Mira: AI 출력물의 신뢰도를 높이는 탈중앙화 검증 네트워크

Ninad Naik
ninad@arohalabs.com

Sidhartha Doddipalli
sid@arohalabs.com

Karan Sirdesai
karan@arohalabs.com

AI는 그럴듯한 결과물을 생성하는 데 뛰어나지만, 대규모 언어 모델이나 확산 모델과 같은 신경망 기반 기술의 확률적 특성으로 인해 종종 부정확한 정보를 만들어냅니다. 이에 본 문서에서는 탈중앙화된 합의 방식으로 AI 생성 결과물의 신뢰성을 검증하는 네트워크를 소개합니다. 이 네트워크는 AI 결과물을 독립적으로 검증 가능한 형태로 가공하여, 여러 AI 모델이 각 주장의 진위를 종합적으로 판단할 수 있도록 합니다. 추론 기반 검증을 수행하는 노드 운영자들은 정직한 검증을 장려하기 위해, 하이브리드 작업 증명/지분 증명 방식에 따라 경제적 보상을 받게 됩니다. 더 나아가, 저희는 오류 없는 결과물을 제공하는 합성 기반 모델로 발전하는 것을 목표로 합니다. 이러한 기반 시설은 AI 시스템이 인간의 개입 없이 자율적으로 작동할 수 있도록 하는 중요한 발판이며, AI가 사회 전반에 걸쳐 혁신적인 잠재력을 실현하는 데 필수적인 조건입니다.

1. 서론

인공지능은 인쇄술, 증기기관, 전기, 인터넷과 같이 인류 문명을 근본적으로 재편성한 기술에 비견될 만큼 혁신적인 힘을 지니고 있습니다. 그러나 현재의 AI는 혁명적 잠재력을 실현하는 데 있어 본질적인 어려움에 직면해 있습니다. AI는 창의적이고 설득력 있는 결과물을 생성하는 데는 뛰어나지만, 오류 없는 결과물을 안정적으로 제공하는 데는 한계가 있습니다. 이러한 제약으로 인해 AI는 주로 인간의 감독이 필요한 작업이나 챗봇과 같이 중요도가 낮은 분야에 제한적으로 활용되고 있으며, 중요한 상황에서 자율적으로 실시간 작업을 처리할 수 있는 AI의 잠재력을 충분히 발휘하지 못하고 있습니다.

AI 신뢰성은 핵심적인 걸림돌입니다. AI 시스템은 환각(실제로 존재하지 않는 정보를 생성하는 현상)과 편향이라는 두 가지 주요 유형의 오류를 겪으며, 이는 모델의 전체 오류율에 영향을 미칩니다. 현재의 오류율은 중요한 상황에서 자율적으로 운영하기에는 여전히 너무 높아, AI의 이론적 역량과 실제 적용 사이에 간극을 발생시킵니다. AI 모델이 더

많은 훈련 데이터와 매개변수화를 통해 지속적으로 발전하고 있음에도 불구하고, 이러한 신뢰성 문제는 훈련 딜레마로 인해 해결되지 않고 있습니다. 이 딜레마는 고전적인 정밀도-재현율(Precision-Recall) 트레이드오프를 반영합니다. 여기서 환각은 정밀도 오류(모델 출력의 일관성 부족)를, 편향은 재현율 오류(실제 사실로부터의 체계적인 왜곡)를 나타냅니다.

모델 개발자가 정밀도를 높이고 환각을 줄이기 위해 훈련 데이터를 선별적으로 선택하면, 선택 과정에서 불가피하게 재현율 오류(편향)가 발생합니다. 반대로, 재현율을 높이기 위해 다양하고 잠재적으로 상충하는 데이터 소스로 훈련하면 모델이 일관성 없는 출력을 생성하여 정밀도가 떨어지고 환각이 증가합니다. 특정 영역에 최적화된 모델은 좁은 범위 내에서 높은 신뢰성을 보이지만, 새로운 지식을 안정적으로 통합하는 데 어려움을 겪는다는 연구 결과가 있습니다. 이는 기존 지식과 일치하는 훈련 데이터보다 새로운 정보를 포함하는 훈련 데이터가 학습 효과가 떨어진다는 것을 의미합니다.

특정 영역에 최적화된 모델은 또한 훈련 영역 외부의 예외적인 상황이나 예측 불가능한 시나리오에 적절히 대응하지 못하므로, 다양한 실제 환경을 처리해야 하는 자율 시스템에는 적합하지 않습니다. 이러한 근본적인 제약은 AI 모델 성능에 고유한 한계를 설정합니다. 즉, 규모나 아키텍처와 관계없이 어떤 단일 모델로도 극복할 수 없는 최소 오류율이 존재합니다.

단일 모델로는 환각과 편향을 모두 최소화할 수 없지만, 집단 지성을 활용하면 해결책을 찾을 수 있습니다. 합의 메커니즘을 통해 여러 모델이 협력하면 개별 모델로는 불가능한 결과를 얻을 수 있습니다. 즉, 집단 검증을 통해 환각을 제거하고 다양한 관점을 통해 개별 편향을 상쇄할 수 있습니다. 이는 신뢰할 수 있는 AI를 위해서는 더 나은 모델뿐만 아니라 모델들의 강점을 결합하고 약점을 보완할 수 있는 더 나은 방법이 필요하다는 것을 시사합니다.

그러나 중앙 집중식 통제하에 모델 앙상블을 구축하는 것만으로는 신뢰성 문제를 완전히 해결할 수 없습니다. 모델 선택 과정 자체가 체계적인 오류를 야기합니다. 즉, 중앙화된 큐레이터의 선택은 불가피하게 특정 관점과 편향을 반영합니다. 더욱이, 많은 진실은 문화, 지역, 영역에 따라 다양하게 변화하는 맥락에 따라 달라집니다. 진정한 신뢰성은 단순히 여러 모델의 결합이 아닌, 분산된 참여를 통해 확보되는 진정으로 다양한 관점에서 비롯됩니다.

따라서 필요한 것은 중앙화된 권위가 아닌 분산된 합의에 기반한 AI 검증 시스템입니다. 이러한 시스템은 단일 신뢰 기관에 의존하지 않고 AI가 생성한 모든 결과물을 검증할 수 있도록 지원합니다. 합의를 조작하는 데 막대한 비용과 노력이 필요한 시스템은 사용자를

신뢰할 수 없는 결과물로부터 보호하고, 전문화된 영역 모델과 다양한 관점을 대변하는 모델 개발을 장려할 것입니다.

본 문서에서는 AI 결과물의 유효성에 대한 증거를 생성하기 위해 다양한 AI 검증자로 구성된 블록체인 기반 네트워크를 활용하여 AI 신뢰성 문제에 대한 해결책을 제시합니다. 네트워크의 보안 프레임워크는 경제적 인센티브, 기술적 안전장치, 게임 이론적 원칙의 조합을 통해 신뢰할 수 있는 검증을 보장합니다. 이러한 접근 방식은 분산 검증 메커니즘을 통해 AI 신뢰성을 향상시키고, 이는 편향과 환각 비율을 모두 감소시킵니다.

2. 네트워크 아키텍처

미라 네트워크는 복잡한 콘텐츠를 독립적으로 검증 가능한 주장 단위로 분해하는 혁신적인 프로토콜을 통해 AI 생성 결과물의 신뢰성을 획기적으로 높입니다. 이렇게 분해된 개별 주장은 다양한 AI 모델 간의 분산된 합의 과정을 거쳐 검증되며, 노드 운영자들은 정직한 검증 수행에 대한 경제적 보상을 받습니다. 이러한 탈중앙화된 구조는 검증 결과의 임의 조작을 방지하고, AI 생성 결과물에 대한 신뢰성 높은 검증을 보장합니다.

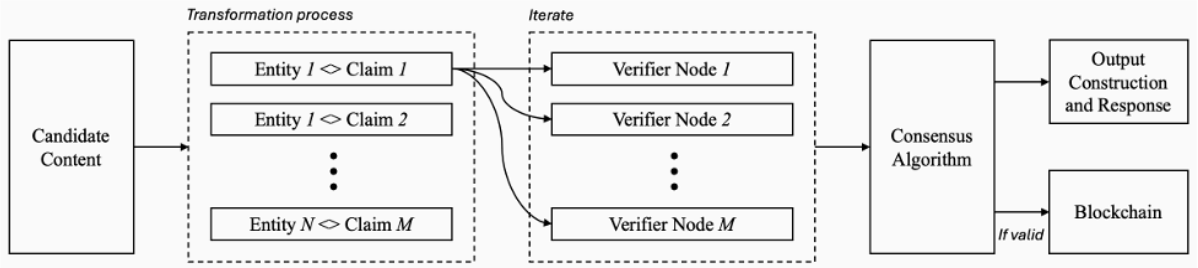
미라 네트워크의 시스템 아키텍처는 콘텐츠 변환, 분산 검증, 합의 도출 메커니즘을 혁신적으로 결합하여 신뢰성 있는 검증을 구현합니다. 이 시스템은 단순한 사실 정보에서부터 기술 문서, 법률 문서, 창작물, 멀티미디어 콘텐츠, 소프트웨어 코드 등 다양한 형태의 정보를 처리할 수 있습니다.

예를 들어, "지구는 태양을 중심으로 공전하고, 달은 지구를 중심으로 공전한다."라는 문장을 생각해 봅시다. 이 문장을 여러 AI 모델을 통해 검증하는 것은 비교적 간단하지만, 문서 전체, 법률 계약서, 프로그램 코드와 같이 복잡하고 긴 콘텐츠를 검증하는 것은 훨씬 더 어려운 문제입니다. 검증할 콘텐츠를 AI 모델에 직접 입력하면 각 모델이 내용의 서로 다른 부분을 자의적으로 해석하고 검증할 수 있습니다. 따라서 체계적인 검증을 위해서는 모든 AI 모델이 동일한 관점에서 동일한 문제를 다루도록 AI 결과물을 표준화하는 과정이 필수적입니다.

미라의 콘텐츠 변환 방식은 바로 이 문제를 해결합니다. 앞서 제시된 예시 문장의 경우, 시스템은 문장을 (1) "지구는 태양을 중심으로 공전한다."와 (2) "달은 지구를 중심으로 공전한다."와 같이 독립적으로 검증 가능한 개별 주장으로 분해합니다. 그런 다음 앙상블 검증을 통해 각 주장의 진위 여부를 확인하고, 검증 결과를 증명하는 암호화된 인증서를 발급합니다. 이 과정은 AI가 생성했는지 사람이 생성했는지에 관계없이 모든 콘텐츠에

동일하게 적용될 수 있으므로, 콘텐츠의 출처와 무관하게 엄격한 검증 기준을 유지할 수 있습니다.

미라 네트워크는 콘텐츠 변환, 주장 분배, 합의 관리, 네트워크 운영을 담당합니다. 노드 구조는 검증 모델을 실행하고, 주장을 처리하며, 검증 결과를 제출하는 독립적인 운영자들로 구성됩니다. 각 노드는 자율적으로 운영되지만, 네트워크 참여를 위해서는 일정 수준 이상의 성능과 신뢰도를 유지해야 합니다.



검증 과정은 다음과 같이 체계적으로 진행됩니다. 사용자는 검증을 요청할 콘텐츠를 제출하고, 의료, 법률 등 특정 분야, 절대적 합의, N/M 합의 등 합의 기준과 같은 검증 요건을 설정합니다. 네트워크는 콘텐츠를 검증 가능한 주장 단위로 변환하고 논리적 관계를 유지한 채, 검증을 위해 노드에 주장을 분배하고, 결과를 취합하여 합의를 도출합니다. 그런 다음, 각 주장에 대해 어떤 모델들이 합의에 도달했는지 등을 기록한 암호화폐 인증서를 생성하여 검증 결과와 함께 사용자에게 전달합니다.

3. 경제적 보안 모델

미라 네트워크의 경제적 보안 모델은 작업 증명(PoW)과 지분 증명(PoS) 메커니즘을 결합하여 정직한 검증에 대한 지속 가능한 인센티브를 제공하고, 실질적인 경제적 가치를 창출하여 분배합니다. 이러한 하이브리드 방식은 AI 결과물 검증에서 발생하는 특수한 문제들을 해결합니다.

네트워크는 검증을 통해 AI 오류율을 감소시켜 실질적인 경제적 가치를 창출합니다. 사용자들은 검증된 결과물을 얻기 위해 네트워크 수수료를 지불하고, 네트워크는 이러한 수수료를 노드 운영자와 데이터 제공자와 같은 참여자들에게 검증 보상으로 분배합니다.

기존 블록체인 네트워크에서 작업 증명(PoW)은 극히 낮은 확률로 암호화 퍼즐을 푸는 방식이지만, 미라 네트워크는 검증을 표준화된 객관식 문제로 변환합니다. 이러한 표준화는 노드 전체에서 체계적인 검증을 가능하게 하지만, 선택 가능한 응답의 확률 공간이

제한된다는 근본적인 문제 또한 야기합니다. 예를 들어, 검증 작업이 이진 선택(예/아니오)인 경우 무작위로 정답을 맞힐 확률은 50%이고, 4지선다형인 경우 25%입니다. 이로 인해 무작위 추측이 잠재적으로 매력적인 전략이 될 수 있으며, 실제 연산 없이도 높은 보상을 얻을 가능성이 있습니다.

이러한 문제를 완화하기 위해 노드는 검증에 참여하기 위해 일정량의 가치를 스테이킹(예치)해야 합니다. 노드가 지속적으로 합의에서 벗어나거나, 실제 추론보다는 무작위 응답을 하는 경향을 보이는 경우, 해당 스테이킹된 가치는 삭감(슬래싱)될 수 있습니다. 이러한 경제적 페널티는 무작위 응답으로 시스템을 속이려는 시도가 경제적으로 불합리하도록 보장합니다.

따라서 네트워크의 경제적 보안 모델은 세 가지 기본 원칙에 따라 작동합니다. 첫째, 노드 운영자는 스테이킹한 가치가 삭감 페널티를 통해 손실될 위험에 처할 수 있으므로 경제적 유인에 합리적으로 반응합니다. 둘째, 정직한 운영자가 스테이킹된 가치의 과반수를 통제하는 한 네트워크 보안이 유지되어 조작 시도가 경제적으로 매우 많은 비용이 들게 됩니다. 셋째, 네트워크가 확장됨에 따라 검증자 모델의 자연스러운 다양성이 통계적 편향을 줄여줍니다. 다양한 모델이 서로 다른 훈련 방식과 지식 기반을 갖추기 때문입니다. 이러한 원칙은 서로를 강화합니다. 경제적 유인은 다양한 참여자를 유치하고, 다양한 관점은 보안을 강화하며, 이는 다시 경제 모델을 지원합니다.

Table 1 은 답변 선택지의 수에 따라 정답을 성공적으로 추측할 확률을 보여줍니다**.**

검증	**2**개의 답변 옵션	**4**개의 답변 옵션	**6**개의 답변 옵션	**8**개의 답변 옵션	**10**개의 답변 옵션
1	50.0000%	25.0000%	16.6667%	12.5000%	10.0000%
2	25.2500%	6.2500%	2.7778%	1.5625%	1.0000%
3	12.5000%	1.5625%	0.4630%	0.1953%	0.1000%
4	6.2500%	0.3906%	0.0772%	0.0244%	0.0100%
5	3.1250%	0.0977%	0.0129%	0.0031%	0.0010%
6	1.5625%	0.0244%	0.0021%	0.0004%	0.0001%
7	0.7813%	0.0061%	0.0004%	0.0000%	0.0000%
8	0.3906%	0.0015%	0.0001%	0.0000%	0.0000%
9	0.1953%	0.0004%	0.0000%	0.0000%	0.0000%
10	0.0977%	0.0001%	0.0000%	0.0000%	0.0000%

네트워크 초기 단계에서는 네트워크 무결성을 보장하기 위해 노드 운영자를 신중하게 검증합니다. 두 번째 단계에서는 동일한 검증 모델의 여러 인스턴스가 각 검증 요청을 처리하는 계획된 중복 방식을 통해 네트워크가 분산화되기 시작합니다. 이러한 중복은 검증 비용을 증가시키지만, 악의적이거나 태만한 운영자를 효과적으로 식별할 수 있게 합니다. 네트워크가 안정화되어 성숙함에 따라 검증 요청은 노드 전체에 무작위로 분산되어 담합이 점점 더 어렵고 비용이 많이 들게 됩니다.

네트워크의 샤딩(분할) 메커니즘은 또 다른 보안 계층을 제공합니다. 노드 간의 응답 패턴과 유사성 지표를 분석함으로써 시스템은 잠재적인 담합을 식별할 수 있습니다. 악의적인 행위자는 결과를 조작하기 위해 네트워크의 스테이킹된 가치의 상당 부분을 통제해야 하며, 이 시점에서 그들의 경제적 유인은 오히려 정직한 운영과 일치하게 됩니다.

노드 운영자는 가장 낮은 비용으로 정확한 답변을 제공함으로써 성공을 거둡니다. 특화된 모델이 특정 검증 작업에서 더 큰 모델과 비슷한 성능을 달성하면, 합리적인 최적화 기회가 창출됩니다. 이러한 기회는 특정 도메인에 최적화된 효율적인 작업별 모델 개발을 촉진하여, 더 높은 정확도, 낮은 비용, 짧은 지연 시간으로 전체 생태계에 이익을 줍니다.

네트워크의 경제 모델은 여러 긍정적 순환 고리를 통해 이러한 긍정적인 역학을 강화합니다. 네트워크 사용량이 증가함에 따라 수수료 수입이 늘어나고, 이는 더 나은 검증 보상을 가능하게 하여 더 많은 노드 운영자를 유치하고 정확도, 비용, 지연 시간의 개선을 촉진합니다. 이러한 성장은 네트워크 보안을 자연스럽게 강화합니다. 스테이킹 요구 사항은 네트워크 가치와 함께 증가하고, 모델 다양성은 전문화와 관점의 근본적인 차이를 통해 확장되며, 축적된 검증 이력은 점점 더 정교한 이상 징후 탐지를 가능하게 합니다. 이러한 복합적인 효과는 정직한 검증과 지속적인 혁신이 주된 전략으로 부상하고, 악의적인 조작이 경제적으로 불합리하고 기술적으로 불가능하게 되는 강력한 게임 이론적 균형을 만들어냅니다.

4. 프라이버시

앞서 설명한 보안 기반 위에, 미라 네트워크는 프라이버시 보호를 핵심 설계 원칙으로 삼고 있습니다. 프라이버시 보호는 네트워크의 콘텐츠 변환 방식에서부터 시작됩니다. 복잡한 콘텐츠는 개체-주장 쌍으로 분해되어 노드 전체에 무작위로 분산됩니다. 이를 통해 단일 노드 운영자가 전체 콘텐츠를 재구성할 수 없게 되어 검증의 무결성을 유지하면서도 사용자의 프라이버시를 보호합니다.

프라이버시 모델은 여러 보호 계층을 통해 강화됩니다. 노드의 검증 응답은 합의가 이루어질 때까지 비공개로 유지되어 검증 과정 중 정보 유출을 방지합니다. 합의가 이루어지면 네트워크는 필요한 검증 세부 정보만 포함하는 인증서를 생성하여 데이터 최소화를 통해 프라이버시를 더욱 강화합니다.

네트워크 발전 초기 단계에서는 중앙화된 변환 소프트웨어가 자연스러운 프라이버시 경계를 제공합니다. 네트워크 로드맵에는 암호화 프로토콜과 보안 연산 기술을 통해 강력한 프라이버시 보장을 유지하면서 이 구성 요소의 점진적인 탈중앙화가 포함되어 있습니다.

5. 네트워크 발전 방향

미라 네트워크의 발전은 AI 시스템의 작동 방식을 근본적으로 재구성할 종합적인 AI 검증 및 생성 플랫폼을 향한 자연스러운 진화 과정을 따릅니다. 우리의 비전은 단순한 검증을 넘어, 검증이 생성 과정에 내재된 새로운 유형의 기반 모델 창출로 이어지는 것입니다. 이는 AI가 혁신적인 잠재력을 실현하기 위해 필요한 핵심적인 돌파구입니다.

초기에는 의료, 법률, 금융과 같이 사실적 정확성이 중요하고 편향 위험이 최소화된 분야에 집중하다가, 점진적으로 코드, 구조화된 데이터, 멀티미디어 콘텐츠를 포함한 더 복잡한 콘텐츠 유형을 처리할 수 있도록 확장됩니다. 네트워크는 데이터 가용성 계층과 보완 기술을 통해 개인 데이터와 추가적인 맥락에 대한 검증 기능을 확장하여 기본 네트워크를 복잡하게 만들지 않고도 안전하고 효율적인 검증을 가능하게 합니다. 이러한 각 확장은 단순히 더 넓은 범위를 포괄하는 것뿐만 아니라, 더욱 정교하고 신뢰할 수 있는 AI 시스템을 향한 도약입니다.

검증 기능은 단순한 유효성 확인에서 유효하지 않은 콘텐츠의 종합적인 재구성을 거쳐, 궁극적으로는 검증된 결과물의 직접 생성으로 발전합니다. 이러한 발전은 생성 속도와 정확성 사이의 전통적인 상충 관계를 해소하고, 엄격한 검증 기준을 유지하면서도 실시간에 가까운 성능을 달성합니다.

직접적인 검증을 넘어, 블록체인에 경제적으로 보장된 사실의 축적은 강력한 파생 응용 프로그램을 가능하게 합니다. 이 검증된 지식 기반은 네트워크의 보안 보장을 상속받는 결정론적 사실 확인 시스템과 오라클 서비스를 지원할 수 있습니다. 더 근본적으로, 진실 검증을 위한 경제적 유인을 창출함으로써 네트워크는 원시 데이터를 가치 기반 사실로 변환하는 새로운 모델을 구축합니다. 이는 신뢰할 수 있는 AI 시스템의 핵심 구성 요소입니다.

기술적 역량과 경제적 유인 모두의 지속적인 발전을 통해, 네트워크는 전례 없는 신뢰성을 제공하는 새로운 세대의 AI 애플리케이션을 실현할 것입니다. 이는 AI 시스템의 단순한 점진적 개선을 넘어, 인간의 감독 없이도 오류 없는 작동이 가능한 새로운 패러다임을 구축하여 AI가 마침내 자율적으로 작동할 수 있도록 합니다.

6. 결론

오늘날 AI 시스템은 근본적인 난제에 직면해 있습니다. 창의적이고 그럴듯한 결과물을 생성하는 데 탁월하지만, 오류 없는 결과물을 일관되게 제공하지 못하여 인간의 감독이 필수적입니다. 미라 네트워크는 콘텐츠 변환과 암호경제학적 유인으로 가능해진 분산 합의라는 혁신적인 조합을 통해 이 문제를 해결합니다. 이는 조작을 기술적으로나 경제적으로 실행 불가능하게 만듭니다. 임의의 퍼즐을 푸는 기존의 작업 증명(PoW) 방식과는 달리, 미라 네트워크는 정직한 운영을 보장하기 위해 스테이킹된 가치로 뒷받침되는 의미 있는 추론 연산을 요구합니다.

검증을 넘어, 우리의 비전은 검증을 생성 과정에 직접 통합하는 합성 기반 모델을 구축하는 것입니다. 이 간소화된 접근 방식은 생성과 검증 사이의 경계를 허물고 오류 없는 결과물을 제공합니다. 검증을 인센티브가 부여된 운영자들의 탈중앙화된 네트워크 전체에 분산시킴으로써, 우리는 중앙 집중식 통제에 본질적으로 저항하는 인프라를 구축합니다. 이는 패러다임의 전환을 의미합니다. AI 시스템이 인간의 감독 없이도 작동할 수 있도록 함으로써, 우리는 진정한 인공지능의 토대를 마련합니다. 이는 사회 전반에 걸쳐 AI의 혁신적인 잠재력을 실현하기 위한 중요한 발걸음입니다.

레퍼런스:

[1] E. Bender, T. Gebru, A. McMillan-Major, and S. Shmitchell, “On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?”, https://dl.acm.org/doi/pdf/10.1145/3442188.3445922, 2021

[2] M. Chelli, J. Descamps, V. Lavoué, C. Trojani, M. Azar, M. Deckert, J. Raynier, G. Clowez, P. Boileau, C. Ruetsch-Chelli, “Hallucination Rates and Reference Accuracy of ChatGPT and Bard for Systematic Reviews: Comparative Analysis”, https://www.jmir.org/2024/1/e53164/?t, 2024

[3] Z. Gekhman, G. Yona, R. Aharoni, M. Eyal, A. Feder, R. Reichart, J. Herzig, “Does Fine-Tuning LLMs on New Knowledge Encourage Hallucinations?”, https://arxiv.org/pdf/2405.05904, 2024

[4] N. Naik, “Probabilistic Consensus through Ensemble Validation: A Framework for LLM Reliability”, https://mira.network/research/ensemble-validation.pdf, 2024

[5] S. King, S. Nadal, “PPCoin: Peer-to-Peer Crypto-Currency with Proof-of-Stake”, https://www.peercoin.net/read/papers/peercoin-paper.pdf, 2012

[6] S. Nakamoto, “Bitcoin: A Peer-to-Peer Electronic Cash System”, https://bitcoin.org/bitcoin.pdf, 2009

[7] X. Zuwei, S. Jain, M. Kankanhalli, “Hallucination is Inevitable: An Innate Limitation of Large Language Models”, https://arxiv.org/abs/2401.11817, 2024