양자내성암호 전환을 위한 멀티-에이전트 LLM 프레임워크에 관한 연구

하준수*, 정진호*, 서정민*, 신찬희*, 문재현*, 윤기순*

*한국항공대(대학생), *중부대(대학생), *중앙대(대학생), *이화여자대(대학생),
*한국거래기술사회(교신저자), *NSHC(교신저자)

Research on a Multi-Agent LLM Framework for the Transition to Post-Quantum

Junsu Ha*, Jihno Jung*, Jungmin Seo*, Chanhee Shin*,
Moonjae Hyun*, Kisoon Yoon*

*Korea Aerospace Univ.(Undergraduate), *Joongbu Univ.(Undergraduate),
*Chungang Univ.(Undergraduate), *Ewha Womans Univ.(Undergraduate),
*Korea Technology Transfer Agent Association(Corresponding author),
*NSHC(Corresponding author)

요 약

양자 컴퓨팅 기술의 발전은 기존 공개키 암호체계에 대한 실존적 위협으로 부상하고 있다. 이에 따라 양자내성암호(PQC)로의 전환이 시급한 과제로 인식되고 있으나, 방대한 시스템 전반에 분포된 양자 취약 암호 자산을 식별하는 것은 어려운 초기 단계이다. 본 연구는 이 문제에 대응하기 위해 거대 언어 모델(LLM)을 활용한 계층적 멀티-에이전트 시스템(MAS)을 제안한다. 종합된 결과를 바탕으로 Non-PQC 알고리즘을 식별하고 PQC 전환 가이드를 생성한다. 본 연구는 멀티-에이전트 아키텍처가 PQC 전환의 초기 단계를 자동화하고 가속화하는 실용적이고 확장 가능한 프레임워크임을 입증한다.

I. 서론

양자 컴퓨터(CRQC)의 등장은 RSA, ECC와 같은 현대 암호체계의 보안 시스템을 근본적으로 위협한다.[1] 그러나 양자내성암호(PQC)로의 전환 과정에서 가장 큰 난관 중 하나는 방대한 소프트웨어 자산 속에 내재된 양자취약 암호화(Non-PQC)구현을 정확히 식별하는 것이다. 이러한 이유로 미국 국립표준기술연구소(NIST)와 미국 사이버보안 및 인프라 안보국(CISA)은 성공적인 마이그레이션의 필수적인 첫 단계로 암호 자산의 발견 및 목록화를 강조하고 있다.[2]

본 연구는 이러한 문제점에 대해 거대 언어 모델(LLM) 기반의 멀티-에이전트 시스템을 제안한다. 이 시스템은 초기 분석에서 식별된 Non-PQC 후보군에 대해 심층적인 검증과 분석을 수행하며, 최종적으로 실행 가능한 마이그레이션 지침을 제공한다. 단일 LLM이 갖는 구조적 한계를 개선하고자, 각기 다른 전문성을 지닌 다수의 에이전트가 협력하는 AI 오케스트레이션 아키텍처를 활용한다.

II. 관련 연구

2.1 자동화된 코드 분석

자동화된 코드 분석 분야는 LLM의 등장으로 중요한 전환점을 맞이했다. 그러나 보안 취약점 탐지와 같은 전문 영역에 단일 LLM을 직접 적용하는 것은 제한된 컨텍스트 창, 추론의 비일관성, 그리고 특정 보안 지식의 부족과 같은 한계가 존재한다.[3]

2.2 제안 시스템 구조

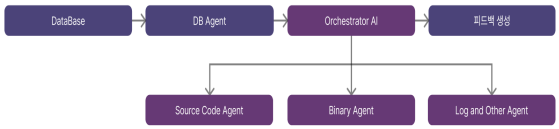
이러한 한계를 극복하기 위해, 본 연구는 다

수의 전문 에이전트가 협력하는 멀티-에이전트 시스템(MAS) 아키텍처를 채택한다.[4] 이는 단일 범용 모델에 의존하는 대신, 각기 다른 분석 관점(소스코드, 바이너리, 설정파일 등)을 가진 전문가 집단의 협력을 모방하여 보다 신뢰도 높은 결론을 도출하는 것을 목표로 한다.

III. Muitl-Agent LLM Framework 설계

3.1 LLM 오케스트레이션 구조

[Fig 1] Framework Flowchart



본 프레임워크는 [Fig 1]과 같이 초기 분석에서 Non-PQC 사용 가능성이 있는 것으로 분류된 데이터를 DB를 통해 입력받아 심층 분석을 수행한다. 시스템은 계층형 MAS 구조를 따르며, 중앙 오케스트레이션 에이전트와 네 개의 전문 에이전트로 구성된다.

3.2 멀티 에이전트 프레임워크

오케스트레이션 에이전트는 시스템의 중앙 조정자로서, 분석 대상을 기반으로 작업을 세분화하여 각 전문 에이전트에게 할당한다. 이후 각 에이전트로부터 수집된 분석 결과를 종합하여 최종 결론을 도출하고, DB 에이전트와 연동하여 마이그레이션 가이드를 생성한다.

소스코드 에이전트는 고수준 소스 코드 내의 암호화 API 호출, 알고리즘 식별자, 키 활용 방식 등 코드의 의미론적 맥락을 분석하는 역할을 담당한다. 어셈블리 에이전트는 디스어셈블된 바이너리 코드에서 암호화 라이브러리 호출이나 알고리즘 고유의 상수 값과 같은 저수준의 증거를 식별한다. 로그 및 설정 에이전트는 빌드 스크립트나 의존성 명세서 등에서 암호화 라이브러리 의존성과 같은 선언적 증거를 탐색한다. DB 에이전트는 Non-PQC 알고리즘과 NIST가 승인한 PQC 대체 알고리즘 간의 매핑 정보와 같은 동적 지식 베이스 역할을 수행하며, 오케스트레이션 에이전트에 의해 지속적으로 최신 정보로 갱신될 수 있다.

IV. 실험적 평가

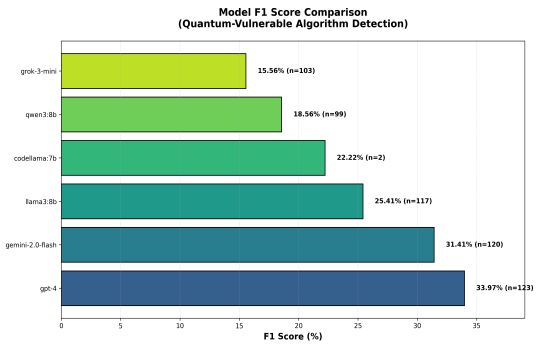
4.1 실험 설계

본 연구의 벤치마크는 두 가지 주요 배포 시나리오를 고려하여 모델을 선정했다. 첫째, 일반 사용자를 대상으로는 상용 클라우드 API를 활용하여 접근성과 신속한 서비스 제공을 목표로 한다. 둘째, 민감 정보를 다루는 정부나 기관을 위해서는 정보 유출 위험을 최소화할 수 있도록 폐쇄망 환경에서 독립적으로 구동 가능한 오픈소스 LLM을 제공하는 것을 목표로 한다. 이러한 이원적 접근 방식을 통해 gpt-4, gemini-2.0-flash, grok-3-mini와 같은 상용 모델과 llama3:8b, qwen3:8b, codellama:7b등 오픈소스 모델을 포함한 총 6개 모델을 평가 대상으로 선정했다.

테스트 케이스는 소스코드, 어셈블리 바이너리, 로그 및 설정파일의 세 가지 에이전트 유형에 맞춰 약 1,000번 이상의 테스트를 수행했다.

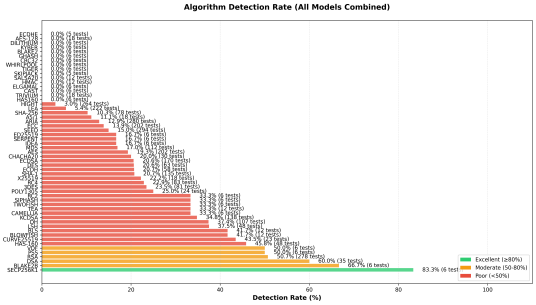
4.2. 결과 및 논의

[Fig 2] Model F1 Score



[Fig 2]는 선정한 모델의 정확성인 F1 Score을 보여준다. F1-Score는 Precision(정밀도)와 Recall(재현율)의 조화평균을 의미한다. 실험 결과, gpt-4와 같은 상용 API 모델은 높은 F1 Score를 기록하며 정확도가 높은 분석 결과를 보였으나, 오픈 소스모델들은 비교적 낮은 정확도를 보였다. 이는 모델 사이즈의 크기나 학습 데이터 수준이 알고리즘 분석 성능에 영향이 있음을 나타낸다.

[Fig 3] Algorithm Detection Rate



[Fig 3]은 전체 모델의 알고리즘별 탐지율을 종합한 결과이다. RSA, DES 등과 같은 국제 표준 알고리즘에서는 50% 이상의 비교적 높은 탐지율을 보였다. 이는 해당 알고리즘들이 LLM의 주요 학습 데이터셋의 양이 방대하기 때문으로 분석된다.

반면, SEED, ARIA와 같은 국내 표준 암호 알고리즘에 대해서는 약 10% 대의 전반적으로 낮은 탐지율을 기록했다. 이는 대부분의 LLM이 영미권 중심의 데이터로 학습되어, 나라별 특화된 알고리즘에 대한 지식이 부족함을 나타내는 명백한 한계점이다.

V. 향후 연구

본 연구 결과는 몇 가지 향후 연구 방향을 제안한다. 첫째, 국산 암호나 최신 PQC 알고리즘에 대한 탐지율을 높이기 위해 RAG(Retrieval-Augmented Generation) 기술을 도입 하거나, 관련 데이터셋으로 전문 에이전트를 추가로 조정하는 방안을 고려할 수 있다.[5]

둘째, 비용 및 속도 효율성을 극대화하기 위해 지식 증류(Knowledge Distillation) 기법의 도입을 제안한다. 고성능 모델을 이용하여 Non-PQC 탐지 지식을 더 작은 모델로 이전시키는 것이다.[6] 이를 통해 운영 비용과 응답 시간을 단축한 소형 특화 모델인 sLLM(Smaller Large Language Model)을 개발할 수 있을 것이다.

VI. 결론

본 논문은 PQC 전환의 핵심 선결 과제인 양자 취약 암호 자산 식별을 위해 LLM 기반의 멀티-에이전트 프레임워크를 제안한다.

중앙 오케스트레이션 에이전트를 중심으로 소스코드, 어셈블리, 설정파일에 특화된 전문 에이전트끼리 협력하는 제안한 아키텍처는 단일 LLM의 한계를 보완하고 다각적인 분석을 가능하게 한다.

포괄적인 벤치마크를 통해 상용 API 모델과 온프레미스 배포가 가능한 오픈소스 모델 모두에서 유의미한 성능을 확인했으며, 제안된 아키텍처의 실용적 가치 및 향후 연구 방향을 제안했다. 추후 제안된 모델의 개선 작업을 진행할 경우 PQC 전환의 첫 단계를 자동화 하는 실질적 시스템을 구성 할 수 있을 것이다.

본 논문은 2025년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임 (RS-2024-00399401, 양자안전 보안인프라 전환 및 대양자 복합 안전성 검증기술 개발 결과물입니다.

참고문헌

[1] Alagic, G., et al. (2020). Status Report on the Second Round of the NIST Post-Quantum Cryptography Standardization Process. NIST Interagency or Internal Report 8309.

[2] CISA. (2025). Strategy for Migrating to Automated Post-Quantum Cryptography Discovery and Inventory Tools.

[3] Santos, J. C. S., & Sabetta, A. (2025). On the Use of Large Language Models for Automated Program Repair. IEEE Security & Privacy.

[4] Zhang, D., et al. (2021). Physical Safety and Cyber Security Analysis of Multi-Agent Systems: A Survey of Recent Advances. IEEE/CAA Journal of Automatica Sinica.

[5] Lewis., et al.(2020). Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. Advances in Neural Information Processing Systems, 33, 9459-9474.

[6] Hinton., et al.(2015). Distilling the Knowledge in a Neural Network. arXiv preprint arXiv:1503.02531.