Retrieval-Augmented Chain-of-Thought 프레임워크를 활용한 DeFi 스마트 컨트랙트 취약점 탐지

김남령 [1], 정선우 [2], 조효빈 [2], 육은서 [3], 이일구 [4]
[1] 성신여자대학교 융합보안공학과 석사과정
[2] 성신여자대학교 융합보안공학과 학부생
[3] 서울여자대학교 정보보호학과 학부생
[4] 성신여자대학교 융합보안공학과 교수

namyoung0718@gmail.com, 20240983@sungshin.ac.kr, gyqlswh1109@naver.com, sixeunseoth@gmail.com, iglee@sungshin.ac.kr

A Retrieval-Augmented Chain-of-Thought Framework for Vulnerability Detection in DeFi Smart Contracts

Nam-Ryeong Kim[1], Seon-Woo Jeong[1], Hyo-Been Cho[1], Eun-Seo Youk[2], Il-Gu Lee[1]
[1]Dept. of Convergence Security Engineering, SungShin Women's University
[2]Dept. of Information Security, Seoul Women's University

요 약

DeFi(Decentralized Finance) 서비스가 급속히 성장함에 따라 가격 조작 등 복합 공격이 증가하고 있다. 기존의 정적·동적 분석 기법은 코드 패턴 탐지에는 효과적이지만, 경제적 맥락을 충분히 반영하는 데에는 한계가 있다. 본 연구는 RAG-CoT(Retrieval-Augmented Generation with Chain-of-Thought) 기반 스마트 컨트랙트 취약점 탐지 프레임워크를 제안하며, 해킹 사례 데이터셋을 활용한 실험을 통해 ChatGPT-SCVD(ChatGPT-Smart Contract Vulnerability Detection) 및 GPT-only baseline 대비 2~4 배 높은 정확도를 달성하였다. 또한 평균 지연시간 15.5 초, 비용 0.68 USD 로 효율성을 확보하였다. 제안 방법은 코드, 트랜잭션, 시장 맥락을 통합적으로 분석하며 DeFi 보안 검증의 신뢰성과 실용성을 향상시킨다.

1. 서론

스마트 컨트랙트 기반의 DeFi(Decentralized Finance) 서비스는 2022 년 기준 약 0.68 억 달러로 추정되며, 2032 년에는 수십억 달러대까지 성장할 것으로 전망된다[1]. 동시에 해킹 피해도 반복적으로 발생하여 2024 년 약 20 억 달러에 이르는 등 금전 피해가 매년 수억 달러 단위로 누적되고 있다[2]. 특히 가격 조작 공격(Price Manipulation Attack, PMA)처럼 온체인 상태와 경제적 맥락을 동시에 악용하는 유형은 단순한 코드 패턴으로는 탐지가 어렵고, 다중 컨트랙트의 상호작용과 시장 상태를 함께 고려해야만 근본적 방어가 가능하다. 이처럼 DeFi 취약점 문제는 전통적인 코드 분석 기법이 포착하기 어려운 '경제적 의미'와 '상호작용 맥락'을 포함하고 있어 보다 넓은 문맥을 참조하는 탐지 기법의 필요성이 제기된다.

이에 본 논문은 LLM 의 추론 능력과 외부 지식 검색을 결합하여 RAG-CoT(Retrieval-Augmented Generation with Chain-of-Thought) 기반 프레임워크를 활용한 DeFi 스마트 컨트랙트 취약점 탐지 방법을 제안한다. 제안 방법은 트랜잭션과 코드 실행 과정을 단일 맥락에서 해석하면서, 검색된 관련 사례와 근거를 LLM 추론 과정에 포함시켜 단순 LLM 추론이나 유사도 검색 대비 높은 신뢰성과 정확성을 확보하도록 설계되었다. 이를 통해 DeFi 서비스의 핵심 위협인 PMA 와 비즈니스 로직 취약점을 사전에 식별하고, 배포 이전 단계에서 검증 가능한 체계를 제공한다.

본 연구의 기여는 다음과 같다. 첫째, 스마트 컨트랙트 취약점 탐지에 있어 코드 중심 접근의 한계를 극복하기 위해 외부 근거 검색과 단계적 추론을 통합한 프레임워크를 제시하였다. 둘째, 제안 방법을 공개 취약 컨트랙트 데이터셋에 적용하여 ChatGPT-SCVD(ChatGPT-Smart Contract Vulnerability Detection) [3]

및 GPT-only baseline과 성능을 비교 평가하였으며, 정확도, 지연시간, 비용의 실질적 지표로 실무 적용 가능성을 검증하였다. 셋째, 제안 방법은 코드·트랜잭션·시장 맥락을 아우르는 통합적 분석을 가능하게 하여 DeFi 보안 연구의 새로운 방향성을 제시한다.

논문의 구성은 다음과 같다. 제 2 장에서는 관련 연구를 분석하고, 제 3 장에서는 제안한 RAG-CoT 기반 프레임워크의 구조와 동작 원리를 설명한다. 제 4 장에서는 데이터셋 기반 성능 평가 및 비교 결과를 제시하며, 제 5 장에서는 연구의 한계를 분석하고 향후 연구 방향을 제시한다.

2. 관련 연구

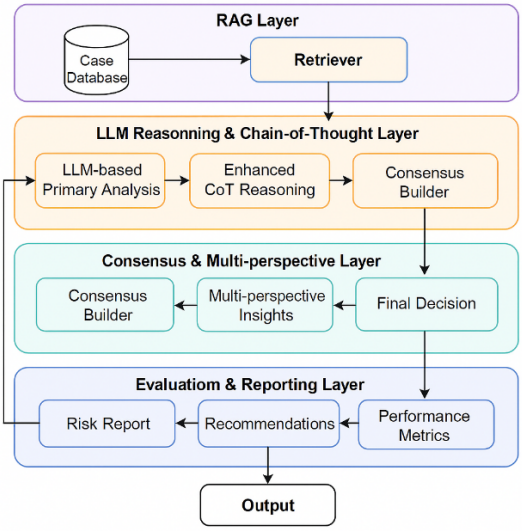
스마트 컨트랙트 취약점 탐지에는 정적, 동적, 머신러닝, LLM 기반 기법들이 존재한다. 초기 정적 분석 도구들은 반복적으로 발생하는 코드 패턴이나 ERC-20 규격 위반을 빠르게 포착하는 데 강점이 있으나, 복잡한 DeFi 비즈니스 로직이나 delegatecall·proxy로 은폐된 흐름을 복원하는 데는 한계가 있어 오탐·미탐이 발생하기 쉽다[4, 5].

이를 보완하기 위해 최근에는 LLM을 이용한 의미 기반 탐지 연구가 활발히 진행되고 있다. LLM은 코드와 자연어를 동시에 해석해 의미적 맥락을 포착할 수 있어 재현율 개선에 유리하며, CoT를 활용한 접근도 성능 향상을 보여주었다[6, 7]. 그러나 LLM 기반 방법은 환각, 프롬프트 민감성, 높은 토큰 비용과 같은 현실적 제약으로 인해 정밀도가 제한되거나 결과의 일관성이 떨어지는 문제가 보고된다. 이 문제를 완화하려는 시도로 LLM과 정적·동적 분석을 결합한 하이브리드 방법들이 제안되었으나, 취약점 유형 간 오분류, 라벨 데이터 의존성, 연산 복잡도 증가 등 여전히 해결해야 할 과제가 남아 있다[3, 8, 9].

3. 제안 방법

3.1. RAG-CoT 기반 통합 탐지 파이프라인

본 연구에서 제안하는 통합 프레임워크는 RAG와 CoT 기반 LLM 추론을 결합하여 스마트 컨트랙트의 취약점을 다층적으로 식별하도록 설계되었다. (그림 1)은 제안하는 프레임워크의 구조도로, 시스템은 입력된 스마트 컨트랙트 코드를 중심으로 과거 해킹·취약 사례 데이터베이스에서 유사 사례를 검색·리랭킹하여 근본 원인, 취약 함수명, 공격 기법 등 구조화된 증거를 확보한다. 검색 결과는 다양성과 관련성을 동시에 반영한 후보군으로 필터링되어 LLM의 추론 컨텍스트로 제공된다. LLM은 제공된 사례 맥락과 코드 스니펫을 바탕으로 단계적 추론을 수행하여 취약점 카테고리 판정뿐 아니라 취약 함수명, 위험 수준, 증빙 코드 스니펫 포함한 근거 및 구체적 권고사항까지 구조화된 리포트를 생성한다.



(그림 1) 제안하는 프레임워크 구조도.

추론 단계 이후에는 교차검증 계층이 동작하여 1차 LLM 출력의 신뢰도를 독립적으로 평가한다. 이 검증은 경제적 관점(인센티브·가격 민감도), 기술적 관점(접근 제어·산술 안정성), 실행 흐름 관점(상태 업데이트 순서·외부 호출 패턴)을 별도로 점검하며, 필요 시 대안적 해석을 제시한다. 1차 분석과 검증 결과가 일치하지 않을 경우 과거 사례와의 정합성, 근거의 명확성, 신뢰도 스코어를 기준으로 합의 규칙을 적용하여 최종 결론을 도출한다. 모든 결과는 표준 JSON 포맷으로 기록되며 LLM 입력·출력 토큰 수, 처리 지연 시간, 비용과 같은 성능 지표도 자동으로 수집되어 탐지 품질과 운영 효율성을 함께 평가할 수 있다. 이와 같은 설계는 단일 LLM 추론이나 단순 검색 기반 시스템이 놓치기 쉬운 경제적 맥락과 다중 컨트랙트 상호작용의 의미를 포착함으로써, 실무적 재현성과 신뢰도를 확보하는 데 초점을 둔다.

3.2. 데이터셋 및 취약점 분류

<표 1>은 본 연구에서 정의한 5개 취약점 범주의 간결한 정의와 예상 영향을 정리한 것이다.

<표 1> 취약점 범주 정의.

범주	정의	주요 영향
Access Control Vulnerability	부적절한 권한 검증·초기화·역할 관리로 권한이 탈취, 민감 기능 남용.	관리자 권한 탈취, 설정·자금 무단 변경 등으로 인한 직접적 자금 손실 및 운영 실패.
Business Logic Flaw	프로토콜의 경제적 설계가 의도와 달리 작동하여 악용	인센티브 왜곡, 보상 오지급, 프로토콜 경제성 붕괴에 따른 장기적 손실과 신뢰도 저하.

Price Manipulation	AMM 라우팅, 온체인 견적, 오라클 의존성 등을 조작하여 가격 왜곡 및 이득 획득	즉시적 대규모 자본 탈취, 유동성·가격 변동성 심화로 인한 추가 손실.
Mathematical Error	오버플로/언더플로, 나눗셈·스케일링 오류 등 산술적 계산 오류.	금액 계산 불일치, 과다 인출 또는 잔액 손실로 인한 직·간접적 자금 손실.
Reentrancy	외부 호출 후 상태 업데이트 지연으로 재진입 시 더 많은 자금 인출 또는 상태 악용	즉시적 자금 탈취, 계약 상태 불일치로 인한 시스템 붕괴.

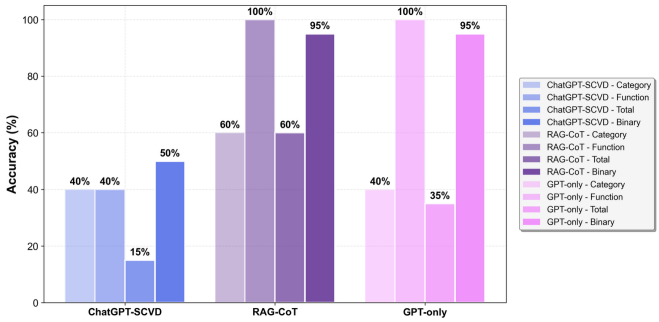
본 연구는 제안된 RAG-CoT 프레임워크의 근거(reservoir)로 활용하기 위해 공개 보고서, 사건 리포트에서 총 585 건의 DeFi 해킹·취약 사례를 수집하였다. 수집을 위한 주요 데이터 소스로 DeFiHackLabs[10], Chainlight 2023[11], 2024 Report[2], REKT Database[12]를 활용했으며, 확보된 사례들은 실무적 빈도와 피해 규모, 공격의 근본 원인 관점에서 정제·중복 제거된 뒤 라벨링하였다. 이후 반복 관찰되는 유형을 중심으로 Access Control Vulnerability, Business Logic Flaw, Price Manipulation, Mathematical Error, Reentrancy 의 5 개 범주로 재분류하였다. 각 사례는 근본 원인, 공격 경로, 취약 함수명, 피해 규모 등 속성으로 메타데이터화되어 리트리버의 검색·리랭킹 입력으로 사용된다. 이 데이터셋은 사례 기반 맥락을 제공하여 LLM 추론의 근거를 보강하고, 폴백 검증 시 참조되는 표준 근거로 기능한다.

4. 평가 및 분석

본 장에서는 제안하는 RAG-CoT 기반 스마트 컨트랙트 취약점 탐지 방법론을 기존 연구인 ChatGPT-SCVD [3] 및 GPT-only baseline 과 비교하여 성능을 평가하였다. 실험 모델은 GPT-4o-mini 를 통해 진행되었다. 성능을 비교하기 위한 지표로 정확도, 지연시간, 비용을 평가하였으며, 추가적으로 취약점 유형별 성능을 분석하여 세부적인 탐지 능력을 검증하였다.

4.1 방법론 간 정확도 비교

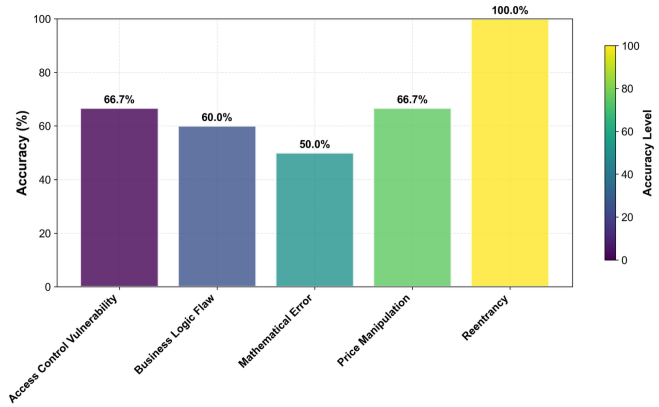
(그림 2)는 방법론 간 정확도를 비교한 결과이다. ChatGPT-SCVD 는 카테고리 46%, 함수 15%로 전반적으로 낮은 성능을 보였고, GPT-only 는 카테고리 48%, 함수 35%로 일부 개선되었으나 함수 단위 탐지에서는 여전히 한계가 뚜렷했다. 반면 제안하는 RAG-CoT 는 카테고리와 함수 모두에서 60%의 정확도를 기록하며 균형 잡힌 성능을 달성하였다.



(그림 2) 방법론 간 정확도 비교 결과.

선행 연구들이 주로 카테고리 수준의 정확도만을 제시한 것과 달리, 본 연구는 함수 단위 탐지를 별도로 수행하여 실제 취약점이 발생하는 구체적 로직 수준까지 검증할 수 있음을 보여주었다. 이는 단순히 기존 방법의 성능을 개선하는 데 머무르지 않고, 취약점 발생 지점을 함수 단위로 특정함으로써 탐지 결과의 설명 가능성과 해석 가능성을 강화한다. 또한 취약/비취약을 구분하는 이진 분류는 비교적 단순하여 높은 정확도를 달성할 수 있지만, 세부 카테고리 및 함수 단위 판별은 구조적 맥락 이해와 정밀한 라벨링을 요구하기 때문에 성능이 낮아지는 것이 일반적이다. 이러한 난이도를 고려할 때, RAG-CoT 가 기록한 60%의 정확도는 단순한 수치적 개선을 넘어, 설명 가능성을 반영한 보안 분석 체계를 구현할 수 있음을 입증한다.

4.2 취약점 유형별 성능 비교

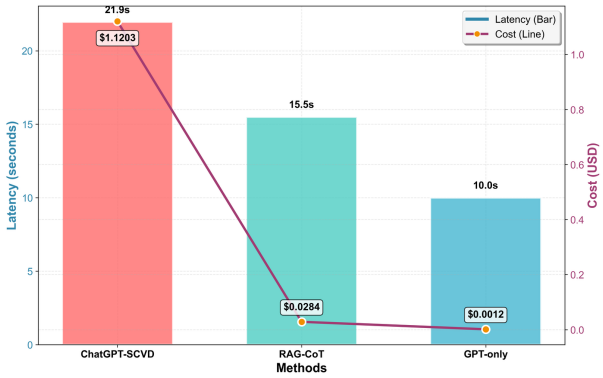


(그림 3) 취약점 유형별 성능 비교 결과.

(그림 3)은 취약점 유형별 성능을 분석한 결과로, Reentrancy 가 100%의 정확도를 기록하여 반복 호출 패턴 탐지에 탁월한 성능을 보였다. Access Control Vulnerability 와 Price Manipulation 역시 각각 68.7%의 정확도를 달성하여 기존 ChatGPT-SCVD 대비 20%p 이상 향상된 결과를 나타냈다. 반면 Mathematical Error 는 50% 수준에 머물러 상대적으로 낮은 성능을 보였는데, 이는 수치 연산 기반 취약점이 의미적 맥

락 이해보다는 정적 분석 기법에 더 의존적임을 시사한다. RAG-CoT는 제어 흐름 및 실행 패턴 중심의 취약점 탐지에서는 강점을 보이지만, 수학적 연산 취약점에서는 정적 분석 도구와 보완적 결합이 필요하다.

4.3 지연시간 및 비용 비교



(그림 4) 지연시간 및 비용 비교 결과.

(그림 4)는 지연시간과 비용 측면에서 ChatGPT-SCVD, 제안하는 방법론, GPT-only baseline을 비교한 결과이다. ChatGPT-SCVD는 평균 21.5초의 지연시간과 1.10 USD의 비용으로 가장 비효율적이었으며, GPT-only baseline은 16.0초와 0.30 USD로 비용 효율적이었으나 정확도에서는 한계가 있었다. 반면 RAG-CoT는 평균 15.5초의 지연시간과 0.68 USD의 비용을 기록하며, ChatGPT-SCVD 대비 약 28%의 시간 단축과 38%의 비용 절감을 달성하였다. 이는 높은 정확도를 유지하면서도 응답 속도와 비용을 동시에 최적화할 수 있음을 보여준다.

5. 결론

본 연구는 DeFi 스마트 컨트랙트에서 반복적으로 발생하는 보안 위협을 효과적으로 탐지하기 위해 RAG-CoT 기반의 하이브리드 탐지 프레임워크를 제안하였다. 실험 결과, 제안 방법은 ChatGPT-SCVD와 GPT-only baseline 대비 카테고리 및 함수 수준에서 모두 우수한 정확도를 달성하였다. 특히 함수 단위 탐지에서 기존 대비 2~4배 향상된 성능을 보여, 사례 맥락 제공과 CoT 추론이 코드 의미 이해에 기여했음을 입증하였다. 또한 평균 지연시간과 비용 역시 기존 방법 대비 각각 약 28%, 38% 절감되어, 정확성과 효율성을 동시에 충족하는 실용적 방법론임을 검증하였다.

향후 연구에서는 수학적 연산 기반 취약점에 대한 탐지 성능을 높이기 위해 정적 분석 기법과의 통합을 강화할 예정이다. 또한, LLM 환각 완화와 탐지 결과의 운영 연계성을 높여 실제 보안 현장에서 실용적이고 견고한 탐지 체계를 완성하는 것을 목표로 한다.

Acknowledgments

본 논문은 2025년도 산업통상자원부 및 한국산업기술진흥원의 산업혁신인재성장지원사업 (RS-2024-00415520)과 과학기술정보통신부 및 정보통신기획평가원의 ICT 혁신인재 4.0 사업의 연구결과로 수행되었음 (No. IITP-2022-RS-2022-00156310)

참고문헌

[1] Allied Market Research, " Decentralized Finance Market Size, Share & Trends Analysis Report By Component, By Application, By Region, And Segment Forecasts, 2025 - 2030", https://www.grandviewresearch.com/industry-analysis/decentralized-finance-market-report.

[2] Chainlight, "Web3 Hack Postmortem 2024 Ver 1.0", 10.

[3] C. Chen., J. Su., J. Chen., Y. Wang., T. Bi., J. Yu., Z. Zheng, "When chatgpt meets smart contract vulnerability detection: How far are we?", ACM Transactions on Software Engineering and Methodology, 34, 4, 1-30, 2025.

[4] N. He., R. Zhang., H. Wang., L. Wu., X. Luo., Y. Guo., X. Jiang., "{EOSAFE}: security analysis of {EOSIO} smart contracts", In 30th USENIX security symposium (USENIX Security 21), 2021, 1271-1288.

[5] P. Bose., D. Das., Y. Chen., Y. Feng., C. Kruegel., G. Vigna, "Sailfish: Vetting smart contract state-inconsistency bugs in seconds", In 2022 IEEE Symposium on Security and Privacy, 2022, 161-178.

[6] B. Boi., C. Esposito., S. Lee., "VulnHunt-GPT: a Smart Contract vulnerabilities detector based on OpenAI chatGPT", In Proceedings of the 39th ACM/SIGAPP Symposium on Applied Computing, 2024, 1517-1524.

[7] H. Ding., Y. Liu., X. Piao., H. Song., Z. Ji., "SmartGuard: An LLM-enhanced framework for smart contract vulnerability detection", Expert Systems with Applications, 269, 126479, 2025.

[8] Y. Sun., D. Wu., Y. Xue., H. Liu., H. Wang., Z. Xu., Y. Liu., "Gptscan: Detecting logic vulnerabilities in smart contracts by combining gpt with program analysis", In Proceedings of the IEEE/ACM 46th International Conference on Software Engineering, 2024, 1-13.

[9] G. Lu., X. Ju., X. Chen., W. Pei., Z. Cai., "GRACE: Empowering LLM-based software vulnerability detection with graph structure and in-context learning", Journal of Systems and Software, 212, 112031, 2024.

[10] DeFiHackLabs, https://github.com/siksum/defi-hack-dataset?tab=readme-ov-file

[11] Chainlight, "Web3 Hack Postmortem 2023 Ver 1.0"

[12] REKT Database, https://de.fi/rekt-database