ДЕРЖАВНИЙ УНІВЕРСИТЕТ ІНФОРМАЦІЙНО-КОМУНІКАЦІЙНИХ ТЕХНОЛОГІЙ

НАВЧАЛЬНО-НАУКОВИЙ ІНСТИТУТ КІБЕРБЕЗПЕКИ ТА ЗАХИСТУ ІНФОРМАЦІЇ
КАФЕДРА СИСТЕМ ТА ТЕХНОЛОГІЙ КІБЕРБЕЗПЕКИ

ВСЕУКРАЇНСЬКА НАУКОВО-ПРАКТИЧНА КОНФЕРЕНЦІЯ



«АКТУАЛЬНІ ПРОБЛЕМИ КІБЕРБЕЗПЕКИ»

Тези доповідей

29 жовтня
2026

м. Київ

Ланде Дмитро Володимирович, доктор технічних наук, професор, завідувач кафедри інформаційної безпеки
Кузьмінський Віктор Сергійович, аспірант кафедри інформаційної безпеки
Національний технічний університет України «Київський політехнічний інститут імені Ігоря Сікорського», Київ

ВИЛУЧЕННЯ ТРИПЛЕТІВ ПРО КІБЕРЗАГРОЗИ З ТЕКСТОВИХ ПОВІДОМЛЕНЬ ІЗ ГАРАНТОВАНОЮ МЕЖЕЮ ЧАСТКИ ПОМИЛОК

Анотація. Запропоновано метод вилучення триплетів «суб'єкт — зв'язок — об'єкт» про кіберзагрози з текстових повідомлень, який приймає лише кандидатів з оцінкою, не нижчою за поріг, перевірений на розмічених текстах джерела. Гарантовано, що середня по текстах частка хибних серед прийнятих з тексту триплетів не перевищує заданого рівня α з імовірністю щонайменше $1 - \delta$. Під нове джерело поріг перекалібровують без перенавчання великих моделей. На англомовних звітах CTINexus і українських публікаціях CERT-UA рівень перевищувався не більш як у 1,3 % повторів, а влучність за $\alpha = 0{,}3$ становила 80–81 % проти 45–74 % у мовної моделі без фільтра.

Графи знань про кіберзагрози дедалі частіше наповнюють автоматично: велика мовна модель (LLM) читає звіт і записує з нього триплети «суб'єкт — зв'язок — об'єкт», наприклад «UAC-0050 — використовує — RemcosRAT». Скільки серед них хибних, наперед невідомо, і ця частка залежить від джерела: у моделі Ministral 3 8B вона становила 26 % на англомовних звітах CTINexus [1] і 49 % на українських публікаціях CERT-UA. Чимало помилок стосуються не зв'язку, а самих сутностей: замість назви діяча чи шкідливої програми модель ставить загальне слово або уламок назви. Власна впевненість LLM у задачах про кіберзагрози погано узгоджується з часткою її правильних відповідей [2], а поріг, підібраний на розмічених текстах без статистичної перевірки, у наших дослідах перевищував заданий рівень помилок у 38–54 % випадкових поділів текстів.

Мета роботи — метод, який приймає з тексту лише ті триплети, для яких середня частка хибних не перевищує заданого аналітиком рівня α, і зберігає цю властивість під час переходу до нового джерела повідомлень.

Кандидати і оцінки. Кандидати бувають двох видів. Перші — пари сутностей: компактна модель GLiNER2 знаходить згадки сутностей, LLM відповідає, чи варта кожна з них вузла графа, а для пари — чи є між сутностями зв'язок, якого він типу і яка з них суб'єкт. Другі — триплети, які LLM виписує з тексту сама. Дві логістичні регресії зводять ознаки кандидатів до оцінок від 0 до 1, і про всіх кандидатів вирішує один спільний поріг (рис. 1).

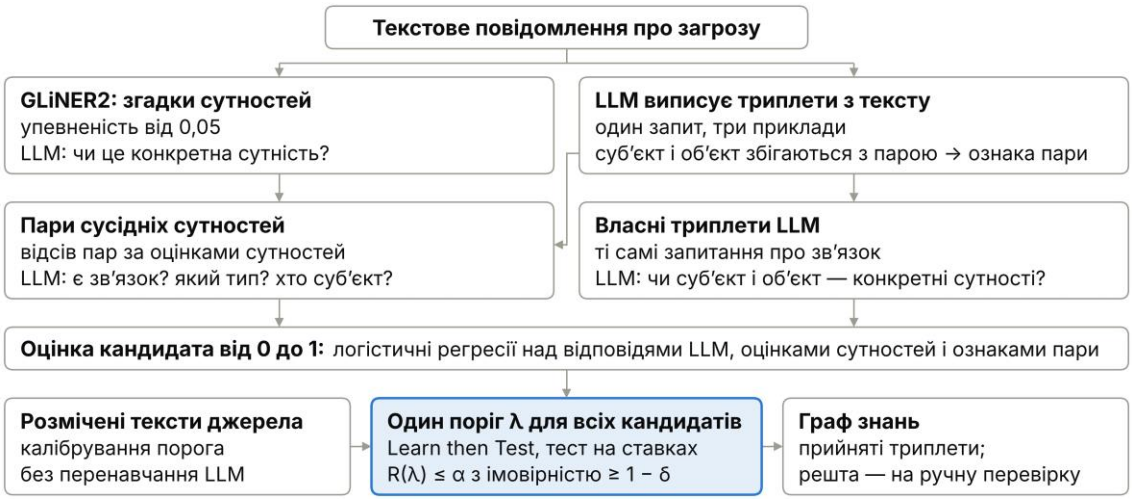


Рис. 1. Схема методу: один перевірений поріг для двох видів кандидатів

Строге означення правильного триплета. Триплет правильний, якщо текст прямо стверджує зв'язок між суб'єктом і об'єктом і обидва вони є конкретними сутностями, тобто називають річ, яку аналітик лишив би вузлом графа. Поріг перевіряють саме за такими мітками, тому гарантія охоплює і зв'язок, і обидві сутності, а оцінки сутностей лише допомагають прийняти більше правильних триплетів. Для порівняння, на CTINexus два послідовні фільтри (спершу сутності, потім зв'язки) з тим самим рівнем для кожного перевищували рівень для триплетів у 58 % повторів, а оцінки сутностей у складі оцінки пари збільшували відносну повноту (частку знайдених зі спільного переліку правильних зв'язків) в 1,3–4 рази.

Вибір і перевірка порога. Одиницею гарантії є текст, бо помилки в межах одного звіту залежні: у них спільні сутності, контекст і запит до моделі. Ризик порога — сподівана частка хибних серед прийнятих з тексту триплетів, усереднена по текстах, з яких щось прийнято. Розмічені калібрувальні тексти ділять на дві частини: чверть пропонує кілька порогів, решта перевіряє їх тестом на основі ставок [3] за схемою Learn then Test [4]. Якщо тексти незалежні й однаково розподілені, ймовірність прийняти поріг із ризиком, не меншим за α, не перевищує заданого δ; якщо перевірки не пройшов жоден поріг, метод нічого не приймає. Пройти перевірку можна лише за достатньої кількості текстів, з яких щось прийнято: за $\delta = 0{,}1$ це щонайменше 88, 43, 20 і 13 текстів для $\alpha = 0{,}05$; 0,1; 0,2; 0,3. Ця межа задає найменший обсяг розмітки під нове джерело.

Адаптація до джерела. Великі моделі під нове джерело не перенавчають: досить заново відкалібрувати поріг на розмічених текстах цього джерела, а щоб приймати більше, перенавчити й логістичні регресії. Без перекалібрування гарантія не переноситься: у постановці з еталонними сутностями пороги, перевірені на CTINexus, на текстах CERT-UA перевищували рівень у 55–71 % повторів, а після перекалібрування на текстах CERT-UA — не більш як в 1 % повторів.

Результати. Метод перевірено на англомовному наборі CTINexus і на корпусі зі 158 публікацій CERT-UA за 2018–2026 роки, зібраному в цій роботі, з двома LLM на 8 млрд параметрів (Ministral 3 8B і Llama 3.1 8B), що працюють локально; на CERT-UA — за протоколом, записаним наперед. Середня по тестових текстах частка хибних перевищувала α не більш як у 1,3 % із 300 випадкових поділів (з Llama на CERT-UA — у 4,3 %). За α = 0,3 влучність методу (частка правильних серед прийнятих) становила 80–81 % на обох наборах проти 73–74 % на CTINexus і 51 % (Ministral) чи 45 % (Llama) на CERT-UA у LLM без фільтра. Перевага не пояснюється лише більшими обчисленнями: самоперевірка і самоузгодженість LLM за зіставних витрат часу мали на CERT-UA з Ministral на 19–23 пункти нижчу влучність. Рівень α задає компроміс між влучністю і відносною повнотою: за α = 0,3 на CERT-UA з Ministral метод знаходить удвічі менше зв’язків, ніж LLM без фільтра, а за α = 0,5 — і більше зв’язків, і з вищою влучністю. Мітки для калібрування поставила мовна модель Claude за письмовим правилом; за мітками незалежного фахівця з кібербезпеки на вибірці текстів CERT-UA середня частка хибних за α = 0,3 становить 0,218 (верхня одностороння 95-відсоткова межа 0,265), тобто нижча за α.

Висновки. Метод дає аналітикові один керований параметр — допустиму частку помилок — і триплети для графа знань із передбачуваною середньою часткою хибних. Під нове джерело потрібні лише розмічені тексти цього джерела, а конвеєр вилучення працює локально, тож тексти не залишають організацію. Код, мітки й протоколи відкрито (https://github.com/datch1588/cti-gate-experiment). Подальші дослідження: більші LLM і послідовна перевірка, що вказує, коли потрібне перекалібрування.

Список використаних джерел

1. Cheng Y., Bajaber O., Tsegai S. A., Song D., Gao P. CTINexus: Automatic cyber threat intelligence knowledge graph construction using large language models. *Proceedings of the 10th IEEE European Symposium on Security and Privacy (EuroS&P)*. 2025. URL: https://arxiv.org/abs/2410.21060.

2. Mezzi E., Massacci F., Tuma K. Large language models are unreliable for cyber threat intelligence. *Availability, Reliability and Security (ARES 2025)*. LNCS, vol. 15993. Springer, 2025. P. 343–364. DOI: 10.1007/978-3-032-00627-1_17.

3. Waudby-Smith I., Ramdas A. Estimating means of bounded random variables by betting. *Journal of the Royal Statistical Society Series B*. 2024. Vol. 86, no. 1. P. 1–27. DOI: 10.1093/jrsssb/qkad009.

4. Angelopoulos A. N., Bates S., Candès E. J., Jordan M. I., Lei L. Learn then test: Calibrating predictive algorithms to achieve risk control. *The Annals of Applied Statistics*. 2025. Vol. 19, no. 2. P. 1641–1662. DOI: 10.1214/24-AOAS1998.