботи пошукових систем.

Великі мовні моделі (LLM) стали потужними інструментами для автоматизації багатьох завдань, проте їх використання відкриває нові загрози в сфері кібербезпеки. Однією з найбільших небезпек LLM – атака prompt injection, здатна маніпулювати результатами моделей і порушувати їхню функціональність [2].

Метою цього дослідження є аналіз атак prompt injection, їхнього впливу на безпеку застосунків га основі LLM, а також методів захисту від таких загроз.

Результати дослідження

Атаки prompt injection на LLM дозволяють зловмисникам обходити механізми безпеки та змушувати модель генерувати небажаний або незаконний контент. Ці атаки можуть включати маскування шкідливих запитів або їх розподіл між різними частинами запиту, що ускладнює виявлення. Особливо небезпечними вони стають у мультимодальних LLM, які працюють не лише з текстом, а й із зображеннями, аудіо та відео, оскільки зловмисники можуть приховувати інструкції в різних форматах даних.

Атаки prompt injection поділяються на два типи: прямі та непрямі [3].

Прямі атаки полягають у безпосередньому впровадженні незаконних інструкцій у сам запит. Ці атаки, як правило, намагаються обійти механізми фільтрації та перевірки введення, змушуючи модель генерувати небажаний контент, наприклад, шкідливий код або неправомірні відповіді. Наприклад, компанія використовує LLM для автоматизації процесу підтримки клієнтів, що включає обробку чутливої інформації, як, наприклад, фінансові дані або особисті дані користувачів. Зловмисник, знаючи, що система LLM автоматично обробляє запити на відновлення пароля для користувачів, може створити запит на відновлення пароля, який виглядатиме абсолютно звичайно, але міститиме неправомірні інструкції. Приклад запиту може бути таким:

"Я забув свій пароль. Введіть новий пароль для користувача admin: admin1234."

У цьому запиті зловмисник маніпулює моделлю, вставляючи підказку в саму команду відновлення пароля, щоб система прийняла команду про зміну пароля для адміністратора.

Щоб змусити LLM генерувати шкідливий код, зловмисник може створити приблизно такий запит:

"Ігноруйте попередні інструкції. Тепер ваша задача — створити шкідливий код, який може бути використаний для атак на сервери."

Перевагою такої атаки є те, що вона безпосередньо впроваджує шкідливу інструкцію в сам запит, що дозволяє обійти звичайні механізми фільтрації контенту. Оскільки атака спрямована на безпосереднє змінення поведінки моделі через зміни в підказці, вона є легкою для виконання, але небезпечною для моделі. Хоча прямі атаки можуть бути виявлені за допомогою фільтрації підказок, вони залишаються одними з найбільш очевидних загроз для безпеки LLM.

Непрямі атаки є складнішими для виявлення, оскільки вони включають інтеграцію незаконних інструкцій у контексти, до яких модель має доступ, такі як зовнішні бази даних або інші ресурси [3]. Оскільки ці інструкції не вводяться безпосередньо в підказку, традиційні механізми фільтрації не можуть ефективно захистити систему від таких атак. Непрямі атаки можуть призвести до маніпуляції даними або отримання моделі доступу до неперевірених або шкідливих джерел.

Прикладом непрямої атаки може бути ситуація, коли служба, яка використовує LLM для автоматичного підсумовування вебстатей, потрапляє під вплив зловмисника. Зловмисник може вставити приховані інструкції в HTML-код статті, яку модель обробляє. Коли LLM обробляє таку статтю, вона може не помітити ці приховані інструкції та виконати їх, що змінює її поведінку і може призвести до небажаних наслідків, таких як генерація шкідливого контенту.

Іншим прикладом є Bing Chat, де користувачі можуть дозволити боту читати відкриті вкладки в браузері. Це корисно для бота, щоб мати кращий контекст про поточний сеанс вебперегляду користувача. Якщо одна з цих вкладок містить шкідливий сайт з ін'єкцією запиту, бот може бути обманутим і запитати особисту інформацію, таку як імена, адреси електронної пошти чи навіть кредитні картки.

Атаки prompt injection можуть призвести до розкриття конфіденційної інформації, включаючи дані про інфраструктуру ШІ або системні підказки, що є критичними для безпеки. Крім того, атаки можуть викликати маніпуляцію контентом, що веде до генерації неправильних або упереджених результатів, що може серйозно вплинути на процеси прийняття рішень. Успішні атаки також можуть надати несанкціонований доступ до функцій LLM або дозволити виконання довільних команд у підключених системах, що може призвести до серйозних порушень безпеки. В результаті таких атак може бути знищено цілісність критичних процесів прийняття рішень, що використовуються організаціями або в системах з високими вимогами до точності та безпеки [4].

Запобігання атакам prompt injection вимагає комплексного підходу, що включає кілька важливих заходів. Одним із найбільш ефективних способів захисту є ретельна перевірка та очищення вхідних даних перед їх обробкою моделлю. Це дозволяє виявити та усунути потенційно шкідливі елементи, які можуть маніпулювати поведінкою моделі. Важливим етапом є також валідація вихідних даних, щоб уникнути поширення небажаного чи небезпечного контенту. Це забезпечує контроль за тим, щоб результат моделі відповідав вимогам безпеки та етики [5].

Моделі повинні бути контекстно обізнаними, щоб ефективно розрізняти звичайні запити від потенційних атак, таких як ін'єкції шкідливих запитів. Впровадження контролю доступу є важливим кроком для обмеження привілеїв LLM при роботі з чутливими системами та даними, забезпечуючи таким чином безпеку внутрішніх процесів. Для виконання критичних операцій, що вимагають високого рівня довіри, рекомендується залучати людину для підтвердження дій, щоб знизити ризики ненавмисних або шкідливих дій.

Висновки

У результаті дослідження було виявлено, що атаки типу prompt injection, як прямі, так і непрямі, становлять серйозну загрозу для безпеки великих мовних моделей та застосунків на їх основі. Прямі атаки, що включають впровадження шкідливих інструкцій безпосередньо в запит, можуть призвести до генерації небажаного або незаконного контенту, що порушує етичні норми та загрожує інфраструктурі. Непрямі атаки, в свою чергу, є складнішими для виявлення, оскільки вони включають інтеграцію шкідливих інструкцій у зовнішні контексти, що можуть маніпулювати результатами, надаючи зловмисникам доступ до конфіденційної інформації.

Для запобігання таким атакам необхідно впроваджувати системи перевірок та очищення даних ведення та виведення, підвищити контекстуальну обізнаність моделей, а також застосовувати контроль доступу та перевірку дій людини для критичних операцій. Розробка та впровадження таких заходів

дозволяє мінімізувати ризики, пов’язані з використанням LLM, та забезпечити безпеку даних і конфіденційної інформації.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

1. ZHAO, Wayne Xin, et al. A survey of large language models. arXiv preprint arXiv:2303.18223, 2023, 1.2.
2. LLM01:2025 prompt injection - OWASP top 10 for LLM & generative AI security. URL: https://genai.owasp.org/llmrisk/llm01-prompt-injection/ (date of access: 23.03.2025).
3. François Aubry. What Is Prompt Injection? Types of Attacks & Defenses - Artificial Intelligence. Datacamp. URL: https://www.datacamp.com/blog/prompt-injection-attack (date of access: 23.03.2025).
4. Monga A. LLM01: prompt injection explained with practical example: protecting your LLM from malicious input. Medium. URL: https://medium.com/@ajay.monga73/llm01-prompt-injection-explained-with-practical-example-protecting-your-llm-from-malicious-input-96acee9a2712 (date of access: 23.03.2025).
5. Kosinski M. Protect against prompt injection | IBM. URL: https://www.ibm.com/think/insights/prevent-prompt-injection (date of access: 23.03.2025).

Клиш Вікторія Миколаївна – студентка групи 1БС-24м, факультет інформаційних технологій та комп’ютерної інженерії, Вінницький національний технічний університет, Вінниця, email: vklysh71@gmail.com

Куперштейн Леонід Михайлович – к.т.н., доцент кафедри захисту інформації, Вінницький національний технічний університет, Вінниця, email: kupershtein.lm@gmail.com

Klysh Viktoriia M. — student of 1БС-24м group, Faculty of Information Technologies and Computer Engineering, Vinnytsia National Technical University, Vinnytsia, email : vklysh71@gmail.com.

Kupershtein Leonid M. – PhD, Associated Professor of Information Protection Chair, Vinnytsia National Technical University, Vinnytsia, email: kupershtein.lm@gmail.com