additional challenge is the uneven access to data across different platforms, particularly to moderation logs, interaction histories, and account metadata.

Further research should focus on building a model of a virtual community in which the proportion of AI bots, the intensity of their activity, the level of toxicity, and the speed of moderation vary. This will allow us to identify empirical threshold conditions under which AI bots do not yet disrupt community functioning, as well as critical states under which significant destructive effects, fragmentation of interactions, and a decline in cyberresilience occur.

1. Molodetska K. Information Influence on the Virtual Community: Implementation Features and Method of Detection in Social Internet Services / K. Molodetska, S. Veretiuk, I. Rahimova, S. Milevskyi, V. Khvostenko // 2023 7th International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT). Ankara, Turkiye, 2023. P. 1–6. DOI: 10.1109/ISMSIT58785.2023.10305001.
2. Doshi J., Novacic I., Fletcher C., Borges M., Zhong E., Marino M. C., Gan J., Mager S., Sprague D., Xia M. Sleeper Social Bots: a new generation of AI disinformation bots are already a political threat. arXiv. 2024. DOI: 10.48550/arXiv.2408.12603.
3. ENISA. ENISA Threat Landscape 2025. Luxembourg: Publications Office of the European Union, 2025.

Аналіз атак витоку системних інструкцій у великих мовних моделях

УДК 004.056.5

Віктор Кольченко[1]

Національний університет «Львівська політехніка»,
[1]viktor.v.kolchenko@lpnu.ua

Стрімкий розвиток великих мовних моделей (LLM) та їх інтеграція в інтелектуальних агентів створили нові вектори загроз, серед яких витік системних інструкцій (Prompt Leakage) стає однією з найбільш критичних проблем безпеки. Ця вразливість полягає в ненавмисному розкритті прихованих налаштувань, які визначають логіку поведінки, обмеження безпеки та операційні параметри моделі [1]. В сучасній екосистемі ШІ системні інструкції перетворюються на найцінніший актив інтелектуальної власності, що створює прямі стимули для їх викрадення.

Мета дослідження полягає у проведенні аналізу атак на витік системних інструкцій, дослідженні внутрішніх механізмів їх реалізації для забезпечення безпеки інтелектуальної власності в екосистемах ШІ-агентів.

Системні інструкції у LLM функціонують як прихований шар управління, що задає роль моделі, її тон та межі дозволеної взаємодії. У сучасних агентних системах вони еволюціонують від простих текстових промптів до модульних пакетів навичок, які поєднують робочі процеси, використання інструментів та специфічні доменні знання. Сутність атак типу Prompt Leakage базується на експлуатації фундаментальної здатності моделей до повторення контексту, що

є необхідним для виконання корисних завдань, таких як узагальнення тексту. Зловмисники намагаються обійти фільтри безпеки та змусити модель «забути» про заборону на розголошення інструкцій, використовуючи неоднозначність між даними користувача та керуючими командами. Особливо небезпечними є багатоходові діалоги, де використовується схильність моделі погоджуватись з користувачем та «flip-flop» ефект, що дозволяє підвищити успішність витоку з 17,7% до понад 86% [2].

Методи витоку включають різноманітні стратегії, починаючи з евристичних атак, таких як відновлення здатності моделі до повторення контексту через вгадування початкових токенів (наприклад, «You are ChatGPT») [3]. Більш складні агентні підходи використовують навчання з підкріпленням та кооперативні команди агентів для автоматизованого пошуку вразливостей цільової моделі [1]. Окрему категорію становлять атаки через сторонні канали, такі як PROMPTPEEK, що експлуатують механізм спільного використання KV-cache у багатокористувацьких середовищах для покрокового відновлення токенів чужих запитів [4]. Також поширюються методи «крадіжки навичок», де зловмисники використовують рольові сценарії (наприклад, роль адміністратора) та ін'єкцію «ланцюжка думок» для ексфільтрації пропрієтарних алгоритмів [5].

Методи захисту еволюціонують у напрямку багатошарової оборони. Програмні методи включають «захист сендвічем», дублювання інструкцій, XML-тегування та переписування запитів для видалення шкідливих компонентів [1]. Проте більш надійними є архітектурні рішення, зокрема SysVec (System Vectors), що пропонує кодувати системні промпти як внутрішні вектори активації, повністю видаляючи їх із текстового контексту, що робить їх недоступними для прямого копіювання [3]. Іншим інноваційним підходом є пробінг інтентів, який дозволяє виявити намір моделі здійснити витік через аналіз прихованих станів останнього токена вхідної послідовності ще до початку генерації відповіді з точністю понад 90% [1]. Для вихідної фільтрації застосовуються системи, які поєднують семантичний аналіз за допомогою потужних моделей-суддів із перевіркою посимвольного збігу [5].

Попри ці заходи, існують суттєві проблеми та обмеження методів захисту. Стохастична природа LLM та відсутність чіткої межі між інструкціями та даними роблять повне запобігання витокам наразі недосяжним. Більшість методів аналізу внутрішніх станів вимагають доступу до моделі за принципом «білої скриньки», що неможливо для розробників, які працюють через сторонні API [3]. Крім того, адаптивні атаки через переклад на інші мови або складне перефразування часто обходять фільтри схожості, а жорсткі обмеження можуть негативно впливати на корисність моделі та швидкість її відповіді [1].

Аналіз витоків системних інструкцій підкреслює необхідність впровадження ешелонованої оборони. Вона повинна поєднувати мінімізацію привілеїв агентів, архітектурне приховування інструкцій, безперервний моніторинг аномалій у внутрішніх станах та ретельну фільтрацію вихідних даних [3, 5]. Тільки комплексна комбінація цих підходів дозволить забезпечити стійкість комерційних ШІ-систем до сучасних автоматизованих атак.

1. Sternak T., Runje D., Granoša D., Wang C. Automating Prompt Leakage Attacks on Large Language Models Using Agentic Approach – 2025. – URL: https://arxiv.org/pdf/2502.12630
2. Agarwal D., Fabbri A., Risher B., Laban P., Joty S., Wu C.-S. Prompt Leakage Effect and Mitigation Strategies for Multi-turn LLM Applications // Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track. – Miami, Florida, USA, 2024. – C. 1255–1275. – URL: https://aclanthology.org/2024.emnlp-industry.94/
3. Cao B., Li C., Cao Y., Ge Y., Wang T., Chen J. You Can't Steal Nothing: Mitigating Prompt Leakages in LLMs via System Vectors // Proceedings of the 2025 ACM SIGSAC Conference on Computer and Communications Security (CCS '25). – Taipei, Taiwan, 2025. – URL: https://arxiv.org/abs/2509.21884
4. Wu G., Zhang Z., Zhang Y., Wang W., Niu J., Wu Y., Zhang Y. I Know What You Asked: Prompt Leakage via KV-Cache Sharing in Multi-Tenant LLM Serving // Network and Distributed System Security Symposium (NDSS) 2025. – San Diego, CA, USA, 2025. – URL: https://www.ndss-symposium.org/ndss-paper/i-know-what-you-asked-prompt-leakage-via-kv-cache-sharing-in-multi-tenant-llm-serving/
5. Wang Z., Zhang R., Liu Y., Liu C., Zhao Q., Li H., Xu G. Black-Box Skill Stealing Attack from Proprietary LLM Agents: An Empirical Study // arXiv. – 2026. – URL: https://arxiv.org/abs/2604.21829

Еволюція стратегії ЄС щодо протидії іноземному втручанню та маніпулюванню інформацією (FIMI)

УДК 327:004.056

Сергій Кондратюк

Державний університет інформаційно-комунікаційних технологій, s.kondratiuk@duikt.edu.ua

У період 2015-2026 рр. стратегія Європейського Союзу у сфері протидії іноземним маніпуляціям і втручанням (FIMI) пройшла глибоку еволюцію — від поодиноких реакцій на дезінформаційні кампанії до створення узгодженої, багаторівневої екосистеми проактивної протидії. У березні 2015 р. на саміті ЄС у Брюсселі було ухвалене рішення про необхідність протидії постійним дезінформаційним кампаніям Росії, що зумовило створення East StratCom Task Force у межах Європейської служби зовнішніх справ (EEAS).

До 2022 р. на рівні ЄС було напрацьовано базову систему захисту від FIMI. Вона включала Платформу EUvsDisinfo; План дій проти дезінформації 2018 р., який формалізував FIMI як гостру внутрішню загрозу демократії ЄС; та Кодекс практики щодо дезінформації. Важливим етапом стало створення у середині 2016 р. Центру аналізу гібридних загроз ЄС (EU Hybrid Fusion Cell) у складі EEAS для аналізу розвідувальної інформації та OSINT. У квітні 2017 р. в Гельсінкі засновано Європейський центр передового досвіду з протидії гібридним загрозам (Hybrid CoE) — міжнародний хаб для країн ЄС та НАТО.