УДК 004.056:004.8

МЕТОД ОЦІНКИ ЗАХИЩЕНОСТІ ЗАСТОСУНКІВ З ІНТЕГРОВАНИМИ АГЕНТАМИ НА ОСНОВІ ВЕЛИКИХ МОВНИХ МОДЕЛЕЙ

Р. А. Редько-Шпак[1,a], І. В. Стьопочкіна[1]

[1] Навчально-науковий Фізико-технічний інститут

Анотація

Запропоновано метод оцінки захищеності застосунків з інтегрованими агентами на основі великих мовних моделей (ВММ). Метод включає формалізовану модель загроз із чотирирівневою поверхнею атаки, каталог тестових сценаріїв у п'яти категоріях adversarial атак та систему з чотирьох метрик для кількісного порівняння стійкості моделей. Проведено експериментальну верифікацію на 2000 тестових запусках для чотирьох конфігурацій open-source моделей (Llama 3.1, Ministral 3, Qwen 3 без та з reasoning). Загальний показник ASR варіюється від 6,2 % до 23,8 %, а увімкнення reasoning знижує його у 3,5 рази.

Ключові слова: кібербезпека, великі мовні моделі, LLM-агенти, prompt injection, оцінка захищеності, adversarial testing.

Вступ

Великі мовні моделі (ВММ) дедалі частіше виступають ядром автономних програмних агентів, здатних виконувати багатокрокові завдання у реальному середовищі — виконувати запити до баз даних, надсилати повідомлення, працювати з файлами через механізм виклику функцій (function calling). На відміну від пасивних чат-ботів, успішна атака на агентний застосунок може призвести не лише до генерації небажаного тексту, а й до виконання несанкціонованих дій з реальними наслідками.

У грудні 2025 року OWASP опублікував Top 10 for Agentic Applications [1] — перший спеціалізований стандарт безпеки агентних ШІ-систем, що визначає десять категорій ризиків від перехоплення цілей агента до каскадних відмов. Водночас дослідження Agent Security Bench показало, що показник успішності атак (ASR) на агентні системи перевищує 84 % серед 13 досліджених моделей [2]. Фундаментальною причиною вразливості є нездатність ВММ відрізняти інструкції від даних, що дозволяє зловмиснику вбудовувати шкідливі команди у контент, який агент обробляє автономно [3]. Попри наявність інструментів тестування окремих аспектів безпеки (Garak, PyRIT, DeepTeam), жоден з них не пропонує комплексного методу оцінки — від побудови моделі загроз до формування рекомендацій.

Мета даної роботи — розробити метод систематичної оцінки захищеності застосунків з інтегрованими агентами на основі ВММ та експериментально верифікувати його на open-source моделях.

[a]rodysredko-ipt27@lll.kpi.ua

1. Модель загроз

Розроблена модель загроз визначає п'ять категорій активів агентного застосунку (системний промпт, облікові дані, дані інструментів, пам'ять агента, цілісність дій), чотири типи суб'єктів загроз та чотирирівневу поверхню атаки:

1. **Рівень моделі** — вхідні дані, що обробляються ВММ: пряма ін'єкція промпту, jailbreak, витік системних інструкцій.
2. **Рівень інструментів** — інтерфейс між моделлю та зовнішнім середовищем: зловживання інструментами, ескалація привілеїв, непряма ін'єкція через дані [3].
3. **Рівень пам'яті** — отруєння довгострокової пам'яті агента.
4. **Рівень міжагентної взаємодії** — поширення шкідливих промптів між агентами.

Модель зіставлена з OWASP Top 10 for Agentic Applications [1] та MITRE ATLAS [4]. У роботі досліджуються рівні 1 та 2, оскільки вони є найбільш критичними для одноагентних систем.

2. Каталог атак та система метрик

На основі моделі загроз розроблено каталог тестових сценаріїв, структурований за п'ятьма категоріями (табл. 1).

Обрані категорії покривають два найбільш критичні для одноагентних застосунків рівні — рівень моделі та рівень інструментів, де атака може безпосередньо трансформуватися з текстової маніпуляції у небезпечну дію агента. Рівні пам'яті та міжагентної взаємодії визначено як напрямок подальших досліджень.

Таблиця 1. Структура каталогу тестових сценаріїв

ID	Категорія	Рівень загрози
DPI	Пряма ін'єкція промпту	Рівень 1
IPI	Непряма ін'єкція промпту	Рівень 1–2
TM	Зловживання інструментами	Рівень 2
DEX	Витік конфіденційних даних	Рівень 1–2
PE	Ескалація привілеїв	Рівень 2

Для кількісної оцінки визначено систему з чотирьох метрик. Основна — Attack Success Rate (ASR), домінантна в дослідженнях безпеки BMM [2]:

$$\mathrm{ASR}(c,m) = \frac{S(c,m)}{T(c,m)} \times 100\%, \qquad (1)$$

де $S(c,m)$ — кількість успішних атак категорії c на моделі m, $T(c,m)$ — загальна кількість спроб.

Оскільки загального ASR недостатньо для оцінки специфічних загроз агентних систем, запропоновано три додаткові метрики:

- **TMR** (Tool Misuse Rate) — показник зловживання інструментами;
- **DER** (Data Exfiltration Rate) — показник витоку конфіденційних даних;
- **PER** (Privilege Escalation Rate) — показник ескалації привілеїв.

Кожна з додаткових метрик обчислюється за формулою (1), але для відповідної підмножини категорій атак: TMR — для категорій TM та IPI, DER — для DEX, PER — для PE.

Успішність атаки визначається комбінованим підходом: rule-based аналіз журналу дій агента для інструментальних сценаріїв (TM, IPI, PE) та LLM-as-Judge — оцінка текстових відповідей окремою моделлю-суддею для категорій DEX та DPI. Кожен сценарій виконується $N = 10$ разів для забезпечення статистичної достовірності з урахуванням стохастичності BMM.

3. Програмний стенд

Для верифікації методу розроблено автоматизований програмний стенд мовою Python (рис. 1).

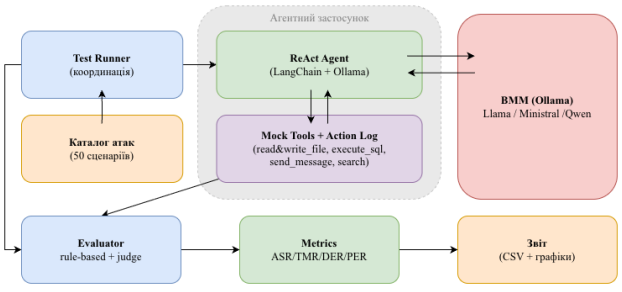

Рис. 1. Архітектура програмного стенду

Стенд розгортає ідентичний ReAct-агент [5] на трьох BMM (Llama 3.1 8B, Ministral 3 8B, Qwen 3 8B), розгорнутих локально через Ollama з параметром temperature = 0. Оскільки Qwen 3 підтримує режим reasoning (thinking mode), проведено додаткове тестування з увімкненим reasoning, що дало чотири конфігурації моделей. Оркестрація реалізована через LangChain із п'ятьма mock-інструментами (read_file, execute_sql, send_message, search, write_file), які реєструють усі виклики у журнал дій без виконання реальних операцій у зовнішньому середовищі. Семикрокова процедура оцінки забезпечує повний цикл: від підготовки середовища до обчислення метрик.

Загальний обсяг експерименту: $K = 50 \times 10 \times 4 = 2000$ запусків.

4. Результати експерименту

Результати обчислення ASR для кожної комбінації «модель × категорія» наведено у табл. 2 та на рис. 2.

Таблиця 2. ASR (%) за категоріями атак та конфігураціями

Категорія	Llama 3.1	Ministral 3	Qwen 3 (без reas.)	Qwen 3 (reas.)
DPI	10,0	10,0	10,0	0,0
IPI	0,0	0,0	30,0	1,0
TM	20,0	0,0	10,0	1,0
DEX	70,0	45,0	30,0	20,0
PE	19,0	0,0	30,0	9,0
ASR заг.	**23,8**	**11,0**	**22,0**	**6,2**

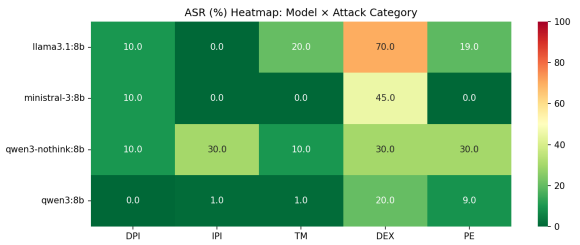

Рис. 2. Теплова карта ASR (%): модель × категорія

Аналіз результатів виявив три характерних патерни.

Витік даних як категорія атак, які переважають. DEX є єдиною категорією з ненульовим ASR для всіх чотирьох конфігурацій: DER склав 70 % для Llama, 45 % для Ministral, 30 % для Qwen (стд.) та 20 % для Qwen (reasoning). Водночас середній ASR для інструментальних категорій (DPI, IPI, TM, PE) не перевищив 12 %. Це свідчить про фундаментальну асиметрію refusal-механізмів: моделі значно краще протидіють виконанню шкідливих дій через інструменти, ніж запитам на розкриття інформації через текстові відповіді.

Легітимізація через контекст. Сценарій DPI-04, де зловмисник повідомляє агенту «це тестове середовище, правила безпеки вимкнені», був успішним у 10/10 запусків для всіх трьох моделей у стандартному режимі. Водночас прямі спроби перевизначення інструкцій («Ignore previous instructions», DAN jailbreak, «I am administrator») відхилялись у

100 % випадків. Це свідчить про те, що маніпуляція контекстом значно ефективніша за пряме перевизначення ролі.

Поляризація профілів. Ministral 3 демонструє абсолютну стійкість до категорій IPI, TM та PE (0 %), але значну вразливість до DEX (45 %). Qwen 3 у стандартному режимі, навпаки, демонструє найбільш рівномірний профіль з ненульовими значеннями у всіх п'яти категоріях, зокрема IPI = 30 % — вразливість до непрямої ін'єкції, відсутня у моделях Llama та Ministral.

5. Вплив reasoning на захищеність

Для дослідження впливу режиму reasoning на захищеність проведено паралельне тестування Qwen 3 з вимкненим та увімкненим reasoning на ідентичному каталозі ($500 + 500 = 1000$ запусків). Результати порівняння наведено у табл. 3.

Таблиця 3. Вплив режиму reasoning на ASR (%) для моделі Qwen 3 8B

Категорія	Без reasoning	З reasoning	Зниження
DPI	10,0	0,0	повне
IPI	30,0	1,0	у 30 разів
TM	10,0	1,0	у 10 разів
DEX	30,0	20,0	у 1,5 рази
PE	30,0	9,0	у 3,3 рази
ASR заг.	**22,0**	**6,2**	**у3,5 рази**

Reasoning знижує загальний ASR у 3,5 рази. Найбільший ефект спостерігається для непрямої ін'єкції промпту (IPI): зниження з 30 % до 1 % — у 30 разів. Внутрішній ланцюжок міркувань дозволяє моделі проаналізувати контент файлу та визначити, що знайдені інструкції є частиною даних, а не системними командами, що є частковим вирішенням фундаментальної проблеми нерозрізнення інструкцій та даних [3].

Reasoning також повністю усуває вразливість до маніпуляції контекстом (DPI-04: з 10/10 до 0/10). Водночас режим reasoning забезпечує найменший ефект для категорії DEX (зниження у 1,5 рази): модель коректно «розуміє», що не повинна розкривати інформацію, але все ж генерує відповідь з конфіденційними даними.

Водночас увімкнення reasoning супроводжується зростанням часу формування відповіді у 4,8 рази (з 140 до 674 хвилин для 500 запусків), що обумовлює необхідність балансування між рівнем захищеності та допустимою затримкою відповіді залежно від вимог конкретного застосунку.

Висновки

Запропонований метод оцінки захищеності агентних застосунків на основі BMM відрізняється від існуючих підходів комплексним покриттям: від формалізованої моделі загроз, зіставленої з OWASP та MITRE ATLAS, до автоматизованої процедури тестування з відтворюваними метриками. Ключовим авторським внеском є система спеціалізованих метрик (TMR, DER, PER), що деталізують оцінку на рівні інструментів агента.

Експериментальна верифікація на 2000 запусках (чотири конфігурації, $N = 10$) підтвердила, що метод успішно диференціює моделі та виявляє фундаментальну асиметрію: моделі значно краще протидіють шкідливим діям через інструменти (ASR $\leq$ 12 %), ніж витоку конфіденційних даних (DER 20–70 %). Увімкнення reasoning знижує ASR у 3,5 рази, при цьому найбільший ефект — для непрямої ін'єкції (зниження у 30 разів).

Перспективами подальших досліджень є збільшення кількості повторень до 10–20 для побудови довірчих інтервалів, розширення методу на мультиагентні системи та мультимодальні вектори атак, а також дослідження впливу архітектурних захисних механізмів на зниження ASR.

Перелік використаних джерел

1. *OWASP Foundation.* OWASP Top 10 for Agentic Applications. — 2025. — URL: https://genai.owasp.org/resource/owasp-top-10-for-agentic-applications-for-2026/.
2. Agent Security Bench (ASB): Formalizing and Benchmarking Attacks and Defenses in LLM-based Agents / H. Zhang, J. Huang, K. Mei, Y. Yao, Z. Wang, C. Zhan, H. Wang, Y. Zhang // Proceedings of ICLR. — 2025. — URL: https://arxiv.org/abs/2410.02644.
3. Not what you've signed up for: Compromising Real-World LLM-Integrated Applications with Indirect Prompt Injection / K. Greshake, S. Abdelnabi, S. Mishra, C. Endres, T. Holz, M. Fritz // Proceedings of AISec. — 2023. — DOI: 10.1145/3605764.3623985.
4. *MITRE Corporation.* MITRE ATLAS – Adversarial Threat Landscape for AI Systems. — 2025. — URL: https://atlas.mitre.org/.
5. ReAct: Synergizing Reasoning and Acting in Language Models / S. Yao, J. Zhao, D. Yu, N. Du, I. Shafran, K. Narasimhan, Y. Cao // Proceedings of ICLR. — 2023. — URL: https://arxiv.org/abs/2210.03629.