심볼릭 실행 기반 LLM 에이전트를 통한 허니팟 스마트 컨트랙트 탐지 방법론 제안

이동하*, 김수민*, 한보윤*, 서정택**

*가천대학교 스마트보안학과 (학생)

**가천대학교 스마트보안학과 (교수)

LLM Agent-Driven Honeypot Smart Contract Detection Using Symbolic Execution

DongHa Lee*, SuMin Kim*, BoYun Han*, Jung Taek Seo**

*Department of Smart Security, Gachon University (Undergraduate student)

**Department of Smart Security, Gachon University (Professor)

요 약

허니팟 스마트 컨트랙트는 공격자가 의도적으로 취약하게 작성한 컨트랙트로, 숨겨진 조건이나 로직에 의해 자금 인출이 불가능하도록 설계된 새로운 사기 형태이다. 이를 자동으로 탐지하기 위해 심볼릭 실행과 휴리스틱을 결합한 연구들이 제안되었으나, 경로 폭발 문제와 규칙 기반 접근의 한계로 인해 복잡한 구조의 허니팟을 정확히 식별하는 데 근본적인 제약이 있다. 본 연구에서는 이를 극복하기 위해 CFG 추출과 심볼릭 실행으로 수집한 경로 조건을 바탕으로 ReAct 기반 LLM 에이전트가 SMT 솔버를 활용하여 허니팟 여부를 판정하는 탐지 파이프라인을 제안하며, 기존 연구 대비 정확도를 67.2%p 향상시키며 그 효과를 입증하였다.

I. 서론

허니팟 스마트 컨트랙트는 공격자가 의도적으로 취약해 보이도록 설계되어 사용자의 자금 투입을 유도하지만, 숨겨진 조건이나 로직으로 인해 실제 인출은 불가능한 사기형 컨트랙트이다. 이를 자동으로 탐지하기 위해 HoneyBadger[1]와 같은 심볼릭 실행·휴리스틱 기반 기법이 제안되었으나, 경로 폭발 문제와 규칙 기반 접근의 한계로 인해 복잡하거나 변형된 허니팟을 식별하는 데 제약이 있다.

심볼릭 실행은 실행 경로와 상태 정보를 체계적으로 생성하는 데 강점이 있지만, 이를 의미론적으로 해석하는 능력은 제한적이다. 반면 LLM은 코드와 실행 정보에 대한 맥락적 이해에 강점을 가지므로, 두 기법의 결합은 기존 한계를 보완할 수 있다.

본 연구에서는 CFG 추출과 심볼릭 실행을 통해 EVM 바이트코드에서 경로 조건과 실행 정황을 추출하고, 이를 LLM이 해석 가능한 고수준 경로 표현으로 구조화하는 허니팟 탐지 파이프라인을 제안한다. ReAct 기반 에이전트는 SMT 솔버를 활용해 경로 조건의 만족 가능성을 검증하고, 실제 도달 가능한 의심 경로에 대해서만 근거 기반 판단을 수행한다. 그 결과, Original 데이터셋에서는 재현율 84.4%, F1 91.5%, Stealth 데이터셋에서는 재현율 90.6%, F1 95.1%를 달성하여 기존 방법 대비 탐지 성능이 크게 향상되었다.

II. 배경 및 관련 연구

HoneyBadger[1]는 EVM 바이트코드로부터 CFG를 구성하고 심볼릭 실행을 수행하여, 자금의 유입, 유출 여부를 경로 조건 기반으로 분석하는 허니팟 탐지 연구이다. 심볼릭 실행 결과를 사전에 정의된 8가지 휴리스틱 허니팟 패턴과 비교하여 허니팟 여부를 판정한다. 해당 연구 이후 허니팟 탐지 문제를 다양한 관점에서 확장하는 연구들이 제안되었다. 적대적 허니팟 연구[2]는 기존 탐지 도구가 의존하는 코드 및 행동 기반 특징이 공격자에 의해 우회될 수 있음을 보이며 규칙 기반 탐지의 강건성 한계를 지적했다. Greed Trap[3]은 자금 귀속 관계를 기준으로 허니팟을 분석하는 새로운 접근을 제시했다. SCH-Hunter[4]는 테인트 분석, 퍼징, 심볼릭 실행을 결합하여 탐지 범위를 확장하였으나, 복잡한 허니팟 구조에 대한 의미론적 해석의 한계는 여전히 남아 있다.

III. LLM 기반 허니팟 스마트 컨트랙트 탐지 방법론 제안

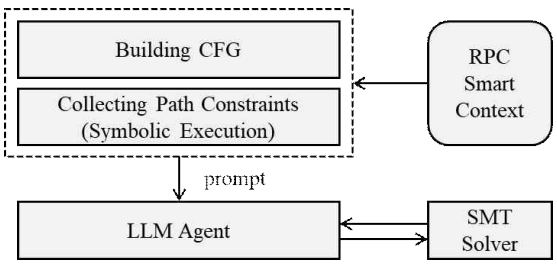

Fig 1. Overview of the proposed method

Fig 1은 허니팟 탐지를 위해 LLM 기반 분석 파이프라인을 보여준다. 본 방법은 LLM이 탐지 규칙을 생성하는 방식이 아니라, CFG 추출·심볼릭 실행·SMT 솔버 검증을 통해 얻은 구조화된 실행 정보를 규칙 기반 분석 대신 LLM이 해석하도록 함으로써, 실행 경로 기반 의미 해석과 탐지 정확도를 함께 향상시킨다.

3.1 경로 제약조건 수집

경로 제약조건 수집 모듈은 RPC(Remote Procedure Call)을 통해 획득한 스마트 컨트랙트의 바이트코드와 상태 정보를 입력으로 받아 CFG를 구성하고, 심볼릭 실행을 수행하여 각 실행 경로에 대한 경로 제약조건과 실행 문맥을 추출한다. 심볼릭 실행은 실제 값 대신 기호를 입력으로 사용하여 가능한 실행 경로를 탐색하며, 자금의 유입이나 유출이 발생하는 경로를 중심으로 분기 조건, 상태 제약, 외부 호출 조건과 같은 핵심 정보를 수집한다. 추출된 결과는 바이트코드 형태가 아닌 LLM이 해석 가능한 고수준 경로 표현으로 구조화되어 LLM 에이전트에 프롬프트로 전달된다.

3.2 LLM 에이전트

LLM 에이전트는 구조화된 경로 조건과 실행 문맥을 입력으로 받아 ReAct 구조를 기반으로 허니팟 여부를 반복적으로 추론한다. 구체적으로, 에이전트는 경로 조건과 실행 문맥을 요약하고 잠재적 허니팟 시나리오를 도출한 후, 핵심 경로 조건에 대해 SMT 솔버를 호출하여 만족 가능성을 검증한다. 검증 결과를 바탕으로 실행 가능성이 없는 경로를 제거하고, 실제로 도달 가능한 의심 경로에 대해서만 추론 결과와 판단 근거를 갱신하는 과정을 반복함으로써 정확한 허니팟 판정을 수행한다.

IV. 실험 및 검증

4.1 실험 환경

실험에 사용된 데이터셋은 Original과 Stealth 두 가지로 구성된다. Original 데이터셋은 HoneyBadger[1]에서 평가에 사용한 32개의 허니팟 스마트 컨트랙트로 구성되며, Stealth 데이터셋은 규칙 기반 탐지를 우회하도록 설계된 연구[2]를 기반으로 제작한 컨트랙트 32개로 구성된다. 비교 대상은 HoneyBadger의 규칙 기반 탐지 방식과 제안하는 LLM 기반 탐지 방식이며, 탐지 성능 평가를 위해 정밀도(Precision), 재현율(Recall), 정확도(Accuracy), 그리고 F1 스코어를 성능 지표로 사용했다. 실험에 사용된 LLM 모델은 gpt-5.3-codex이며, 구현에는 z3-solver, pyevmasm 등을 활용했다.

4.2 탐지 정확도 비교

실험	N	방법	Precision	Recall	F1	Acc
Original	32	기존	100%	40.6%	57.8%	40.6%
		제안	**100%**	**84.4%**	**91.5%**	**84.4%**
Stealth	32	기존	-	0.0%	0.0%	0.0%
		제안	**100%**	**90.6%**	**95.1%**	**90.6%**

Table 1. Benchmark Results

Table 1은 HoneyBadger(기존)와 제안 방법론의 탐지 성능을 비교한 결과이다. Original 데이터셋(32개)에서 기존 방법은 재현율 40.6%, F1 57.8%에 그쳤으나, 제안 방법론은 재현율 84.4%, F1 91.5%로 크게 향상된 성능을 보였다. 두 방법 모두 탐지한 샘플에 대해서는 오탐이 없어 정밀도는 동일하게 100%를 기록하였다. Stealth 데이터셋(32개)에서는 기존 방법이 단 한 건도 탐지하지 못한 반면(재현율 0%), 제안 방법론은 3건의 미탐지에 그쳐 재현율 90.6%, F1 95.1%를 달성하였다. 이는 규칙 기반 탐지가 우회형 허니팟 패턴에 전혀 대응하지 못하는 반면, LLM 기반 의미론적 해석이 변형된 패턴에도 강건하게 동작함을 명확히 보여준다.

4.3 시사점

심볼릭 실행은 컨트랙트의 복잡도가 증가할수록 탐색 경로 수가 크게 늘어나 분석 시간이 급격히 증가하는 한계를 가진다. 실제 실험에서도 기존 방식은 Original 데이터셋 32개 중 10개(31.3%)에서 실행 경로를 산출하지 못했다. 그럼에도 LLM 에이전트는 부분적으로 수집된 경로 조건과 실행 문맥을 바탕으로 허니팟 여부를 판별할 수 있었으며, 이는 심볼릭 실행의 한계를 보완할 수 있음을 시사한다.

V. 결론 및 향후 연구

본 연구에서는 규칙 기반 허니팟 탐지 방식의 한계를 극복하기 위해, CFG 추출과 심볼릭 실행을 통해 경로 제약조건을 수집하고 LLM 에이전트가 이를 해석하여 허니팟 여부를 판정하는 탐지 파이프라인을 제안하였다. 심볼릭 실행은 경로 탐색에 집중하고, 의미론적 해석은 ReAct 기반 에이전트가 SMT 솔버를 활용하여 수행하는 역할 분리 구조를 통해 기존 방식의 한계를 보완하였다. 벤치마크 실험에서 기존 허니팟 패턴에 대해서는 동등한 성능을 유지하면서도, 규칙 기반 탐지를 우회하도록 설계된 Stealth 데이터셋에서 탐지율 0%에 그친 기존 방법과 달리 재현율 90.6%, F1 95.1%를 달성하여 그 효과를 실증하였다. 향후 연구로 경로 폭발 문제를 완화하기 위한 탐색 전략의 최적화, 더 다양한 우회 패턴을 포함하는 데이터셋의 확장, 그리고 온체인 실시간 탐지 환경으로의 적용 가능성 검토를 진행할 예정이다.

[Acknowledgment]

이 논문은 2026년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원(RS-2025-02218141, 선박 사이버 침해사고 분석 기술 및 탐지·대응 기술개발)을 받아 수행된 연구임.

[참고문헌]

[1] The Art of the Scam: Demystifying Honeypots in Ethereum Smart Contracts, 28th USENIX Security Symposium (USENIX Security), 2019.

[2] Yu Han, Tiantian Ji, Zhongru Wang, Hao Liu, Hai Jiang, Wendi Wang, and Xiang Cui, An Adversarial Smart Contract Honeypot in Ethereum, Computer Modeling in Engineering & Sciences, 2021.

[3] Mahtab Norouzi, Mounir Elgharabawy, and Kaiwen Zhang, The Greed Trap: Uncovering Intrinsic Ethereum Honeypots through Symbolic Execution, 2023 Fifth International Conference on Blockchain Computing and Applications (BCCA), 2023.

[4] Haoyu Zhang, Baotong Wang, Wenhao Fu, and Leyi Shi, SCH-Hunter: A Taint-Based Hybrid Fuzzing Framework for Smart Contract Honeypots, Information, 2025.