Smart Media Journal / Vol.15, No.6 / ISSN:2287-1322
https://dx.doi.org/10.30693/SMJ.2026.15.6.100

블록체인 합의 파라미터 제어를 위한 SMT 안전 실드 통합 Safe RL 프레임워크의 실증적 검증

(Empirical Validation of a Safe Reinforcement Learning Framework with SMT-Based Safety Shield for Blockchain Consensus Parameter Control)

이지운

(Ji Woon Lee)

요 약

블록체인 합의 프로토콜의 파라미터를 동적으로 조정하는 강화학습(RL) 기반 접근법은 성능 향상의 가능성을 보여 왔으나, 학습 과정에서의 안전성 보장은 미해결 과제로 남아 있다. 본 연구는 SMT(Satisfiability Modulo Theories) 솔버 기반 안전 실드(Safety Shield)를 Deep Q-Network(DQN) 학습 루프에 통합한 Safe RL 프레임워크를 제안하고, QuadTree 기반 하이브리드 합의(QBHC) 환경에서 그 유효성을 실증한다. 실험 결과, 안전 실드는 합의 안전 속성 위반을 완전히 제거하면서 처리량을 대폭 향상시켰다. Z3 SMT 기반 정확 검증과 규칙 기반 다항식 근사 검증은 대상 속성에 대해 동일한 판정 결과를 보이면서도 검증 지연시간에서 유의미한 차이를 나타냈다. DQN 에이전트는 학습 과정에서 안전 제약을 내재화한 정책을 습득하였으며, 부하 변동 시나리오에서 정적 정책 대비 우수한 적응적 제어를 수행하였다. 이상의 결과는 형식 검증 기반 안전 실드가 안전 보장과 성능 향상을 동시에 달성하며, 에이전트의 정책 형성을 안전한 방향으로 유도하는 학습 가드레일로 기능할 수 있음을 보여준다.

■ 중심어 : 블록체인 합의 ; 심층 강화학습 ; 안전 강화학습 ; 형식 검증 ; 안전 실드

Abstract

Reinforcement learning (RL)-based approaches for dynamically tuning blockchain consensus protocol parameters have shown promising performance gains, but ensuring safety during learning remains a challenge. This study proposes a Safe RL framework that integrates an SMT (Satisfiability Modulo Theories)-based Safety Shield into the DQN (Deep Q-Network) training loop and validates it in a QuadTree-based Hybrid Consensus (QBHC) environment. Results show that the safety shield eliminates consensus safety violations while improving transaction throughput. The Z3 SMT-based verifier and a rule-based polynomial approximation verifier yield identical results but differ significantly in latency. The DQN agent learns safety-aware policies and outperforms static policies under dynamic loads, suggesting that formal verification-based safety shields can ensure both safety and performance while guiding inherently safe policy learning.

■ keywords : Blockchain Consensus ; Deep Reinforcement Learning ; Safe Reinforcement Learning ; Formal Verification ; Safety Shield

Ⅰ. 서 론

블록체인 네트워크의 확장성 한계를 극복하기 위해 샤딩(Sharding) 기술이 널리 연구되고 있다. 선행 연구[1]에서는 로컬(L1) 합의에 Raft를 사용하여 각 샤드 내에서 트랜잭션을 빠르게 합의한 뒤, 합의된 proposal을 글로벌(L2) 합의인 PBFT로 최종 완결하는 이중 계층 하이브리드 합의 구조를 설계하고, Z3/PySMT 기반 형식 검증을 통해 Safety 및 Liveness 속성이 Byzantine Fault 환경에서도 보장됨을 확인하였다. 그러나

정회원, 한국기술교육대학교 미래융합학부 기술연구원

접수일자 : 2026년 04월 08일
게재확정일 : 2026년 05월 14일
교신저자 : 이지운 e-mail : jwlee7@koreatech.ac.kr

해당 연구에서는 합의 프로토콜의 파라미터(타임아웃, 배치 크기, 위원회 크기, 샤드 깊이 등)가 정적으로 고정되어 있어, 부하 급증이나 네트워크 지연 변동 등 동적 환경 변화에 대한 적응이 불가능하였다. 선행 연구의 결론에서도 향후 과제로 강화학습 기반 동적 파라미터 제어와 실제 네트워크 환경 검증을 명시적으로 제안한 바 있다. 이 선행 연구를 직접 확장하여, 형식 검증된 하이브리드 합의 구조 위에 심층강화학습(DQN, Deep Q-Network) 기반 동적 파라미터 제어를 통합한 Safe RL 프레임워크를 제안한다. 선행 연구에서 합의 안전성 검증에 사용된 SMT 기반 형식 검증을 안전 실드(Safety Shield)로 재구성하여, RL 에이전트의 학습 루프에 직접 통합하는 구조를 채택하였다. 이 구조에서 학습 과정의 탐색(exploration)은 합의 안전 속성을 위반하지 않으면서도 동적 환경에 적응하는 정책 학습이 가능하다.

기존의 RL 기반 블록체인/네트워크 자원 관리 연구[2-4]는 처리량이나 지연시간 등 성능 지표의 최적화에 집중하면서, 합의 프로토콜의 형식적 안전 속성을 보장하는 문제를 명시적으로 다루지 않았다. 안전 강화학습(Safe RL) 분야에서는 Constrained Policy Optimization(CPO)[5]이나 Lagrangian 기반 제약 충족 방법[6]이 제안되었으나, 이들은 주로 연속 제어 환경에 초점을 맞추며 블록체인 합의의 이산적 안전 속성 검증에는 적합하지 않다. SMT 솔버를 활용한 형식 검증과 RL의 통합은 로봇 제어 분야에서 이론적으로 제안된 바 있으나[7], 블록체인 합의 환경에서의 실증은 전무하다. 선행 연구[1]는 하이브리드 합의 구조의 형식적 안전성을 검증하였으나, 동적 파라미터 제어와 RL 통합은 미래 과제로 남겨두었다.

이번 연구에서는 SMT 기반 형식 검증 안전 실드를 DQN 학습 루프에 직접 통합한 Safe RL 프레임워크를 제안하고, 다음 세 가지 연구 질문에 답한다. (1) “안전 실드가 합의 프로토콜의 안전 속성을 보장하면서도 시스템 성능을 유지 또는 향상시키는가?” 안전 실드의 행동 차단이 처리량 저하를 초래하는지, 아니면 위반 복구 비용을 절감하여 성능을 향상시키는지를 검증한다. (2) “Z3 SMT 기반 정확 검증과 규칙 기반 근사 검증이 동치인 조건은 무엇이며, 근사 검증의 실시간 적용이 가능한가?” 두 실드 구현의 판정 일치 조건을 실증적으로 확인하고, 각 구현의 검증 지연시간을 비교하여 런타임 배포 아키텍처를 제시한다. (3) “DQN 에이전트가 안전 실드와의 상호작용을 통해 안전 제약을 내재화한 정책을 학습하는가?” 학습 완료 후 에이전트가 안전 실드의 개입 없이도 안전한 행동을 자율적으로 선택하는지를 분석한다.

주요 기여는 다음과 같다. (1) SMT 솔버 기반 형식 검증 안전 실드와 DQN을 통합한 Safe RL 프레임워크를 블록체인 합의 환경에 최초로 설계하고 구현하였다. (2) Z3 정확 검증과 규칙 기반 근사 검증이 안전 속성(P1-P5)에 대해 동치임을 5개 시드에 걸쳐 실증적으로 확인하였다. (3) DQN이 안전 실드와의 학습 상호작용을 통해 안전 제약을 내재화하는 현상을 발견하고 정량화하였다. (4) 6단계 부하 변동 시나리오에서 4개의 실험을 통해 프레임워크의 유효성을 검증하였다.

II. 관 련 연 구

1. 블록체인 샤딩 및 합의 프로토콜

블록체인 샤딩은 처리량 확장성을 위한 핵심 기술로 활발히 연구되고 있다. Ethereum 2.0은 초기 설계에서 Beacon Chain 기반 위원회 샤딩을 채택하였고 각 위원회는 샤드 블록의 가용성과 유효성을 검증하는 역할을 수행한다[9], OmniLedger[10]와 RapidChain[11]은 각각 에포크 기반 랜덤 재배정과 Cuckoo 규칙 기반 재구성을 통해 보안성을 강화하였다. 그러나 이들 시스템은 주로 정적 배정 또는 고정 주기 재배정

방식을 채택하며, 실시간 부하 변화에 대한 동적 파라미터 적응을 명시적으로 지원하지 않는다. Monoxide[12]는 비동기 합의 구역(Asynchronous Consensus Zone)을 제안하여 샤드 간 독립적 합의를 가능하게 하여 확장성을 개선하였으나, 합의 파라미터의 런타임 조정은 다루지 않았다. 최근에는 심층강화학습(DRL) 기반 샤드 배치 최적화가 활발히 연구되고 있는데, SPRING[17]은 DRL 기반 상태 배치(state placement)를 통해 크로스 샤드 트랜잭션을 최소화하는 프레임워크를 제안하였다. 그러나 SPRING을 포함한 기존 DRL 샤딩 연구들은 합의 프로토콜의 형식적 안전 속성 보장을 다루지 않아, 학습 과정에서 안전 위반이 발생할 가능성이 열려 있다.

합의 프로토콜의 형식 검증 측면에서는, SMT 솔버를 활용하여 Hyperledger Fabric의 합의 규칙을 검증한 사례[18]가 있으며, Tendermint BFT의 비결정적 행동을 모델 체킹으로 분석한 연구[19]도 보고되었다. 하이브리드 합의 구조에서는 로컬 합의와 글로벌 합의를 분리하여 효율성과 안전성을 함께 추구하는 접근이 있다. 선행 연구[1]는 L1(로컬) 합의에 Raft를, L2(글로벌) 합의에 PBFT를 적용하여, 각 샤드에서 Raft로 빠르게 합의한 proposal을 PBFT로 최종 완결하는 이중 계층 RAFT-PBFT 하이브리드 합의 구조를 설계하고, Z3/PySMT 기반 bounded model checking을 통해 Safety 5개 속성과 Liveness를 형식적으로 검증하였다. 해당 연구는 합의 프로토콜의 정합성을 입증하였으나, 합의 파라미터가 정적으로 고정되어 있어 동적 환경 적응이 불가능하다는 한계가 있다.

QBHC(QuadTree-Based Hybrid Consensus) 아키텍처는 이 선행 연구의 하이브리드 합의 구조를 기반으로 하면서, 합의 파라미터의 동적 조정을 RL 에이전트에게 위임하고 형식 검증을 안전 실드로 재구성하여 학습 루프에 통합하는 구조를 제안한다.

2. 강화학습 기반 블록체인 및 네트워크 자원 관리

DQN[13]은 이산 행동 공간에서의 가치 추정과 경험 재현을 통한 샘플 효율성으로 네트워크 자원 관리에 광범위하게 적용되어 왔다. SDN 환경에서의 라우팅 최적화[2], 에지 컴퓨팅의 태스크 오프로딩[3], 데이터센터 부하 균형[4] 등에서 정적 기준선 대비 유의한 성능 개선이 보고되었다.

블록체인 합의에 RL을 직접 적용하는 연구도 최근 활발히 진행되고 있다. Ameri 등 [15]은 학습 오토마타(learning automata)를 BFT 프로토콜에 통합하여 블록 크기와 전파 지연을 동적으로 조정하는 인지적 블록체인 아키텍처를 제안하였다. 적대적 환경에서의 블록체인 합의 최적화 연구[15]는 RL을 활용하여 블록체인 파라미터를 동적으로 조정하고 Byzantine 공격 환경에서의 견고성을 검증하였다. Liu 등[16]은 인과 모델 기반 DRL(C-DRL) 알고리즘으로 컨소시엄 블록체인의 파라미터 구성을 다목적 최적화 문제로 정의하고, DQN 대비 높은 설명 가능성을 달성하였다. 컨소시엄 블록체인 파라미터 조정 연구[16]는 인과 모델 기반 DRL(C-DRL) 알고리즘으로 파라미터 구성을 다목적 최적화 문제로 정의하고, DQN 대비 높은 설명 가능성을 달성하였다. 그러나 이들 연구는 합의 프로토콜의 형식적 안전 속성을 학습 과정에서 보장하는 메커니즘을 포함하지 않았다. 부하 증가, 유지, 감소의 전체 생명주기에 걸친 정책 행동 분석 역시 미흡하며, 학습 초기 탐색 과정에서의 안전 위반 위험에 대한 대책이 결여되어 있다.

3. 안전 강화학습(Safe RL)

안전 강화학습(Safe RL)은 보상 최대화와 제약 충족을 병행하는 프레임워크이다. CPO[5]는 신뢰 영역(trust region) 내에서 제약을 만족하는 정책 갱신을 보장하며, Lagrangian 기반 방법[6]

은 제약을 보상에 페널티로 변환한다. 사전 검증(pre-verification) 방식으로는 action masking과 안전 실드(safety shield)가 있다. Alshiekh 등[7]은 안전 실드의 개념을 최초로 제안하여, 에이전트가 선택한 행동을 실행 전에 형식적으로 검증하고 안전하지 않은 행동을 차단하는 구조를 확립하였다. 안전 실드의 개념을 제안한 연구[7]는 에이전트가 선택한 행동을 실행 전에 형식적으로 검증하고 안전하지 않은 행동을 차단하는 구조를 확립하였다. 이후 Corsi 등[20]은 형식 검증 기반으로 실드를 자동 합성하는 verification-guided shielding을 제안하여, 검증 유도 실딩(verification-guided shielding)은 형식 검증 기반으로 실드를 자동 합성하는 방향[20]을 제안하여, 수동 규칙 정의 없이도 안전 보장을 달성하는 방향으로 발전하였다. Yang 등[14]은 확률적 논리 실드(probabilistic logic shield)를 통해 확률적 센서 데이터 환경에서의 안전 제약 충족을 다루었다. 확률적 센서 데이터 환경에서의 안전 제약 충족[14]을 다루었다. Z3[8]와 같은 SMT 솔버는 복합 논리식의 만족 가능성을 효율적으로 판정하여, 합의 프로토콜의 안전 속성을 형식적으로 검증하는 데 적합하다.

제안하는 프레임워크는 기존 Safe RL 접근법과 다음 세 가지 측면에서 구분된다. (1) 선행 연구[1]에서 형식 검증된 하이브리드 합의 구조의 안전 속성을 안전 실드로 재구성하여 RL 학습 루프에 직접 통합하고, (2) Z3 정확 검증과 RuleBased 근사 검증의 동치성을 실증하며, (3) 학습된 DQN 에이전트가 안전 실드의 개입 없이도 안전 제약을 자율적으로 준수하는 현상을 검증한다. 기존의 RL 블록체인 연구[15][16]와 비교하면 형식 검증 기반 안전 보장이 추가되고, 기존의 DRL 샤딩 연구[17]와 비교하면 합의 안전 속성의 런타임 검증이 결합된다.

Ⅲ. 시스템 설계

1. QBHC 합의 아키텍처

제안하는 QBHC(QuadTree-Based Hybrid Consensus) 시스템은 선행 연구[1]에서 설계 및 검증된 RAFT-PBFT 이중 계층 합의 구조를 기반으로 한다. L1(로컬) 합의에서 각 샤드가 Raft를 통해 트랜잭션을 빠르게 합의하고, 합의된 proposal을 L2(글로벌) 합의인 PBFT로 최종 완결하는 구조이다. 선행 연구에서 이 합의 구조의 Safety 및 Liveness가 형식적으로 검증되었으므로, 합의 프로토콜 자체의 정합성을 전제로 하고 동적 파라미터 제어에 집중한다.

선행 연구에서는 샤드의 구성 방식이 고정되어 있었으나, 동적 파라미터 제어를 도입하려면 샤드 구조 자체도 제어 가능한 변수로 확장할 필요가 있다. L1 합의가 샤딩을 포함하는 구조이므로, 샤드 분할 방식에 따라 L1 합의의 효율성과 L2로 전달되는 proposal의 품질이 달라진다. 이에 QuadTree 기반 공간 분할을 도입하였다. QuadTree는 2차원 공간을 재귀적으로 4등분하는 계층적 자료구조로, 트랜잭션의 지리적 또는 논리적 인접성에 따라 동일 샤드로 배정함으로써 L1 합의의 지역성(locality)을 극대화한다. 그 결과 각 샤드에서 생성되는 proposal의 응집도가 높아지고, L2 PBFT가 GlobalBlock으로 최종화할 때 샤드 간 의존성이 줄어들어 합의 효율이 향상된다. QuadTree의 계층적 분할 구조는 L1-L2 합의 계층과 자연스럽게 대응되어, 샤드 깊이에 따른 합의 범위를 체계적으로 관리할 수 있다.

QuadTree의 분할 깊이(depth)는 RL 에이전트의 행동 공간에 포함된다(action 5: depth+1, action 6: depth-1). DQN 에이전트가 부하 상황에 따라 샤드 수를 동적으로 조절할 수 있으므로, 선행 연구의 정적 합의 구조를 적응적 합의 구조로 전환하는 토대가 된다. 시스템 아키텍처는 순수 함수형 코어(Functional Core)와 상태 관리 셸(Imperative Shell)의 분리 원칙을 따르

며, 합의 알고리즘, 보상 계산, 안전 검증 등 핵심 로직은 부작용 없는 순수 함수로 구현되어 테스트 용이성과 형식 검증 적합성을 갖추었다. 모든 데이터 타입은 불변(frozen) 데이터클래스로 정의되어 상태 변이로 인한 오류를 방지한다.

2. Safety Shield 설계

안전 실드는 DQN 에이전트가 선택한 행동을 실행 전에 검증하여, 합의 프로토콜의 안전 속성을 위반하는 행동을 사전 차단하는 사전 검증(pre-verification) 모듈이다. 선행 연구[1]에서 Z3/PySMT bounded model checking으로 검증한 Safety 속성들을 런타임 안전 실드로 재구성하였으며, 5개의 합의 안전 속성(P1-P5)을 정의하고 세 가지 실드 구현을 제공한다.

안전 속성 P1-P5는 다음과 같이 정의된다. (1) P1(합의 정족수)은 PBFT 위원회 크기가 Byzantine 내성 조건을 만족하는지 검증한다(n >= 3f+1). (2) P2(타임아웃 적정성)는 Raft 타임아웃이 네트워크 왕복 시간, 배치 처리 시간, 큐 대기 시간을 고려한 최소 임계값 이상인지 확인한다. (3) P3(PBFT 타임아웃)은 PBFT 타임아웃이 네트워크 왕복 시간의 4배 이상인지 검증한다. (4) P4(배치 유효성)는 배치 크기가 최소 1 이상인지 확인한다. (5) P5(샤드 깊이 경계)는 샤드 분할 또는 병합 행동이 허용된 깊이 범위 내에서 수행되는지 검증한다.

Z3Shield는 Z3 SMT 솔버를 이용하여 각 안전 속성을 논리식으로 인코딩한 후, 만족 가능성(satisfiability)을 정확히 판정하는 형식 검증 구현이다. P1의 경우 정수 산술 제약 n_pbft >= 3 × f_est + 1을 Z3 심볼로 표현하여 검증하며, P2는 실수 산술 부등식 timeout_raft >= 1.5 × (rtt × 3 + batch × 0.005 + queue × 0.001)을 인코딩한다. Z3Shield는 모든 속성에 대해 수학적으로 정확한 판정을 보장하지만, SMT 솔버 호출에 따른 지연시간이 발생한다.

RuleBasedShield는 동일한 5개 속성을 직접적인 산술 비교로 검증하는 다항식 상수 시간 근사 구현이다. P1-P5에 대해 Z3Shield와 동일한 부등식을 평가하되, SMT 솔버를 거치지 않고 Python 산술 연산으로 직접 계산한다. P1-P5의 수식 구조가 선형 또는 단순 비교이므로, RuleBasedShield는 Z3Shield와 수학적으로 해당 속성 범위 내에서 동일한 판정 결과를 산출한다.

NoShield는 모든 행동을 무조건 허용하는 기준선 구현으로, 안전 실드가 없는 환경에서의 성능과 안전성을 측정하기 위해 사용된다.

실드 프로토콜은 verify(action, state) -> ShieldResult 인터페이스를 따르며, ShieldResult는 행동 허용 여부(allowed), 위반된 속성 목록(violated_properties), 사유(reason)를 포함하는 불변 데이터클래스이다. 행동이 차단되면 에이전트는 무행동(action_id=0)을 대체 수행하며, 차단 사실은 다음 관측 벡터에 prev_blocked 플래그로 전달되어 에이전트의 학습에 반영된다.

3. DQN 에이전트 설계

DQN 에이전트는 구현한 QBHC 합의 시뮬레이션 환경에서 Gymnasium 인터페이스를 통해 학습한다. 관측 공간은 11차원 정규화 벡터로 구성되며, 평균 지연시간, 처리량, 메시지 오버헤드, 큐 길이, Raft 타임아웃, 배치 크기, PBFT committee 크기, PBFT 타임아웃의 8개 기본 차원에 phase 인덱스, phase 내 진행률, 이전 차단 여부의 3개 맥락 차원이 추가된다.

행동 공간은 11개 이산 행동으로 구성된다. 무행동(action 0), Raft 타임아웃 증감(actions 1-2), 배치 크기 증감(actions 3-4), 샤드 깊이 증감(actions 5-6), PBFT committee 크기 증감(actions 7-8), PBFT 타임아웃 증감(actions 9-10)이다. 각 행동은 현재 합의 파라미터에 대해 고정 단위(타임아웃 ±0.5, 배치 ±5, committee/깊이 ±1)의 증감을 적용하는 순수 함

수로 구현된다.

보상 함수는 정규화된 처리량에서 지연시간 패널티와 메시지 오버헤드 패널티를 감산하는 구조로 다음과 같이 정의 된다.

r = tps_norm × w_tps − lat_norm × w_lat − msg_norm × w_msg − blocked × w_shield

안전 실드에 의해 행동이 차단되면 추가 감점(shield_penalty)이 부과되어, 에이전트가 차단을 회피하는 방향으로 학습하도록 유도한다. 또한 전 스텝 대비 처리량 향상에 대한 보너스 항(delta bonus)이 추가된다.

4. 학습-배포 파이프라인

제안하는 프레임워크의 학습-배포 파이프라인은 두 단계로 구성된다. 학습 단계에서는 DQN 에이전트가 Z3Shield와 함께 학습하여 안전 제약을 포함한 정책을 습득한다. Z3Shield는 수학적으로 정확한 판정을 보장하므로, 에이전트는 정확한 안전 경계를 학습할 수 있다. 배포 단계에서는 학습된 정책에 RuleBasedShield를 래핑하여 런타임 안전성을 확보한다.

RuleBasedShield는 Z3Shield와 동치인 판정을 0.93 마이크로초의 지연으로 수행하므로, 실시간 합의 환경에서의 적용이 가능하다.

재현성 확보를 위해 결정론적 이산 사건 시뮬레이션(DES) 실행기(executor)를 사용하며, 시드 기반 의사 난수 생성을 통해 동일 시드에서 동일한 시뮬레이션 궤적이 재현되도록 하였다.

Ⅳ. 실험 설계

1. 실험 환경 및 시나리오

실험 환경은 강화학습용 합의 환경으로 구현된 시뮬레이터이며, Gymnasium 인터페이스를 통해 RL 에이전트와 상호작용 한다. 전체 에피소드는 500 스텝으로 구성되며, 6개 부하 구간(phase)이 순차적으로 진행된다.

표 1. 실험 시나리오 구간 정의

구간	스텝 수	TX/스텝	네트워크 지연범위(s)	비잔틴 노드	주요 특성
Baseline	80	4	0.00–0.02	0	기저부하, 정책초기화검증
Load Ramp	80	16	0.00–0.02	0	고부하전환, 처리량적응요구
Jitter Under Load	80	16	0.08–0.25	0	고부하+네트워크지연변동
Calm Before Storm	80	4	0.00–0.02	0	부하감소 후 안정구간
Byzantine Surge	100	16	0.03–0.08	1	Byzantine 장애+고부하
Recovery	80	4	0.00–0.02	0	Byzantine 해소, 정상복귀

6개 구간의 합은 500 스텝이며, Byzantine_Surge 구간에 100 스텝을 배정하여 공격 시나리오에 대한 충분한 관측을 확보하였다. 각 구간의 TX/스텝(트랜잭션 주입률)은 구간 전환 시 즉시 적용되는 계단 함수(step function)로, 구간 내에서는 일정하다. 이 6단계 시나리오는 실제 블록체인 네트워크에서 관찰되는 부하 생명주기를 모사한 것으로, 정상 운영에서 스트레스 상황을 거쳐 복원에 이르는 전체 과정을 포괄한다.

2. 실험 구성

프레임워크의 유효성을 검증하기 위해 4개의 실험을 설계하였으며, 각 실험은 독립된 연구 질문에 대응한다.

(1) 실험 A (Shield 비교): 안전 실드의 효과를 격리하기 위해, 무작위 정책 하에서 Z3Shield, RuleBasedShield, NoShield 세 가지 실드를 비교한다. 정책을 무작위로 고정함으로써 실드의 순수 효과만을 측정한다. 시드당 2,000 스텝, 5개 시드로 총 10,000 스텝을 수집하며, 처리량(TPS), 차단 횟수(blocked), 안전 위반 횟수(violations)를 측정한다. (2) 실험 B (DQN 학

습): Z3Shield 하에서 DQN 에이전트의 학습 수렴을 검증한다. 시드당 50,000 타임스텝(약 100 에피소드)의 학습을 수행하며, 에피소드별 보상, 처리량, 차단 횟수, 탐색률(epsilon)을 기록한다. 수렴 판정은 10 에피소드 이동 평균의 변동이 허용치(tolerance=0.1) 이하로 안정화되는 시점으로 정의한다. (3) 실험 C (정책 비교): 실험 B에서 학습된 DQN 모델을 로드하여 DQN, Random, Static 세 정책을 Z3Shield 하에서 비교 평가한다. 시드당 2,000 스텝, 5개 시드로 정책별 10,000 스텝을 수집하며, 처리량, 보상, 차단률, Jain 공정성 지수, 샤드 깊이, 메시지 오버헤드를 phase별로 측정한다. (4) 실험 D (SMT 오버헤드): Z3Shield, RuleBasedShield, NoShield의 검증 호출 지연시간을 측정한다. PBFT 위원회 크기 n=4(f=1일 때 최소 BFT 충족 조건)로 고정하고, 시드당 500회 검증 호출, 5개 시드로 실드당 2,500회 측정을 수행하며, P1-P5 속성만을 대상으로 한다.

3. 하이퍼파라미터

DQN 에이전트의 하이퍼파라미터는 [표 2]에 요약하였다.

표 2. DQN 하이퍼파라미터

Parameter	Value
Network architecture	2-layer MLP (128 x 128)
Learning rate	1e-3 -> 0 (linear decay)
Batch size	64
Replay buffer size	50,000
Target network update interval	250 steps
Exploration(e)	1.0 -> 0.05 (fraction 0.5)
Learning starts	1,000 steps
Discount factor(g)	0.99
Random seeds	42, 123, 456, 789, 1024

4. 통계 분석

정책 간 성능 차이의 통계적 유의성을 검증하기 위해 Bayesian 일반화 선형 혼합 모형(GLMM)을 적용하였다. 시드(Seed)를 랜덤 효과로, 정책(Policy)을 고정 효과로 설정하여 시드 간 변동을 통제하였다(DV ~ Policy + (1|Seed)). 사후 분포의 95% 최고밀도 구간(HDI, Highest Density Interval)와 방향 확률(Probability of Direction)을 기반으로 정책 쌍별 사후 비교를 수행하였다.

V. 실험 결과

1. Safety Shield의 효과

안전 실드의 효과를 격리 검증하기 위해, 무작위 정책 하에서 세 가지 실드를 비교한 결과를 [표 3]에 제시한다.

표 3. Shield별 안전성 및 성능 비교 (무작위 정책, 10,000 스텝)

Shield	Avg TPS (mean +/- std)	Blocked	Violations
NoShield	2.81 +/- 3.23	1,552	7,801
RuleBased	9.78 +/- 0.56	3,110	0
Z3	9.78 +/- 0.56	3,110	0

NoShield 조건에서는 10,000 스텝 중 7,801건의 안전 위반이 발생하였다. 무작위 정책이 선택한 행동의 78%가 P1-P5 중 하나 이상을 위반한 셈이다. NoShield의 평균 처리량은 2.81 TPS에 불과하였으며, 표준편차 3.23은 시드 간 극심한 성능 편차를 나타낸다. 시드별로 보면 시드 42의 처리량은 0.02 TPS, 시드 1024는 7.88 TPS로 편차가 매우 컸다. 주목할 점은 5개 시드의 위반률이 77.5-78.5%로 유사함에도 불구하고, 시드 간 TPS 편차가 390배에 달한다는 사실이다. 시드 42(위반 1,569건)와 시드 1024(위반 1,551건) 사이의 위반 횟수 차이는 18건에 불과하지만, 이 차이가 처리량을 7.88에서 0.02 TPS로 붕괴시켰다. 이는 안전 위반이 합의 상태를 정상 경로에서 이탈시킨 후, 이후의 모든 행동이 추가 위반을 유발하는 양성 피드백 루프를 형성하는 연쇄

실패(cascading failure)의 증거이며, 안전 실드의 조기 개입이 이 루프를 차단하는 데 핵심적임을 보여준다.

반면, Z3Shield와 RuleBasedShield는 안전 위반을 완전히 제거(0건)하면서 평균 처리량을 9.78 TPS로 3.5배 향상시켰다. 표 3의 Blocked 열은 두 종류의 차단을 합산한 값이다. NoShield의 1,552건은 파라미터 경계 제약(샤드 깊이 하한, 배치 크기 최소값 등)에 의한 것으로 안전 속성 검증과 무관하다. 안전 실드가 적용된 조건에서는 이 경계 차단에 P1–P5 검증 차단 1,558건이 추가되어 총 3,110건이 기록되었으며, 이 추가 차단이 7,801건의 안전 위반을 완전히 방지하였다. 차단이 증가했음에도 처리량이 오히려 향상된 것은, 안전 위반 시 합의 라운드 실패와 재시도에 소모되는 복구 비용이 차단의 기회비용(단일 스텝의 무행동 대체)보다 훨씬 크기 때문이다. 표준편차도 3.23에서 0.56으로 감소하여, 실드가 시드 간 성능 편차를 줄이는 안정화 효과를 가짐을 확인하였다.

Z3Shield와 RuleBasedShield는 평균 TPS, 표준편차, 차단 횟수, 위반 횟수의 모든 지표에서 5개 시드 전체에 걸쳐 소수점 이하까지 완전히 동일한 결과를 보였다. 시드별 세부 결과(시드 42: TPS 9.192, 차단 727; 시드 123: TPS 10.23, 차단 431 등)에서도 양 실드 간 차이가 없었다. P1--P5의 수식 구조가 선형 부등식이므로 Z3 SMT 솔버와 직접 산술 비교가 수학적으로 동일한 판정을 산출하며, 이번 실험이 이를 실증적으로 뒷받침한다.

2. Z3-RuleBased 동치성과 실시간 적용 가능성

두 실드의 판정 동치성이 확인된 상태에서, 실시간 적용 가능성을 평가하기 위해 검증 호출 지연시간을 측정하였다.

표 4. Shield별 검증 지연시간 (마이크로초, 실드당 2,500회 측정)

Shield	Mean	Median	P95	P99	Max
NoShield	0.47	0.46	0.50	0.63	14.6
RuleBased	0.93	0.88	1.04	1.71	10.0
Z3	8,232	8,167	8,741	9,159	53,958

RuleBasedShield의 평균 검증 지연시간은 0.93 마이크로초로, NoShield 대비 약 2배의 오버헤드만을 추가한다. P99 지연시간도 1.71 마이크로초에 불과하여, 합의 라운드 시간(통상 수백 밀리초)과 비교하면 무시할 수 있는 수준이다.

Z3Shield의 평균 지연시간은 8,232 마이크로초(약 8.2 밀리초)로, RuleBasedShield 대비 약 8,800배 느리다. Z3 솔버의 사분위 범위(IQR: 8,069–8,310 마이크로초)는 좁아 일관된 성능을 보이나, 최대값 53,958 마이크로초(약 54 밀리초)의 극단적 이상치가 2건(0.1%) 관찰되었다. 이 이상치는 Z3 솔버의 초기화 비용 또는 가비지 컬렉션 영향으로 추정된다. 이상의 결과는 학습–배포 파이프라인의 실용성을 뒷받침한다. 학습 단계에서는 Z3Shield의 정확한 판정으로 에이전트가 정확한 안전 경계를 학습하고, 배포 단계에서는 동치인 RuleBasedShield로 교체하여 마이크로초 수준의 실시간 검증을 수행할 수 있다.

3. DQN 학습 수렴

Z3Shield 하에서 DQN 에이전트의 학습 수렴을 5개 시드에 걸쳐 분석하였다.

표 5. 시드별 학습 수렴 분석

시드(Seed)	수렴 에피소드	최종 보상 평균
42	31	77.22
123	32	43.36
456	15	81.62
789	38	79.37
1024	0	78.16

수렴 판정은 10 에피소드 이동평균의 상대 변동(|ma_curr − ma_prev| / |ma_curr|)이 허용치

0.1 이하로 안정화되는 최초 시점으로 정의하였다. 시드 42, 123, 456, 789의 4개 시드는 에피소드 15-38 범위에서 수렴하여, 50,000 타임스텝의 학습 예산이 충분함을 확인하였다. 탐색률(e)은 약 에피소드 27에서 최종값 0.05에 도달하며, 수렴 시점과 대략 일치한다.

시드 1024는 수렴 에피소드가 0으로 기록되었다. 이는 학습 초기부터 이미 수렴한 것이 아니라, 첫 번째 윈도우 쌍(에피소드 1-10 vs 11-20)의 이동평균 상대 변동이 허용치 미만이었기 때문이다. 실제로 시드 1024의 초기 에피소드 보상은 -124.6에서 +1.1까지 큰 분산을 보였으나, 두 윈도우의 평균이 우연히 비슷하여 수렴 조건을 조기 충족하였다. 최종 보상 평균 78.16이 다른 시드(77-82)와 동등한 수준이므로 학습 자체는 정상적으로 진행되었으며, 이 현상은 이동평균 기반 수렴 판정의 한계로 이해할 수 있다.

시드 456은 에피소드 15에서 가장 빠르게 수렴하여 최종 보상 81.62를 달성하였고, 시드 789는 에피소드 38에서 가장 늦게 수렴하였으나 최종 보상 79.37로 양호한 결과를 보였다. 반면 시드 123은 최종 보상 평균이 43.36으로 다른 시드(77-82)에 비해 현저히 낮았다. 시드 123의 학습 궤적을 살펴보면, 학습 초기 에피소드 1에서 보상 -472.2의 극단적 하락을 경험한 후 국소 최적(local minimum)에 갇혀 충분히 회복하지 못한 것으로 분석된다. 또한 학습 초기의 안전 실드 대량 차단으로 누적된 음의 패널티가 이후 정책 수렴 방향을 제약한 것으로 해석할 수 있다.

수렴 이후(e=0.05)에도 간헐적으로 큰 보상 하락이 관찰되었다. 시드 789의 에피소드 66에서 보상 -481.6, 시드 123의 에피소드 59에서 -151.9 등이 대표적이다. 학습 안정성을 정량화하기 위해 최근 20 에피소드의 보상 변동계수(CV)를 측정한 결과, 시드 42(CV=0.04)와 456(CV=0.04)만이 안정적이었고, 시드 123(CV=0.58), 789(CV=0.92), 1024(CV=0.46)는 불안정 기준(CV>0.2)을 초과하였다. 시드 789는 수렴 후 4회의 보상 급락(에피소드 보상이 이동평균으로부터 2σ 이상 하락하는 급락)을 경험하였다. 이러한 수렴 후 불안정성은 e=0.05의 잔여 무작위 탐색이 Byzantine Surge 구간에서 연쇄적 안전 실드 차단을 유발하는 구조적 현상이며, 배포 시 e=0으로 설정하면 해소될 것으로 판단된다.

4. 정책 비교: Phase별 적응

학습된 DQN 모델을 포함하여 세 정책의 phase별 성능을 비교하였다.

표 6. Phase별 DQN vs Static vs Random 성능 비교

구간	DQN TPS	Static TPS	Delta TPS	DQN Reward	Static Reward	Delta Reward
Baseline	4.0	4.0	+0.0%	+0.088	+0.088	+0.000
Load Ramp	15.8	10.0	+58.0%	+0.673	+0.378	+0.294
Jitter Under Load	16.2	10.0	+62.0%	−0.212	−0.517	+0.305
Calm Before Storm	4.0	10.0	−60.0%	+0.067	+0.376	−0.309
Byzantine Surge	16.0	10.0	+60.0%	−0.017	−0.113	+0.096
Recovery	4.0	10.0	−60.0%	+0.082	+0.375	−0.293

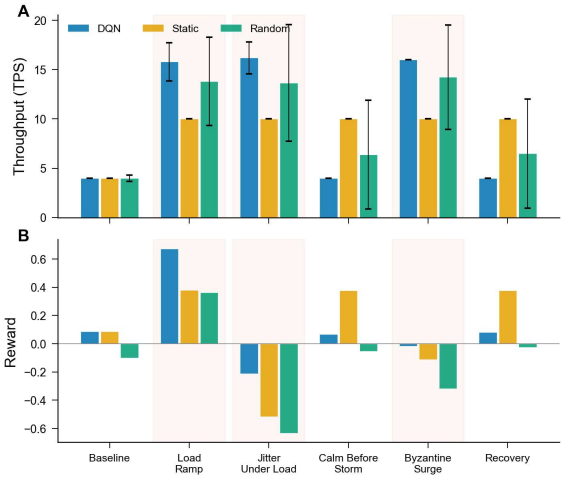

그림 1. Phase별 정책 비교: (A) 처리량(TPS), (B) 에피소드 보상

DQN은 스트레스 구간(Load Ramp, Jitter Under Load, Byzantine Surge)에서 TPS를 15.8-16.2로 높여 Static(10.0) 대비 58-62%의 처리량 향상을 달성하였다. Load Ramp에서의 보상 차이(+0.294)와 Jitter Under Load에서의 보상 차이(+0.305)는 DQN이 부하 증가 패턴을 학

습하여 효과적으로 대응하였음을 나타낸다. Byzantine Surge에서도 DQN의 보상이 Static 대비 0.096 높아, 공격 상황에서의 적응적 대응 능력을 확인할 수 있었다. 특히 Calm Before Storm → Byzantine Surge 전환 시 DQN의 보상 delta는 +0.487인 반면 Static은 -0.497로, DQN이 공격 국면 전환에 이미 준비된 상태임을 보여준다.

수렴한 DQN 정책의 행동 분포를 분석한 결과, 전체 행동의 97.8%가 무행동(NOOP)이었다. 파라미터 변경 행동은 Load Ramp 구간에 집중(활성 행동률 8.7%)되었으며, 나머지 5개 구간에서는 1.2% 이하였다. 샤드 깊이 변경은 단 한 건도 발생하지 않아, DQN이 depth=1을 유지하면서 TPS 조정만으로 적응하는 최소 개입 전략을 학습했음을 알 수 있다. 이는 DQN이 환경이 자체적으로 안정화되는 구간에서는 개입을 억제하고, 부하 급변 시에만 선택적으로 파라미터를 조정하는 정책을 습득했음을 의미한다.

안정 구간(Calm Before Storm, Recovery)에서는 DQN의 TPS가 4.0으로 Static(10.0)보다 낮았다. DQN이 부하 감소 구간에서 보수적으로 행동하는 정책 관성(policy inertia)을 보인 것이며, 보상 함수에서 탈에스컬레이션 인센티브가 부족한 데 기인한다. 다만 이 현상은 프레임워크 자체의 한계가 아닌 보상 설계(reward engineering) 문제이다. 한편, 안정 구간에서의 낮은 TPS는 입력 부하 대비 처리 여유를 감소시킬 수 있으며, 이는 이후 급격한 부하 증가 시 대응 여력에 영향을 줄 가능성이 있다. 따라서 실배포 시에는 탈에스컬레이션 유인을 보상 함수에 명시적으로 포함하는 것이 권장된다.

Bayesian GLMM 분석에서 보상 변수에 대한 정책 간 대조 결과, DQN은 Random 대비 99.98%의 확률로 우위를 보였고(평균 차이 +0.242, HDI [0.139, 0.346]), Static도 Random 대비 99.96%의 확률로 우위를 보였다. DQN과 Static 간의 전체 보상 차이는 방향 확률 64.76%로 통계적 유의성은 제한적이었으나, 스트레스 구간의 우위와 안정 구간의 열위가 전체 평균에서 상쇄된 결과이다.

5. Safety Constraint 내재화

프레임워크의 주요 발견으로, DQN 에이전트의 안전 실드 차단률을 시드별로 분석하였다.

표 7. 정책별 시드별 안전 실드 차단률 (%)

정책	Seed 42	Seed 123	Seed 456	Seed 789	Seed 1024	평균
DQN	0.0	22.0	0.0	0.0	0.0	4.4
Random	25.0	21.4	29.3	29.7	35.4	28.2
Static	0.0	0.0	0.0	0.0	0.0	0.0

DQN의 전체 평균 차단률은 4.4%이나, 시드 123의 22.0%가 이 수치를 지배한다. 나머지 4개 시드(42, 456, 789, 1024)에서는 차단이 단 한 건도 발생하지 않았다. DQN이 TPS를 4에서 16으로 동적으로 전환하는 적극적 정책을 수행하면서도, 안전 실드의 제약 경계를 학습하여 위반 행동 자체를 회피하는 정책을 습득한 것이다.

Static 정책의 0.0% 차단률은 고정 TPS=10이 항상 안전 범위 내에 있으므로 예측 가능한 결과이다. DQN은 행동 공간을 적극적으로 탐색하면서도 4/5 시드에서 0%를 달성했다는 점에서 질적으로 다르다. Random 정책의 28.2% 차단률과 대조하면, DQN의 안전 제약 내재화가 더욱 두드러진다.

시드 123의 예외적 행동은 해당 시드의 DQN 모델이 Byzantine Surge 구간에서 비적응적 정책을 학습한 결과이다. 시드 123의 Byzantine Surge 구간 보상은 -0.823으로, 다른 시드(+0.179 ~ +0.189)와 큰 차이를 보였다. 구간별 차단 내역을 분석하면, 시드 123은 Byzantine Surge 400스텝 전체가 차단되어 해당 구간 차단률이 100%에 달하였고, Calm Before Storm에서 32건(10%), Recovery에서 8건(2.5%)이 추가로 차단되었다. 이는 시드 123의 DQN이 Byzantine

공격 구간에 대한 적응에 완전히 실패하여, 매 스텝 안전 위반 행동을 선택한 것을 의미한다. 이 결과는 DQN 학습의 시드 민감성(seed sensitivity)을 보여주는 결과로, 다중 시드 앙상블 또는 robust training 기법을 통한 개선이 필요하다.

이 해석을 뒷받침하는 추가 근거로, 학습 과정에서 에피소드별 차단 횟수(blocked count)와 보상 사이의 Pearson 상관계수는 r=-0.989로 극도로 강한 음적 상관을 보였으며, 5개 시드 모두에서 $r < -0.98$이었다. 차단 횟수와 에피소드 보상 사이의 이 극도로 강한 음적 상관은, 보상 최대화와 안전 제약 준수라는 두 목적이 학습 과정에서 사실상 정렬되어 있음을 보여준다.

안전 실드의 역할을 종합하면, 실드는 단순한 런타임 안전장치를 넘어 학습 과정의 가드레일로 기능한다. 학습 중 실드의 차단과 패널티 신호가 에이전트에게 안전 경계의 위치를 알려주고, 에이전트는 이 신호를 바탕으로 점차 실드의 개입 없이도 안전한 행동을 선택하게 된다. 이것이 안전 실드가 불필요해졌다는 뜻은 아니다. 실드가 학습을 올바르게 유도한 결과이다.

Ⅵ. 분석 및 시사점

1. 안전성-성능 트레이드 오프의 부재

가장 주목할 결과는 안전 실드가 성능을 저하시키지 않고 오히려 향상시켰다는 점이다. 무작위 정책 하에서 실드는 1,558건의 안전 위반 행동을 추가로 차단하였음에도 처리량이 2.81에서 9.78 TPS로 증가하였다. 합의 프로토콜 환경에서는 안전 위반의 연쇄 효과(cascading effect)가 차단의 기회비용을 압도하기 때문이다. 위반이 발생하면 합의 라운드 실패, 재시도, 타임아웃 등의 복구 비용이 누적되어 시스템 전체의 처리량을 급격히 저하시키는 반면, 위반 행동을 무행동으로 대체하는 비용은 단일 스텝의 진행 포기에 불과하다. 표준편차가 3.23에서 0.56으로 감소한 것도 이 해석을 뒷받침하는데, 위반의 연쇄 효과가 시드별로 다른 시점에서 발생하여 성능 편차를 증폭시키기 때문이다.

학습-배포 파이프라인의 실용성 측면에서, Z3와 RuleBased의 판정 동치성과 RuleBased의 마이크로초 지연시간(0.93 µs)은 형식 검증 수준의 런타임 안전성이 합의 라운드에 무시할 수 있는 오버헤드로 달성 가능함을 보여주었다.

DQN이 4/5 시드에서 차단률 0%를 기록한 것은 안전 실드가 단순한 행동 필터를 넘어, 학습 과정에서 에이전트의 정책 자체를 안전한 방향으로 형성하는 가드레일로 작용한 결과이다.

DQN의 적극적 TPS 조정이 공정성에 미치는 영향도 주목할 필요가 있다. TPS와 Jain 공정성 지수 사이의 Pearson 상관계수는 DQN에서 r=-0.25, Random에서 r=-0.44로, DQN이 Random 대비 절반 수준의 공정성 비용으로 TPS 최적화를 달성하였다. DQN의 고TPS 구간(TPS>15)에서도 Jain 지수는 0.86 이상을 유지하여, 처리량과 공정성 사이의 트레이드 오프가 실용적으로 수용 가능한 범위에 있음을 확인하였다.

2. Z3-RuleBased 동치성의 조건과 한계

P1-P5에 대해 관찰된 Z3 기반 SMT 실드와 RuleBased 실드의 동치성은, 이 속성들이 모두 선형 부등식 형태로 표현 가능하다는 점에서 기인한다. Z3 SMT 솔버가 판정하는 논리식과 RuleBased가 수행하는 직접 산술 비교가 이 영역에서는 수학적으로 등가이다. 다만 이러한 동치성은 P1-P5와 quadtree depth < 2 설정에 국한되며, 보다 복잡한 구조로 확장될 경우에는 자동으로 보장되지 않는다. 따라서 본 프레임워크의 적용 범위는 선형 속성 P1-P5와 quadtree depth ≤ 2로 한정된다. 비선형 속성이나 더 깊은 샤딩 구조로 확장 시 두 실드 간 판정 불일치가 발생할 수 있으므로, 동치성 재검증 또는

Z3Shield 직접 배포가 권장된다.

한편, 쿼드트리 기반 QBHC 시스템에서는 쿼드트리 깊이에 따라 메시지 복잡도와 토폴로지 인식 타임아웃이 비선형적으로 증가하는 추가 안전 속성을 정의할 수 있다. 이러한 비선형 추가 안전 속성에 대해서는 Z3 기반 SMT 실드가 정확한 비선형 구조를 인코딩하는 반면, RuleBased 실드는 계산 비용을 줄이기 위해 선형 또는 다항식 수준의 근사를 적용하는 설계가 자연스럽다. 이 경우 quadtree depth가 커질수록 두 실드의 판정이 점진적으로 발산할 가능성이 있으며, 이러한 표현력 - 비용 트레이드오프에 대한 보다 심화된 분석은 향후 연구 과제로 남는다.

3. 정책 관성과 보상 함수 설계

DQN이 안정 구간에서 TPS=4에 고착되는 정책 관성 현상은 현재 보상 함수의 구조적 특성에서 비롯된다. 현재 보상 함수는 처리량 증가에 양의 인센티브를 부여하고 실드 차단에 감점을 부과하지만, 부하 감소 이후 적정 수준으로 복귀하는 데 대한 명시적 보상 항이 없다. 이로 인해 DQN은 안전 범위 내에서 가장 보수적인 행동(무행동 또는 최소 TPS)을 선택하는 경향이 나타난다. 이 현상은 프레임워크의 구조적 한계가 아닌 보상 설계(reward engineering)의 과제로, 탈에스컬레이션 구간에서 목표 TPS와의 편차를 패널티로 반영하는 보상 항을 추가하면 개선이 가능하다.

Ⅳ. 결 론

본 연구는 형식 검증된 하이브리드 합의 구조[1] 위에 SMT 기반 안전 실드와 DQN을 통합한 Safe RL 프레임워크를 제안하였다. 4개의 실험에서 도출된 핵심 결론은, 안전 제약의 사전 적용이 성능을 희생시키지 않고 오히려 위반 복구 비용을 제거하여 순 처리량을 3.5배 향상시킨다는 점이다. 이는 블록체인 합의처럼 위반의 연쇄 비용이 큰 환경에서 Safe RL이 특히 유효한 접근임을 보여준다.

그러나 다음의 한계가 존재한다. (1) 단일 시뮬레이터 환경에서 수행된 결과이며, 실제 분산 네트워크에서의 네트워크 파티셔닝, 노드 장애의 확률적 분포, 크로스 샤드 트랜잭션 비율 변화 등의 요인이 결과에 미치는 영향은 추가 검증이 필요하다. (2) 5개 시드의 통계적 검정력이 제한적이며, 시드 123의 이상치가 전체 결과에 미치는 영향이 크다. (3) Bayesian GLMM 분석에서 보상 외 6개 종속변수(TPS, 지연시간, 차단률, Jain 공정성, 샤드 깊이, 메시지 오버헤드)에 수렴 문제(divergence)가 발생하여, 해당 변수에 대한 통계적 추론은 기술통계 수준으로 제한된다. 관측치 수의 부족과 일부 변수의 이봉(bimodal) 분포가 원인으로 판단된다. (4) e=0.05의 잔여 탐색에 의한 수렴 후 불안정성이 관찰되었으며, e=0 배포 환경에서의 별도 평가가 필요하다.

실용적 관점에서, Z3 검증과 규칙 기반 근사 검증의 P1–P5 동치성은 학습 시 정확한 안전 경계를 학습하면서도 배포 시 마이크로초 수준의 실시간 검증이 가능한 파이프라인 구조를 뒷받침한다. DQN이 4/5 시드에서 차단률 0%를 달성하면서 스트레스 구간에서 58–62%의 처리량 향상을 보인 것은, 안전 실드가 정책 학습 자체를 안전한 방향으로 유도하는 부가적 효과를 보여준다.

안정 구간의 정책 관성과 시드 123의 Byzantine 구간 100% 차단은 보상 함수 설계와 학습 안정성에 대한 후속 연구의 필요성을 제기한다. 향후 연구에서는 비선형 속성에서의 Z3–RuleBased 발산 특성 분석, 탈에스컬레이션 보상 항 설계, 다중 시드 앙상블을 통한 robust training, 그리고 실제 BFT 네트워크 환경에서의 검증을 수행할 계획이다.

REFERENCES

[1] 이지운, and 서희석. "고성능· 비잔틴 내성을 위한 RAFT-PBFT 하이브리드 합의 구조의 설계 및 검증." *스마트미디어저널,* 제14권, 제8호, 91-101쪽, 2025년

[2] Valadarsky, Asaf, et al. "Learning to route." *Proceedings of the 16th ACM workshop on hot topics in networks.* 2017.

[3] Huang, Liang, Suzhi Bi, and Ying-Jun Angela Zhang. "Deep reinforcement learning for online computation offloading in wireless powered mobile-edge computing networks." *IEEE Transactions on Mobile Computing* 19.11 (2019): 2581-2593.

[4] Mao, Hongzi, et al. "Resource management with deep reinforcement learning." *Proceedings of the 15th ACM workshop on hot topics in networks.* 2016.

[5] Achiam, Joshua, et al. "Constrained policy optimization." *International conference on machine learning.* Pmlr, 2017.

[6] Garcıa, Javier, and Fernando Fernández. "A comprehensive survey on safe reinforcement learning." *Journal of Machine Learning Research* 16.1 (2015): 1437-1480.

[7] Alshiekh, Mohammed, et al. "Safe reinforcement learning via shielding." *Proceedings of the AAAI conference on artificial intelligence.* Vol. 32. No. 1. 2018.

[8] De Moura, Leonardo, and Nikolaj Bjørner. "Z3: An efficient SMT solver." *International conference on Tools and Algorithms for the Construction and Analysis of Systems.* Berlin, Heidelberg: Springer Berlin Heidelberg, 2008.

[9] V. Buterin, "Ethereum 2.0 specifications," *Ethereum Foundation*, 2020.

[10] Kokoris-Kogias, Eleftherios, et al. "Omniledger: A secure, scale-out, decentralized ledger via sharding." *2018 IEEE symposium on security and privacy* (SP). IEEE, 2018.

[11] Zamani, Mahdi, Mahnush Movahedi, and Mariana Raykova. "Rapidchain: Scaling blockchain via full sharding." *Proceedings of the 2018 ACM SIGSAC conference on computer and communications security.* 2018.

[12] Wang, Jiaping, and Hao Wang. "Monoxide: Scale out blockchains with asynchronous consensus zones." *16th USENIX symposium on networked systems design and implementation (NSDI 19).* 2019.

[13] Mnih, Volodymyr, et al. "Human-level control through deep reinforcement learning." *nature* 518.7540 (2015): 529-533.

[14] F. Yang, D. Lechner, T. A. Henzinger, M. Lukina, and E. Bartocci, "Safe reinforcement learning via probabilistic logic shields," in *Proc. IJCAI,* pp. 5739-5749, 2023.

[15] Gutierrez R, Villegas-Ch W, Govea J. Adaptive consensus optimization in blockchain using reinforcement learning and validation in adversarial environments. *Front Artif Intell.* 2025 Sep 30;8:1672273.

[16] Y. Liu, L. Dong, and X. Wang, "An explainable deep reinforcement learning algorithm for the parameter configuration and adjustment in the consortium blockchain," *Engineering Applications of Artificial Intelligence,* vol. 128, 107541, 2024.

[17] Li, Pengze, et al. "Spring: Improving the throughput of sharding blockchain via deep reinforcement learning based state placement." *Proceedings of the ACM Web Conference 2024.* 2024.

[18] Kawahara, Ryo. "Verification of customizable blockchain consensus rule using a formal method." *2020 IEEE International Conference on Blockchain and Cryptocurrency (ICBC).* IEEE, 2020..

[19] Yu, Yisong, et al. "Model Checking Nondeterministic Behaviours in the Tendermint Byzantine Fault Tolerant Blockchain Consensus Protocol." *International Conference on Engineering of Complex Computer Systems.* Cham: Springer Nature Switzerland, 2025.

[20] D. Corsi, G. Amir, R. Bloem, G. Katz, B. Konighofer, and S. Niekum, "Verification-guided shielding for deep reinforcement learning," *in Proc. Reinforcement Learning Conference (RLC)*, 2024.

저 자 소 개

이지운(정회원)

2016년 서강대학교 정보통신대학원 석사 졸업

2021년 한국기술교육대학교 컴퓨터공학과 박사 졸업

2021년 한양대학교 산학협력단(서울) 연구원

2022년 ~ 현재: 한국기술교육대학교 미래융합학부 기술연구원

<주관심분야 : 인공지능, 강화학습, 블록체인, 클라우드 컴퓨팅, 정보보안>