АРХІТЕКТУРНА ОПТИМІЗАЦІЯ ЛОКАЛЬНИХ RAG-СИСТЕМ ДЛЯ АВТОМАТИЗАЦІЇ КОМПЛАЄНСУ NIST SP 800-53 НА БАЗІ APPLE SILICON

Надточий М. М., Балагура Д. С.
Харківський національний університет радіоелектроніки, Харків, Україна

Сучасні процеси SecOps (Security Operations) вимагають швидкого доступу до нормативної документації, зокрема до стандарту NIST SP 800-53, що містить сотні контролів безпеки. Використання публічних великих мовних моделей (LLM), таких як ChatGPT, для аналізу внутрішньої інфраструктури на відповідність стандартам є неприпустимим через критичні ризики витоку чутливих даних. Ефективним рішенням є розгортання локальних AI-асистентів на базі Retrieval-Augmented Generation (RAG), проте спроби використання стандартних "коробкових" рішень часто призводять до галюцинацій або відмов моделі генерувати відповідь

Метою доповіді є визначення оптимальної архітектури локальної RAG-системи, здатної коректно інтерпретувати ієрархічні стандарти безпеки на обмежених обчислювальних ресурсах на базі Apple Silicon, та вирішення проблем стандартного парсингу документів.

Дослідження проводилося на апаратній платформі Apple MacBook Pro (процесор M1 Pro, Unified Memory), яка забезпечує високу пропускну здатність пам'яті для швидкого інференсу квантованих моделей. Програмний стек включав Ollama, LangChain та ChromaDB; тестувалися моделі Llama 3.1 (8B), Qwen 2.5 (7B) та Gemma 2 (9B). На першому етапі було реалізовано стандартний підхід "Naive RAG" (парсинг офіційного PDF-файлу з розбиттям на частини), який показав незадовільні результати. Було зафіксовано, що моделі невеликого розміру, зокрема Llama 3.2 3B, у 100% випадків відмовлялися давати відповідь ("I don't know") на специфічні запити щодо багаторівневих контролів, таких як SC-7 Boundary Protection.

Аналіз виявив проблему "Фрагментації контексту": механічний поділ тексту розриває семантичний зв'язок між ідентифікатором контролю безпеки та його описом. Крім того, було виявлено феномен "витіснення батьківського контексту" (Vector Dilution), коли векторна база повертала підконтролі, залишаючи загальний опис батьківського контролю нижче порогу відсікання Top-K.

Для розв'язання цих проблем було розроблено покращений конвеєр обробки даних (ETL Pipeline). По-перше, здійснено перехід від PDF до структурованих даних у форматі CSV із застосуванням методу "Context Injection", де кожен векторний документ примусово включав метадані у тілі тексту, що зробило фрагменти самодостатніми. По-друге, впроваджено алгоритм гібридного пошуку Hybrid Retrieval, який ефективно поєднує Regex-пошук ідентифікаторів, прямий запит у базу метаданих (Hard Filtering) та

традиційний семантичний векторний пошук. Таке дворівневе архітектурне рішення не лише повністю усунуло феномен "витіснення батьківського контексту", але й по-різному вплинуло на продуктивність інференсу квантованих моделей у форматі GGUF на базі Ollama.

Детальніший аналіз впровадженого конвеєра обробки даних ETL Pipeline та результатів тестування розкриває специфіку роботи алгоритму гібридного пошуку та реакцію різних архітектур мовних моделей на структурований контекст. Індустріальний стандарт Llama 3.1 (8B Instruct) продемонструвала "парадокс швидкості" зі зменшенням загального часу генерації на 16,4% та максимальною експертною оцінкою якості RAG у 5.0 балів, що пояснюється тим, що за наявності чіткого контексту модель припиняє генерувати надлишковий текст та галюцинації, видаючи точну й лаконічну відповідь. Для інших перевірених архітектур було зафіксовано так званий "Time Penalty" — затримку через обробку розширеного промпту. Для перспективної моделі Qwen 2.5 (7B Instruct) час генерації збільшився на 5,8%, хоча якість згенерованої відповіді також залишилася на еталонному рівні у 5.0 балів без жодних галюцинацій. Водночас модель Gemma 2 (9B) виявилася найменш оптимізованою для даного RAG-пайплайну: її час відповіді суттєво зріс на 30,7%, а експертна оцінка якості знизилася до 4.0 балів. Ці порівняльні метрики переконливо доводять, що саме моделі розміром 7-8 мільярдів параметрів є оптимальним вибором для локальних SecOps-завдань, оскільки вони забезпечують ідеальний баланс між швидкістю інференсу та абсолютною точністю цитування стандарту. Усі протестовані моделі класу 7B-9B стабільно працювали на 16GB Unified Memory, що підтверджує високу апаратну ефективність Apple Silicon для локальних задач SecOps.

Практична значущість отриманих результатів полягає у доведенні того факту, що побудова ефективного локального AI-асистента для завдань кібербезпеки неможлива шляхом простого поєднання неструктурованих документів ("PDF + LLM").

Критичним фактором успіху виступає саме якість підготовки даних (Data Engineering) та обов'язкове використання гібридного пошуку (Lexical/Metadata + Semantic).

Впровадження запропонованої архітектури локального RAG не лише гарантує приватність чутливих даних, але й оптимізує час отримання критично важливої інформації за рахунок генерації моделлю відповідей без надлишкового тексту.

Список літератури

1. Lewis, Patrick, et al. "Retrieval-augmented generation for knowledge-intensive nlp tasks." Advances in neural information processing systems 33 (2020): 9459-9474. DOI: https://doi.org/10.48550/arXiv.2005.11401

2. Security and Privacy Controls for Information Systems and Organizations. NIST Special Publication 800-53, Revision 5. National Institute of Standards and Technology. 2020. DOI: https://doi.org/10.6028/NIST.SP.800-53r5