

실패 뒤에는 무엇을 고쳐야 하는가

수학적 추론에서 수정 범위의 비식별성과 판별 실험 기반 에스컬레이션

유아독존 · ChatGPT

Working Paper · 본문 중심 공개판

본문 1-12장 + 참고문헌

이 공개판은 내부 검증표 · 편집 메모 · 출간 설계 부록을 덜어내고, 독자가 읽을 본문과 참고문헌을 중심으로 다시 편집했습니다.

서울드림의원 · 원장의 머릿속 / 개인 연구 프로젝트

1 서론

1.1 답보다 어려운 다음 행동

문제를 푼다는 통상적 그림은 선형적이다. 문제를 표현하고, 후보 해를 탐색하고, 추론을 전개하고, 정답이 나 증명을 산출한다. 그러나 실제 수학 연구의 시간 대부분은 이 직선 위가 아니라 실패 이후의 분기점에서 소비 된다. 계산이 맞지 않을 때 상수를 고쳐야 하는가, 보조정리를 바꾸어야 하는가, 좌표계나 언어를 바꾸어야 하는가, 명제가 거짓인지 의심해야 하는가, 아니면 현재 자원 아래에서 중단해야 하는가. 이 선택이 틀리면 연구자는 맞는 아이디어를 너무 일찍 버리거나, 이미 막힌 경로를 더 정교하게 반복한다. 수학철학은 증명과 이해가 같지 않다는 점을 오래 지적해 왔다. 서스턴은 수학의 진보가 형식적 정리의 축적만으로 포착되지 않으며, 개념과 소통, 새로운 시야가 핵심이라고 주장했다. 과학적 이해 연구는 이해를 정보를 보유하는 상태가 아니라 표현을 조작하고 반사실적 질문에 답하는 능력으로 분석한다. 23 그림에도 ‘실패 이후 무엇을 얼마나 바꿀 것인가’는 이해의 정의와 별개의 의사결정 문제다. 한 오류를 찾는 것과 적절한 수정의 크기를 고르는 것은 다르다. 본 논문은 이 간극을 수정 범위 선택 문제로 명명한다. 수정 범위란 변경의 물리적 양이 아니라, 현재 추론 체계에서 어떤 구성요소를 유지하고 어떤 구성요소를 대체하는가를 뜻한다. 국소 전개만 고치는 수정, 증명 구조를 바꾸는 수정, 표현을 바꾸는 수정, 목표·수용 기준을 다시 쓰는 수정은 비용과 정보 요구량이 다르다.

1.2 연구 질문

본 연구는 다음 다섯 질문에 답한다. 1. RQ1. 실패 간차나 오류 메시지만으로 적절한 수정 범위를 식별할 수 있는가. 2. RQ2. 반복 실패는 언제 상위 범위로의 전환을 정당화하며, 언제 단지 탐색 부족을 뜻하는가. 3. RQ3. 신념 수정, 진단, 제어, 통계, 인과 추론, 수학철학, AI 자기수정은 이 문제의 어느 부분을 이미 해결했는가. 4. RQ4. 현재 수학 추론·정형 증명·에이전트 벤치마크는 실패 이후의 경로 선택을 어디까지 측정하는가. 5. RQ5. 수정 범위 선택을 독립된 능력으로 평가하려면 어떤 반사실적 과제와 지표가 필요한가.

1.3 핵심 주장과 기여

논문의 주장은 강도에 따라 세 층으로 구분한다. 지위 주장 현재 근거 확립된 배경 오류 검출, 진단, 최소 변화, 반복 수리, 모 원전 및 최신 벤치마크 비교 2

3 본 연구의 최신 문헌 검색 기준일은 2026년 9월 10일이다. 2026년 9월에 공개된 나비에 - 스톡스 관련 결과는 출판 · 공동체 검증이 완료

되지 않은 상태이므로 확정 사실과 분리해 기술하였다.

지위 주장 현재 근거 델 전환, 능동 실험은 각각 성숙한 선행연구가 있다. 본 논문의 이론 명제 현재 관찰에서 구별되지 않는 수정 가설이 존재하면 간차만으로 수정 범위를 식별할 수 없다. 형식화와 구성적 반례 검증할 독창성 후보 관찰적으로 별칭화된 실패쌍에서 수정 범위 선택과 보존 비용을 공동 평가하는 수학 벤치마크가 필요하다. 범위 지정 문헌 검토와 프로토콜 아직 미검증 제안 지표가 최종 정확도보다 실제 장기 연구 성과를 더 잘 예측한다. 향후 파일럿 · 종단 실험 필요 표 1 주장의 증거 수준 기여는 네 가지다. 첫째, 실패 검출 · 국소화 · 수정 선택 · 검증의 분리를 통해 자기수정 담론의 숨은 과업 혼동을 제거한다. 둘째, 동일 간차가 서로 다른 수정 원인과 양립하는 최소 반례를 제시한다. 셋째, 반복 실패를 곧바로 ‘더 깊이 바꾸라’는 명령으로 해석하지 않고 경쟁 가설을 가르는 판별 실험으로 연결한다. 넷째, 이 이론을 최종 정답률과 별도로 측정할 수 있는 RevisionScope - Math 설계를 제안한다.

1.4 독자가 얻는 것

전문 독자는 새로운 보편 이론의 과장된 선언보다, 어디까지가 선행연구이고 어디부터가 검증 가능한 제안 인지 확인할 수 있다. 수학 · 과학 교양 독자는 나비에 - 스톡스 방정식과 대표 난제 열 개를 통해 ‘풀이’가 현실과 연결되는 여러 방식을 이해할 수 있다. AI 연구자는 에이전트가 실패를 반복하는 현상을 최종 성공률의 부수적 오류가

아니라 별도의 진단·전환 능력으로 측정하는 실험 설계를 얻는다. 이 이중 독자 구조는 단행본의 폭과 논문의 엄격성을 분리하지 않고, 본문 핵심 명제와 확장 사례집이라는 층위로 결합한다.

2 연구 범위와 방법

2.1 범위 지정 문헌 검토

본 연구는 체계적 메타분석이 아니라 개념 충돌과 벤치마크 공백을 찾기 위한 범위 지정 검토다. 2026년 9월 10일 까지 출판사·학회·저널·공식 프로젝트·arXiv 원문을 우선 확인했다. 검색 축은 ‘belief revision/minimal change’, ‘model-based diagnosis/fault localization’, ‘adaptive control/model switching’, ‘Bayesian workflow/model criticism’, ‘causal identifiability/intervention’, ‘mathematical understanding/proof’, ‘LLM

self-correction/repair’,

‘formal

theorem

proving

benchmark’,

‘failure recovery/replanning/stagnation/backtracking/partial progress’였다. 기술적 주장은 가능한 한 원문과 공식 문제 설명을 사용했고, 매우 최근 사건의 운영 세부는 원문·공개 저장소와 신뢰할 수 있는 보도를 구분했다.

포함 기준은 실패 이후의 지식·모델·전략·증명 수정, 수정 비용, 진단 가설, 능동 판별, 과정 평가 가운데 하나를 명시적으로 다루는 연구다. 단순 최종 정확도 보고, 실패가 있었음을 언급하지만 수정 행동을 관찰하지 않는 연구, 수학적 추론과 연결할 수 없는 순수 응용 성능 보고는 비교표의 핵심 집합에서 제외했다. 최신 연구의 빠른 변화 때문에 ‘어떤 연구도 없다’는 절대적 결론은 내리지 않는다. 여기서 공백은 검토 집합 안에서 확인된 결합적 측정 공백이다.

2.2 개념 개발 절차

개념은 다음 네 번의 공격적 축소를 거쳤다. 단계 초기 후보 선행연구 공격 결과 1 이해 프로파일 $U=(P,E,I,G,C)$ 예측·설명·개입·일반화·압축은 과학적 이해와 설명 문헌에 널리 존재 단행본 지도에는 유용하나 논문 핵심으로 축소 2 교정 가능한 지식 순환 베이지안 워크플로, 제어, 조직 학습, 자기수정과 중복 용어와 범위를 좁힘 3 보편적 수정 계층 매개변수 조정, 모델 전환, 표현 변경, 목표 변경이 각각 독립된 통을 가짐 보편 계층 주장 폐기 4 실패 패턴이 수정 깊이를 직접 지시 같은 간차가 서로 다른 원인에서 발생 가능 비식별성 명제와 판별 실험으로 전환 표 2 아이디어의 축소와 잔존 핵심 이 절차는 부정적 결과를 실패로 숨기지 않는다. ‘이해를 다섯 요소로 정의한다’, ‘실패가 누적되면 한 단계씩 깊게 수정한다’ 같은 문장은 설명력이 높아 보이지만, 기존 이론과의 충돌 및 반례에 취약하다. 살아남은 주장은 더 작다. 실패 증거는 수정 후보를 만들지만, 후보 사이의 선택은 그들을 구별하도록 설계된 추가 관찰에 의존한다. 작은 주장이기 때문에 형식화, 반례, 벤치마크라는 세 종류의 검증을 동시에 요구할 수 있다.

2.3 정보 엔트로피를 낮춘다는 것

이 문서가 목표로 하는 ‘정보 엔트로피가 낮은 글’은 새년 엔트로피를 실제로 계산했다는 뜻이 아니다. 편집 원칙상 동일 주장 반복을 줄이고, 각 절이 하나의 결정 질문에 답하며, 주장마다 증거 지위를 붙이고, 독자가 다음 연구 행동을 선택하는 데 필요한 정보를 밀도 있게 제공한다는 뜻이다. 따라서 사례의 양은 많지만 기능은 중복되지 않는다. 나비에-스토크스는 모델·조건·특이점의 경계를 보여 주고, 열 개 난제는 ‘해결’의 여러 형식을 비교 하며, 벤치마크 절은 이론을 측정 가능한 설계로 바꾼다.

2.4 윤리와 AI 사용 고지

본 초고의 문헌 탐색, 비교표 구성, 개요 설계, 문장 편집에 생성형 AI를 사용하였다. 그러나 인용의 정확성, 최신 사건의 사실 상태, 수식과 반례의 타당성, 투고 규정 준수에 대한 최종 책임은 저자에게 있다. 투고 전에는 모든 인용 원문을 사람이 재확인하고, 대상 학술지의 AI 사용 · 저자 자격 · 데이터 공개 정책에 맞추어 고지문을 수정해야 한다. AI는 저자가 될 수 없으며, 미확인된 자동 생성 설명을 증거로 취급해서는 안 된다.

3 선행연구와 개념적 계보

3.1 증명과 이해

형식적 증명은 명제가 공리와 추론 규칙에서 따라 나온다는 인증이다. 그러나 이해는 왜 그 명제가 참인지, 어떤 가정이 일을 하는지, 조건을 바꾸면 무엇이 달라지는지, 같은 구조가 어디에 재사용되는지를 파악하는 능력을 포함한다. 라카토슈는 추측, 반례, 보조정리 수정이 얽힌 수학의 실제 성장을 ‘증명과 반박’의 대화로 묘사했다. 서스톤은 엄밀한 증명이 수학 공동체가 공유하는 이해의 한 부분이지 전부가 아니라고 보았다. 이 계보는 실패를 제거해야 할 잡음이 아니라 개념 형성의 자료로 본다는 점에서 본 연구의 출발점이다. 4 다만 이해를 예측, 설명, 개입, 일반화, 압축의 다섯 능력으로 묶는 ‘이해 프로파일’은 본 연구의 최종 이론이 아니다. 이러한 요소는 우드워드의 개입주의적 설명, 윌켄펠드의 표현 조작 가능성, 드 레그트의 맥락적 지성이 이미 상당 부분 설명한다. 본 논문은 이를 경쟁 이론으로 제시하지 않고 사례를 배열하는 읽기 지도에 한정한다. 무엇을 이해했는지보다, 실패 뒤에 무엇을 바꿀지를 측정하는 쪽이 더 좁고 검증 가능하다. 56

3.2 신념 수정과 최소 변화

AGM 신념 수정 이론은 새로운 정보가 기존 신념집합과 충돌할 때 확장, 축소, 수정을 어떻게 합리화할지 공리적으로 다룬다. 핵심 직관은 일관성을 회복하되 가능한 한 기존 신념을 보존하는 최소 변화다. 이 전통은 ‘모두 지우고 다시 시작’이 합리적 기본값이 아님을 명료하게 한다. 그러나 무엇이 원인이었는지에 관한 진단 정보와, 다음 실험을 어떻게 설계할지는 별도의 층이다. 같은 모순을 해소하는 최소 수정이 여러 개 존재할 수 있고, 어느 수정이 세계의 구조에 맞는지는 AGM 공리만으로 결정되지 않는다. 7 4 William P. Thurston, “On Proof and Progress in Mathematics,” *Bulletin of the American Mathematical Society* 30 (1994): 161 – 177, <https://arxiv.org/abs/math/9404236>. 그는 수학의 산출물을 증명문에만 환원할 때 실제 진보의 중요한 형식이 사라진다고 논한다. 5 Daniel A. Wilkenfeld, “Understanding as Representation Manipulability,” *Synthese* 190 (2013): 997 – 1016, <https://doi.org/10.1007/s11229-011-0055-x>; Henk W. de Regt, *Understanding Scientific Understanding*, Oxford University Press, 2017, <https://global.oup.com/academic/product/understanding-scientific-understanding-9780190652913>. 6 Imre Lakatos, *Proofs and Refutations: The Logic of Mathematical Discovery*, Cambridge University Press, 1976, <https://books.google.com/books?id=HJ2NCgAAQBAJ>. 라카토슈의 역사적 대화는 반례가 명제 · 정의 · 보조정리를 재조정하게 만드는 과정을 전면화한다. 7 James Woodward, *Making Things Happen: A Theory of Causal Explanation*, Oxford University Press, 2003, <https://books.google.com/books?id=LrAbrrj5te8C>. 설명을 변화 · 개입에 관한 반사실적 정보와 연결한다. 본 연구는 최소 변화 원칙을 폐기하지 않는다. 대신 수정 선택 손실에 ‘버린 검증 성과의 비용’을 넣어 정량화한다. 동시에 최소 수정이 늘 옳다는 주장도 거부한다. 잘못된 표현이나 거짓 목표를 보존하려는 국소 수리는 장기적으로 더 큰 비용을 만들 수 있다. 따라서 최소성은 진단 후의 비용 기준이지, 진단을 대신하는 규칙이 아니다.

3.3 모델 기반 진단과 결함 국소화

라이터의 모델 기반 진단은 시스템 기술, 정상성 가정, 관찰이 모순될 때 어떤 구성요소가 비정상이라고 가정하면 일관성이 회복되는지를 계산한다. 최근 일반화는 진단 후보를 단순 부품 먹집합에 한정하지 않고 선호관 계가 있는 일반 가설 공간에서 최소 진단을 탐색한다. 이 문헌은 실패 국소화와 후보 생성에 가장 가까운 선행연구다. 89 하지만 수학 연구의 수정 대상은 고정된 부품 목록이 아니다. 문제의 표현, 허용 공리, 보조정리 그래프, 탐색 정책, 성공 기준 자체가 바뀔 수 있다. 또한 증명 검사기가 내놓는 첫 오류 위치는 원인의 위치와 다를 수 있다. 앞 단계의 잘못된 정의가 수십 단계 뒤의 타입 오류로 드러나거나, 계산상 오류 없이도 전체 전략이 막힐 수 있다. 본 논문은 진단 전통을 계승하되, 진단 후보가 정해진 뒤에도 어떤 추가 시험으로 후보를 가를지와 어느 수정 연산자를 실행할지를

별도 문제로 둔다.

3.4 단일고리 · 이중고리 학습과 목표 수정

아지리스와 쉰의 단일고리 · 이중고리 학습은 주어진 목표와 규칙 안에서 오류를 고치는 것과, 목표 · 규범 · 기본 가정을 되묻는 것을 구분한다. 이는 매개변수와 목표를 동일한 층에서 다루지 않아야 한다는 중요한 통찰을 준다. 그러나 목표 수정은 더 ‘깊은’ 기술적 수정이 아니다. 성공 기준 자체를 바꾸기 때문에 사실 판단과 가치 판단이 결합된다. 10 따라서 본 논문의 핵심 벤치마크는 목표를 바꾸는 결정을 자동으로 상위 정답으로 취급하지 않는다. 목표가 모호하거나 모순된 과제에서는 ‘명료화 요청’이나 ‘수용 기준 재협상’을 별도 행동으로 평가한다. 순수 수학 명제가 거짓인 경우에도 임의로 쉬운 명제로 바꾸면 성공이 아니다. 정당한 목표 수정은 원래 의도와의 보존 관계를 명시해야 한다.

3.5 적응 제어와 다중 모델 전환

적응 제어는 시스템의 불확실한 동역학을 추정하면서 제어기를 조정한다. 다중 모델 · 스위칭 접근은 하나의 8 Henk W. de Regt et al., “Scientific Understanding across Disciplines: In Varietate Concordia?,” arXiv:2608.14378, 2026, <https://arxiv.org/abs/2608.14378>. 학문별 이해 기준의 차이와 공통점을 비교한 최신 연구이므로 보편 단일척도에 대한 경계 근거로 사용한다. 9 Carlos E. Alchourr n, Peter G rdenfors, and David Makinson, “On the Logic of Theory Change: Partial Meet Contraction and ó ä Revision Functions,” Journal of Symbolic Logic 50, no. 2 (1985): 510 – 530, <https://doi.org/10.2307/2274239>. 10 Raymond Reiter, “A Theory of Diagnosis from First Principles,” Artificial Intelligence 32, no. 1 (1987): 57 – 95, [https://doi.org/10.1016/0004-3702\(87\)90062-2](https://doi.org/10.1016/0004-3702(87)90062-2).

모델을 끝없이 미세 조정하기보다 여러 후보 모델의 식별 오차를 비교하여 전환한다. 이는 ‘국소 보정이 계속 실패하면 모델 가족을 바꾼다’는 직관의 강력한 선행 사례다. 11 그럼에도 제어의 모델 전환 규칙을 수학적 추론에 그대로 이식할 수는 없다. 제어에서는 출력 오차와 안정성 이 비교적 명시적이지만, 수학 연구에서는 어떤 중간 산출물이 진전인지, 명제의 변형이 허용되는지, 표현 변경 이 같은 문제를 보존하는지부터 논쟁적이다. 전환의 존재는 확립되어 있으나 전환 대상의 의미론은 도메인별로 설계되어야 한다.

3.6 베이지안 워크플로와 모델 비판

베이지안 데이터 분석의 실제 흐름은 한 번의 사후분포 계산으로 끝나지 않는다. 모델 구축, 계산 점검, 사후 예측 점검, 검증, 비교, 문제 해결을 반복한다. 잔차 패턴은 모델 부적합을 드러낼 수 있지만 어느 구성요소가 틀렸는지를 자동으로 말하지 않는다. 통계적 적합도 지표가 올바른 잠재 구성과 잘못 지정된 구성을 구별하지 못하는 사례는 ‘좋은 맞춤’과 ‘올바른 구조’의 차이를 보여 준다. 12 본 논문에서 잔차는 넓은 뜻으로 사용한다. 예측과 관찰의 차이뿐 아니라 반례, 증명 검사기 오류, 보조정리의 미충족 조건, 검색 정체, 설명이 요구한 반사실적 질문에 답하지 못하는 현상을 포함한다. 이 확장은 편리하지 만 위험하다. 서로 다른 잔차를 같은 수치로 압축하면 원인 정보가 사라진다. 따라서 벤치마크는 잔차 유형을 보존하면서도, 표면이 같은 사례쌍을 의도적으로 만든다.

3.7 인과 식별과 판별 실험

관찰 자료만으로 여러 인과 그래프가 같은 조건부 독립관계를 만들 수 있다. 개입 자료는 관찰적 마르코프 동치류를 더 잘게 나누어 식별성을 높인다. 이 결과가 본 연구에 주는 교훈은 직접적이다. 현재 실패 이력 아래에서 여러 수정 가설이 같은 예측을 한다면, 같은 종류의 관찰을 더 모으는 것만으로는 구별할 수 없다. 후보가 다른 결과를 예측하도록 입력 · 보조정리 · 표현 · 검사 조건을 바꿔야 한다. 13 베이지안 최적 실험 설계는 실험이 가져올 기대 정보 이득이나 기대 효용을 최대화한다. 본 논문은 이 틀을 수정 가설 판별에 적용한다. 다만 수학적 시험의 비용은 계산량만이 아니다. 새로운 형식화, 전문가 검토, 기존 부분 증명의 폐기, 문제 의미의 변형까지 포함한다. 따라서 정보 이득에서 수정 · 실험 비용을 뺀 비용 민감 기준이 필 요하다. 14 11 Alban Grastien, Patrik Haslum, and Sylvie Thi baux, “A More General Theory of Diagnosis from First Principles,” é arXiv:2309.16180, 2023, <https://arxiv.org/abs/2309.16180>. 12 Chris Argyris and Donald A. Sch n, Organizational Learning: A Theory of Action Perspective, Addison - Wesley, 1978, ö <https://books.google.com/books?id=2aYOAQAAMAAJ>. 13 Kumpati S. Narendran and Jeyendran

Balakrishnan, “Adaptive Control Using Multiple Models,” IEEE Transactions on Automatic Control 42, no. 2 (1997): 171 – 187, <https://doi.org/10.1109/9.554398>. 14 Andrew Gelman et al., “BayesianWorkflow,” arXiv:2011.01808, 2020, <https://arxiv.org/abs/2011.01808>. 이 논문은 추론뿐 아니라 반복적 모델 구축 · 점검 · 검증 · 문제 해결 · 비교를 하나의 얹힌 워크플로로 정리한다.

3.8 AI 자기수정과 도구 피드백

Self - Refine과 Reflexion은 모델이 자신의 출력에 피드백을 주고 반복 개선하거나, 실패 경험을 언어적 기억으로 저장해 다음 시도에 활용하는 방법을 제시했다. CRITIC은 검색기 · 코드 실행기 같은 외부 도구의 피드백을 이용해 수정한다. 반대로 내재적 자기수정만으로는 추론 성능이 악화될 수 있다는 결과도 있다. 이 상반된 결과의 공통점은 피드백의 내용과 출처가 중요하다는 것이다. 1516 그러나 대부분은 ‘수정한다’는 연산을 단일 프롬프트나 재생성으로 구현한다. 무엇을 보존하고 무엇을 바꾸었는지, 더 큰 수정이 정말 필요했는지, 같은 실패 경로를 되풀이했는지, 멈추는 편이 합리적이었는지를 분리해 채점하지 않는다. 본 연구의 측정 단위는 수정된 최종 답이 아니라 실패 직후의 선택과 그 선택이 만든 정보 · 비용 · 보존 효과다.

3.9 정형 증명과 과정 벤치마크

MATH는 단계별 풀이가 있는 경시 수준 문제로 수학 추론 평가를 크게 확장했고, miniF2F · LeanDojo · PutnamBench · miniCTX는 형식 검증과 문맥 의존 증명으로 범위를 넓혔다. ProcessBench는 최종 정답 대신 최초 오류 단계를 찾게 하며, AgentBoard는 부분 진전률을 도입했다. 최신 FormalProofBench는 대학원 수준 증명에서 정확도뿐 아니라 도구 사용, 실패 유형, 비용, 지연을 분석한다. 1718 더 강한 근접 선행연구도 있다. OProver는 실패한 Lean 증명, 컴파일러 피드백, 수리 궤적을 대규모로 저장해 훈련에 사용한다. 2026년의 적응형 증명 검색은 현재 최선의 부분 증명을 보존하고 정체를 감지해 재표집함으로써 효율을 높인다. AgentDebugX는 검출 - 귀인 - 복구 - 재실행의 페루프를 제공하고, ToolMaze는 도구 이상 아래의 동적 재계획과 무의미한 반복을 측정한다. 이 연구들은 본 제안이 ‘누구도 실패 복구를 보지 않았다’고 주장할 수 없게 한다. 1920 15 Alain Hauser and Peter B hlmann, “Characterization and Greedy Learning of Interventional Markov Equivalence Classes of Directed Acyclic Graphs,” Journal of Machine Learning Research 13 (2012): 2409 – 2464, <https://jmlr.org/papers/v13/hauser12a.html>; arXiv version <https://arxiv.org/abs/1104.2808>. 16 Kathryn Chaloner and Isabella Verdinelli, “Bayesian Experimental Design: A Review,” Statistical Science 10, no. 3 (1995): 273 – 304, <https://doi.org/10.1214/ss/1177009939>. 17 Aman Madaan et al., “Self - Refine: Iterative Refinement with Self - Feedback,” arXiv:2303.17651, 2023, <https://arxiv.org/abs/2303.17651>; Noah Shinn et al., “Reflexion: Language Agents with Verbal Reinforcement Learning,” arXiv:2303.11366, 2023, <https://arxiv.org/abs/2303.11366>. 18 Zhibin Gou et al., “CRITIC: Large Language Models Can Self - Correct with Tool - Interactive Critiquing,” ICLR 2024, <https://arxiv.org/abs/2305.11738>; Jie Huang et al., “Large Language Models Cannot Self - Correct Reasoning Yet,” ICLR 2024, <https://arxiv.org/abs/2310.01798>. 19 Dan Hendrycks et al., “Measuring Mathematical Problem Solving with the MATH Dataset,” NeurIPS 2021 Datasets and Benchmarks, <https://arxiv.org/abs/2103.03874>; Kunhao Zheng, Jesse Michael Han, and Stanislas Polu, “miniF2F,” ICLR 2022, <https://arxiv.org/abs/2109.00110>; Kaiyu Yang et al., “LeanDojo,” NeurIPS 2023, <https://arxiv.org/abs/2306.15626>. 20 Chujie Zheng et al., “ProcessBench: Identifying Process Errors in Mathematical Reasoning,” ACL 2025, <https://arxiv.org/abs/2412.06559>; Chang Ma et al., “AgentBoard,” 2024, <https://arxiv.org/abs/2401.13178>; Nikil Ravi et al., “FormalProofBench,” 2026, <https://arxiv.org/abs/2603.26996>.

남는 틈은 더 구체적이다. 검토한 벤치마크들은 오류 위치, 성공 여부, 부분 진전, 수리 후 성공, 재계획, 비용, 지연 가운데 여러 항목을 측정하지만, 동일한 초기 실패 표면을 공유하면서 서로 다른 수정 범위를 요구하는 반사실적 사례쌍을 핵심 단위로 삼지는 않는다. 또한 정답으로 인정된 부분 산출물의 보존, 불필요한 전면 재시작, 너무 늦은 전략 전환, 적당한 중단을 하나의 평가 틀에서 공동 측정하지 않는다. 이 결론은 2026년 9월 10일 의 범위 지정 검토에 한정된 ‘결합 공백’이며 절대적 최초성 선언이 아니다.

3.10 선행연구 비교 요약

전통·벤치마크 이미 해결한 질문 본 연구와의 경계 판정 AGM 신념 수정 일관성·최소 변화의 공리 원인 판별과 시험 선택은 약함 흡수 모델 기반 진단 관찰과 양립하는 최소 결합 후보 가변 표현·목표와 행동 비용 추가 필요 핵심 선행 이중고리 학습 규칙 내부 수정과 규범 수정 구분 정량적 식별·벤치마크 아님 범위 분리 다중 모델 적응 제어 오차 기반 모델 전환 수학적 의미 보존·증명 산출물 보존 미포함 유사 구조 베이지안 워크플로 반복 모델 비판·점점 잔차에서 수정 범위로의 식별은 비 자동 유사 구조 인과 개입·실험 설계 동치 가설을 가르는 개입·정보 이득 수학 추론의 행동·비용으로 구체화 필요 형식 도구 ProcessBench 최초 오류 단계 국소화 복구 행동과 수정 범위는 채점하지 않음 보완 AgentBoard 부분 진전률 진전 보존과 불필요한 리셋 원인은 미분리 근접 OProver·Baldur 컴파일러 피드백 기반 증명 수리 경로 선택의 정당성과 반사실적 별 칭쌈 미평가 강한 근접 적응형 증명 검색 정제 감지·최선 후보 보존·재표집 일반 수정 범위 주식과 목표 명료화 미포함 가장 강한 근접 AgentDebugX·ToolMaze 근본 원인 귀인·복구·재계획 수학 의미론과 증명 구조의 범위 선택이 아님 교차 도메인 근접 RevisionScope - Math 별칭 실패쌍의 수정 범위·비용·보존·중단 데이터와 실증 결과는 아직 없음 독창성 후보

표 3 선행연구 충돌과 잔여 공백

4 연구 프로그램의 전체 구조와 현재 위치

4.1 아홉 노드의 정본 구조

장기 프로젝트 MATHEMATICAL FRONTIERS의 정본 흐름은 세계에서 출발해 수정으로 돌아오는 아홉 노드다. 이 순서는 존재론적 실재가 자동으로 기호가 된다는 뜻이 아니라, 연구자가 어디에서 정보 손실과 판단 오류를 만들 수 있는지 추적하기 위한 감사 경로다. 모든 노드는 정보, 관련성, 규모라는 세 필터를 통과한다. 정보는 무엇이 관찰·계산되었는지, 관련성은 그것이 현재 문제와 어떤 관계인지, 규모는 국소 현상이 전체 결론으로 확대될 수 있는지를 묻는다. 노드 핵심 질문 주요 산출물 대표 실패 WORLD 세계 무엇이 실제 대상·제약인가 현상·공리·데이터·규범 대상과 모델의 혼동 REPRESENTATION 표현 무엇을 어떤 언어로 보존하는가 좌표·기호·형식화·모델 관련 구조의 소실 PROBLEM 문제 정확히 무엇을 묻는가 명제·목표·수용 기준 모호·거짓·과대 목표 SEARCH 탐색 후보 경로를 어떻게 배분하는가 전략 포트폴리오·예산 한 경로 고착·무작위 재시도 INFERENCE 추론 어떤 단계가 무엇에서 따르는가 증명·계산·시뮬레이션 논리 비약·도구 오류 EXPLANATION 설명 왜 작동하며 무엇에 의존하는가 메커니즘·압축·직관 검증과 이해의 혼동 INTERVENTION 개입 조건을 바꾸면 예측이 달라지는가 반사실적 시험·실험 비판별적 추가 관찰 FAILURE 실패 어디서 어떤 방식으로 어긋났는가 잔차·반례·정제 기록 증상과 원인의 혼동 REVISION 수정 무엇을 보존하고 무엇을 바꿀 것인가 국소 수리·전환·재표현 과소 수정·과잉 리셋 표 4 프로젝트 정본 노드와 실패 지점 이 흐름은 직선이 아니라 폐루프다. 수정된 표현과 문제는 다시 탐색과 추론을 거쳐 세계·공리·형식 검사에 부딪힌다. 그러므로 최종 산출물은 ‘정답 하나’보다 수정 가능한 지식 상태다. 다만 이 표현도 이미 사이버네틱스, 베이지안 워크플로, 교정적 학습과 닮아 있다. 본 프로젝트의 논문 가능성은 순환 자체가 아니라, 실패에서 수정으로 넘어가는 경계의 비식별성과 그 평가 방법에 집중할 때 생긴다.

4.2 이해 프로파일의 역할과 한계

$U = (P, E, I, G, C)$ 이해 프로파일 U 는 예측 Prediction, 설명 Explanation, 개입 Intervention, 일반화 Generalization, 압축 Compression을 뜻한다. 이는 단행본에서 서로 다른 난제가 어떤 종류의 이해를 늘렸는지 비교하는 데 유용하다. 예컨대 페렐만의 증명은 분류와 구조적 설명을, 힐베르트 열 번째 문제의 부정적 해법은 계산 가능성의 한계를, 나비에-스톡스 특이점 연구는 모델의 적용 조건과 규모 간 상호작용을 드러낸다.

그러나 U 를 보편적 수치 지수로 합산하면 문제가 생긴다. 다섯 요소의 가중치는 목적과 학문에 따라 달라지고, 압축성과 조작 가능성이 서로 충돌할 수 있으며, 한 요소의 높은 성과가 다른 요소의 결핍을 보상한다고 볼 근거가 없다. 따라서 본 논문에서는 U 를 서술적 프로파일로만 유지하고 핵심 정리나 단일 점수로 사용하지 않는다. 이 제한은 약점이 아니라 과잉 이론화를 피하는 경계 조건이다.

4.3 교정 가능한 지식 순환

Represent → Predict → Observe → Detect → Localize → Revise → Retest 교정 가능한 지식 순환은 표현, 예측, 관찰, 실패 검출, 원인 국소화, 수정, 재검사를 잇는다. 이 흐름은 전체 프로젝트의 운영 프로토콜로 남긴다. 다만 명칭만으로는 새 이론이 아니다. 본 연구가 추가하는 구분은 Detect와 Localize 사이, Localize와 Revise 사이에 불확실성이 존재하며, Retest는 단순 확인뿐 아니라 수정 가설을 가르 는 능동 실험이어야 한다는 점이다. 핵심 등식 실패 검출 ≠ 실패 국소화 ≠ 수정 선택 ≠ 수정 검증

4.4 현재 위치와 다음 방향

후보 현재 판정 이유 다음 행동 보편적 수정 계층 폐기 도메인마다 수정 대상과 순서가 다르고 목표 변경은 기술 수정과 성격이 다른 역사적 배경으로만 보존 잔차가 수정 깊이를 직접 식별 폐기 관찰적으로 같은 잔차가 여러 원인과 양립 반례로 명시 반복 실패를 경보로 사용 조건부 유지 지속성은 이상 신호지만 방향을 주지 못함 판별 시험으로 연결 판별 시험 기반 조건부 에스컬레이션 유지 인과 식별 · 실험 설계와 정합 형식화 · 벤치마크 목표 수정 분리 성공 기준 자체를 바꾸는 규범적 결정 후속 모듈 RevisionScope - Math 핵심 개발 결합 측정 공백이 남음 파일럿 데이터 100 - 300개 구축 표 5 중간 정리와 현재 연구 위치 현재 위치는 개념 발산 단계가 아니라 개념 수렴 단계다. 다음 연구의 우선순위는 새 거대 개념을 더 만드는 것이 아니라, (1) 수정 범위 레이블의 운영 정의, (2) 관찰적 별칭 실패쌍의 제작 가능성, (3) 사람 전문가 간 일치도, (4) 기존 에이전트가 어떤 과소 · 과잉 수정을 하는지에 대한 파일럿이다. 이 네 항목이 성립하면 논문은 철학적 주장과 실증 벤치마크를 연결한다. 성립하지 않으면 수정 범위를 단일 정답으로 주석하는 계획을 버리고 선호 집합 또는 부분순서로 바꾸어야 한다.

5 수정 범위의 형식적 틀

5.1 추론 상태의 구성

현실을 직접 다루는 예측 모델과 수학 문제를 푸는 증명 에이전트를 함께 표현하기 위해 추론 상태를 다음처럼 둔다. O 는 대상 또는 공리적 배경, ρ 는 표현, S 는 구조, θ 는 국소 선택, G 는 목표와 수용 기준, Π 는 탐색 정책이다. $H_t = (O, \rho, S, \theta, G, \Pi, V_t)$ (1) V_t 는 시점 t 까지 검증된 부분 산출물의 집합이다. 예측 문제를 좁혀 쓰면 현재 모델은 $\rho = f(S, (x))$ $\hat{y} = \theta(x)$ ρ 가 된다. 수학적 증명에서는 (x) ρ 가 자연어 명제를 Lean 표현으로 옮긴 결과일 수 있고, S 는 보조정리 의존 그래프, θ 는 구체적 상수 · 전술 · 인스턴스, Π 는 어떤 보조정리와 전술을 먼저 탐색할지 정하는 정책이다. V_t 를 명시하는 이유는 전면 재시작이 버리는 지식의 양을 측정하기 위해서다.

5.2 네 과업의 분리

과업 입력 출력 대표 오류 실패 검출 후보 산출물 · 검사 실패 여부와 증거 e 거짓 음성 · 거짓 양성 실패 국소화 e · 궤적 · 의존관계 원인 후보 집합 C 표면 오류를 근본 원인으로 오인 수정 선택 C · 비용 · 목표 · 보존 집합 행동 d 과소 수정 · 과잉 수정 · 지연 수정 검증 수정 후 산출물 · 새 시험 조건부 채택 · 기각 훈련 사례 과적합 · 국소 성공 오인 표 6 실패 이후 네 과업 ProcessBench가 주로 첫째와 둘째 사이를 평가하고, 정형 증명 수리 시스템이 첫째에서 넷째까지의 일부를 실행하더라도, 셋째의 선택이 적절했는지를 독립적으로 채점하지 않을 수 있다. 최종 성공은 잘못된 선택을 많은 예산으로 덮을 수 있고, 실패는 올바른 선택이었지만 예산이 부족했을 수 있다. 그러므로 최종 결과와 수정 결정은 별도 기록이 필요하다.

5.3 수정 연산자

수정은 절대적 깊이가 아니라 현재 표현과 후보 공간에 상대적인 연산자로 정의한다. 연산자 바꾸는 것 보존 기대 수학 예 R 국소 수리 한 단계 · 상수 · 전술 구조와 표현 전부 부호 수정, 누락 가정 추가 R 구조 수정 보조정리 그래프 · 불변량 · 분해 문제와 표현 귀납 불변량 교체, 직접증명을 모순증명으로 R ρ 표현 수정 좌표 · 언어 · 형식화 · 대상 부호화 의미 보존 관계 기하 문제를 PDE로, 자연어를 Lean으로

연산자 바꾸는 것 보존 기대 수학 예 R Π 탐색 정책 수정 경로 배분 · 도구 · 검색 예산 현재 후보와 검증 성과 정제 경로 중단, 다른 모델 호출 R G 목표 명료화 명제 · 허용 조건 · 성공 기준 원래 의도와 추적성 외력 허용 여부, 근사해 대 고전해 STOP 중단 추가 실행을 멈춤 기록과 불확실성 예산 종료, 독립성 가능성 표지 표 7 후보 수정 연산자

이 표는 ‘항상 R에서 시작해 R, R, R, ..., R 순으로 올라가라’는 규칙이 아니다. 때로는 타입이 맞지 않는 형식화가 즉시 R을 요구하고, 계산 자원 부족은 R만 바꾸면 되며, 명제가 모순되면 국소 수리를 더 하는 것이 낫다. 수정 연산자 사이에는 전역 총순서가 아니라 과제별 부분순서와 비용 함수가 있다.

5.4 수정 범위의 비식별성

정의 1. 두 수정 가설 h_i 와 h 가 현재까지 수행한 시험 집합 A 의 모든 a 에 대해 같은 관찰 분포를 만들면 A 에 관하여 수정-동치라고 한다. $h \sim_A h_i \Leftrightarrow P(Y | h, a, H) = P(Y | h_i, a, H)$ for all $a \in A$

(2) 명제 1. 적어도 두 개의 서로 다른 수정 범위 가설이 A 에 관하여 수정-동치라면, A 에서 얻은 잔차만을 입력으로 하는 어떤 선택 규칙도 두 가설을 일반적으로 식별할 수 없다. 이는 복잡한 정리가 아니라 식별 가능성의 정의에서 바로 따른다. 동일한 가능도 아래에서는 추가 잔차가 사후 오즈를 바꾸지 못하며, 다른 선택 확률이나 임의 규칙에 의존할 뿐이다.

5.5 최소 반례

입력이 $x \in \{0, 1\}$ 로 제한되고 현재 예측이 $\hat{y} = x$, 관찰이 $y = 2x$ 라고 하자. 세 세계를 생각한다. A에서는 같은 표현과 구조에서 계수만 2로 바꾸면 된다. B에서는 참 구조가 $y = x + x^2$ 라서 구조 향이 빠졌다. C에서는 참 표현이 $z = x^2$ 이고 $y = 2z$ 라서 입력 표현이 잘못되었다. 그러나 0과 1에서는 $x = x^2$ 이므로 세 세계 모두 $y = 2x$ 를 만든다. 세계 참 관계 적절한 수정 $x=0 \ x=1 \ x=2 \ A \ y=2x$ 매개변수 $R \ 0 \ 2 \ 4 \ B \ y=x+x^2$ 구조 $R \ 0 \ 2 \ 6 \ C \ z=x^2, \ y=2z$ 표현 $R \ 0 \ 2 \ 8$

표 8 같은 잔차와 다른 수정 원인의 반례 현재 자료에서 $\hat{y} = 2x$ 로 고치면 세 세계를 모두 완벽히 맞춘다. 따라서 국소 수정의 성공은 국소 수정이 원인 수준에서 옳았다는 증거가 아니다. $x=2$ 라는 판별 입력을 추가하면 세 가설은 4, 6, 8이라는 다른 예측을 내므로 구별된다. 결론은 두 줄이다. 첫째, 잔차의 모양은 수정 범위를 직접 지시하지 않는다. 둘째, 국소 수정의 훈련 구간 성공은 올바른 수정 범위의 충분조건이 아니다.

5.6 실패 증거 벡터

실무에서는 하나의 잔차보다 여러 증거를 결합한다. 수정 수준 L 에 대한 증거를 다음 벡터의 함수로 둘 수 있다. $E_L = f(P, C, B, F, M, N)$ (3) 기호 증거 무엇을 시사하는가 왜 단독으로 부족한가 P 지속성 persistence 같은 수정 후 오류가 남음 탐색 · 계산 부족일 수 있음 C 맥락 간 재현 특정 예제보다 일반적 문제 공유된 데이터 편향일 수 있음 B 방향성 편향 체계적 누락 · 부호 · 스케일 여러 구조가 같은 편향 생성 F 국소 수리 실패 더 큰 변경 후보 국소 수리 구현이 나빴을 수 있음 M 모델 가족 전체 실패 표현 · 가정 문제 후보 가족 탐색이 불완전할 수 있음 N 새 특징 의존 기존 표현의 정보 부족 새 특징이 우연 상관일 수 있음 표 9 수정 범위 판단을 위한 실패 증거 어떤 한 성분도 충분조건이 아니다. 증거 벡터의 역할은 곧바로 수정 레이블을 출력하는 것이 아니라 경쟁 가설의 우선순위를 갱신하고 다음 판별 시험을 고르는 것이다. 특히 ‘세 번 실패했으니 표현을 바꾼다’ 같은 고정 임계값은 과제 난이도, 시험의 독립성, 실패 상관구조를 무시한다.

5.7 비용 민감 판별 실험

수정 가설 D 에 대한 불확실성을 줄이는 시험 a 를 선택하되, 시험 비용을 함께 고려한다. 가장 단순한 형태는 기대 상호정보량에서 비용을 뺀다. $a^* = \arg \max_a \{ I(D; Y | a, H) - \lambda c_{\text{test}}(a) \}$ (4) 시험 뒤에는 수정 행동 d 의 기대 손실을 최소화한다. 손실은 편집 비용뿐 아니라 검증된 성과를 버리는 비용, 같은 경로 반복, 전환 지연, 잘못된 결론의 위험을 포함한다.

$d^* = \arg \min_d E[L(d, D) | H, Y]$ (5) $L = c_{\text{edit}} + \alpha c_{\text{discard}} + \beta c_{\text{repeat}} + \gamma c_{\text{delay}} + \delta c_{\text{invalid}}$ (6) 식 (4)는 정보가 많은 시험이면 무엇이든 수행하라는 뜻이 아니다. 새로운 대형 형식화는 큰 정보를 주지만 지나치게 비쌀 수 있고, 작은 경계 사례나 차원 축소 예제가 더 효율적일 수 있다. 식 (6)의 가중치는 보편 상수가 아니다. 안전이 중요한 계산에서는 잘못된 결론 비용 δ 가 크고, 장기 수학 연구에서는 검증된 보조정리 폐기 비용 α 와 반복 경로 비용 β 가 클 수 있다. 이 가중치의 목적은 가치 판단을 숨기지 않는 것이다. 21

5.8 조건부 에스컬레이션 규칙

보편 계층 대신 다음 조건부 규칙을 제안한다. 6. 현재 실패가 검출되면 먼저 오류 증거와 검증된 부분 산출물을 동결해 기록한다. 7. 서로 다른 수정 범위를 대표하는 최소 두 개 이상의 원인 가설을 만든다. 8. 현재 이력에서 가설들이 구별되는지 확인한다. 구별되지 않으면 깊이를 결정하지 않는다. 9. 가설의 예측을 갈라놓는 가장 저비용 시험을 수행한다. 10. 시험 후 기대 손실이 가장 낮은 수정을 선택하되, 보존 가능한 V 를 명시한다. 11. 새 맥락과 보류 자료에서 재검사한다. 국소 성공만으로 원인 진단을 확정하지 않는다. 12. 추가 시험의 기대 가치가 비용보다 낮거나 목표가 정의되지 않으면 중단 또는 명료화를 선택한다. FAILURE → COMPETING HYPOTHESES → DISCRIMINATING TEST → CONDITIONAL REVISION → RETEST

6 나비에 – 스토크스 방정식 사례

6.1 왜 이 사례가 수정 범위 문제인가

나비에 – 스토크스 방정식은 물과 공기 같은 유체의 속도와 압력이 시간에 따라 어떻게 변하는지를 기술한다. 그러나 ‘방정식을 푼다’는 말은 한 가지가 아니다. 특정 초기 · 경계조건의 수치해를 계산하는 것, 약한 해의 존재를 보이는 것, 모든 매끄러운 초기자료에 대해 고전해가 영원히 매끄러운지 증명하는 것, 어떤 조건에서 특이점이 생기는 한계를 구성하는 것은 서로 다른 목표다. 외력의 허용 여부, 공간이 전 공간인지 주기적 상자인지, 해의 정칙성과 에너지 조건이 무엇인지가 바뀌면 같은 명칭 아래의 문제가 달라진다. 따라서 이 사례는 목표 G와 표현, ρ 구조 S를 구별해야 하는 이유를 한꺼번에 보여 준다. 22 21 David Ma et al., “OProver: A Unified Framework for Agentic Formal Theorem Proving,” 2026, <https://arxiv.org/abs/2605.17283>; Zhuo Liu, Ding Yu, and Hangfeng He, “Compiler - Guided Adaptive Proof Search with Cross - Model Synergy,” Findings of EMNLP 2026, <https://arxiv.org/abs/2608.18084>. 22 Kunlun Zhu et al., “AgentDebugX: An Open - Source Toolkit for Failure Observability, Attribution, and Recovery in LLM Agents,” 2026, <https://arxiv.org/abs/2607.18754>; Dongsheng Zhu et al., “When Tools Fail: Benchmarking Dynamic Replanning and Anomaly Recovery in LLM Agents,” 2026, <https://arxiv.org/abs/2606.05806>.

6.2 방정식과 각 항의 물리적 의미

$\partial u + (u \cdot \nabla)u = -(1/\rho)\nabla p + \nu \Delta u + f, \nabla \cdot u = 0$ (7) $u(x,t)$ 는 위치 x 와 시간 t 에서의 속도장, p 는 압력, ρ 는 일정한 밀도, ν 는 운동점성계수, f 는 단위 질량당 외력이다. 식 (7)의 왼쪽은 유체 입자를 따라가며 본 가속도이고, 오른쪽은 그 입자에 작용하는 힘이다. 결국 뉴턴의 제2법칙을 점이 아니라 연속체의 작은 유체 요소에 적용한 것이다. 항 수학적 역할 물리적 현상 직관 ∂u 한 위치에서 속도의 시간 변화 국소 가속 카메라를 고정했을 때 그 점의 흐름이 빨라지거나 느려짐 $(u \cdot \nabla)u$ 속도가 자기 자신의 공간기울기를 운반 비선형 대류 가속 입자가 서로 다른 속도 영역으로 이동해 스스로 가속됨 $(1/\rho)\nabla p$ - ρ 압력의 공간 기울기 압력에 의한 내부 힘 높은 압력에서 낮은 압력 쪽으로 밀림 $\nu \Delta$ 속도장의 라플라시안 점성 확산 빠른 층이 느린 층을 끌어 속도 차를 퍼 줌 f 외부에서 주는 체적력 중력 · 전자기력 · 인위적 구동 유체 밖의 원인이 운동량을 넣거나 뺌 $\nabla \cdot u=0$ 속도장의 발산이 0 비압축성 · 부피 보존 작은 유체 덩어리가 이동해도 부피가 변하지 않음 표 10 비압축성 나비에 – 스토크스 방정식의 각 항

6.3 대류항은 왜 어려운가

대류항 $(u \cdot \nabla)u$ 가 비선형인 이유는 미지수 u 가 미분되는 대상이면서 동시에 그 미분의 방향과 운반 속도를 정하기 때문이다. 속도를 두 배로 하면 이 항은 단순히 두 배가 아니라 대략 네 배로 커진다. 서로 다른 규모의 소용돌이가 결합해 에너지를 전달하고, 작은 오차가 다른 규모로 이동할 수 있다. 선형 방정식에서는 두 해를 더해 도해가 되지만 이 항 때문에 중첩 원리가 깨진다. 비압축성과 적절한 경계조건 아래에서 대류항은 전체 운동에너지를 직접 만들거나 없애지 않는다. 대신 에너지를 위치와 규모 사이에서 재배치한다. 난류에서 큰 소용돌이의 에너지가 작은 소용돌이로 전달되는 캐스케이드와 연결되는 이유다. 문제는 점성이 작은 규모에서 에너지를 소산하기 전에 비선형 집중이 얼마나 빠르게 기울기와 와도를 키울 수 있는가에 있다.

6.4 압력은 단순한 입력이 아니다

압력 p 는 흔히 독립적으로 주어진 힘처럼 보이지만, 비압축성 방정식에서는 $\nabla \cdot \mathbf{u} = 0$ 을 유지하도록 속도장을 전역적으로 조정하는 라그랑주 승수의 역할을 한다. 식 (7)에 발산을 취하면 압력에 대한 포아송 방정식이 생기므로, 한 지점의 압력은 주변 전체 속도 구조와 연결된다. 이 비국소성 때문에 ‘특이점이 한 점에서 생긴다’는 그림도 주변 흐름과 분리해 해석할 수 없다.

6.5 점성항과 오일러 방정식

점성항 $\nu \Delta$ 는 공간적으로 거친 속도 차이를 평균화하는 확산이다. 열방정식이 온도 차를 퍼는 것과 같은 수학적 형식을 갖는다. 점성은 운동에너지를 내부에너지로 전환하여 작은 규모의 요동을 감쇠한다. 반면 $\nu = 0$ 으로 놓으면 오일러 방정식이 된다. 오일러 유체는 이상적인 비점성 유체이며, 경계층과 실제 소산을 직접 표현하지 않는다. 클레이의 나비에-스토크스 문제는 $\nu > 0$ 이고, 오일러 방정식의 정칙성 문제는 중요하지만 별도의 문제다. $\partial \mathbf{u} + (\mathbf{u} \cdot \nabla) \mathbf{u} = - (1/\rho) \nabla p + \mathbf{f}$, $\nabla \cdot \mathbf{u} = 0$ ($\nu = 0$) (8) 점성이 있으니 당연히 모든 거칠기가 사라질 것 같지만, 3차원에서는 비선형 집중이 점성 확산보다 빠르게 작은 규모를 만들 가능성을 일반적으로 배제하지 못했다. 대표 무차원수인 레이놀즈수 $Re = UL/\nu$ 는 관성 효과와 점성 효과의 상대 크기를 나타낸다. Re 가 크면 대류가 상대적으로 강하고 다양한 규모가 활성화되어 난류적 거동이 나타나기 쉽다. 그러나 높은 Re 의 난류 관찰이 곧 수학적 유한시간 blowup의 증거는 아니다.

6.6 에너지와 와도

충분히 매끄럽고 경계항이 사라지는 경우 운동에너지 균형은 다음과 같다. $d/dt [\int |\mathbf{u}|^2 dx] + \nu \int |\nabla \mathbf{u}|^2 dx = \int \mathbf{f} \cdot \mathbf{u} dx$ (9) 왼쪽 첫 항은 운동에너지 변화, 둘째 항은 점성 소산, 오른쪽은 외력이 유체에 공급하는 일률이다. 외력이 없으면 매끄러운 해의 전체 운동에너지는 증가하지 않는다. 그러나 L^2 에너지가 유한하다는 사실만으로 점별 속도나 미분이 유한하다고 보장되지는 않는다. 좁은 영역에 큰 진폭이 집중되면 전체 적분은 유한하면서 최댓값은 커질 수 있다. 최신 blowup 구성에서 ‘에너지는 유계지만 속도 최댓값은 무한대로 간다’는 문장이 가능한 이유다. 와도 $\boldsymbol{\omega} = \nabla \times \mathbf{u}$ 는 유체 요소의 국소 회전을 나타낸다. 3차원 비압축성 흐름의 와도 방정식은 다음과 같다. $\partial \boldsymbol{\omega} + (\mathbf{u} \cdot \nabla) \boldsymbol{\omega} = (\boldsymbol{\omega} \cdot \nabla) \mathbf{u} + \nu \Delta \boldsymbol{\omega} + \nabla \times \mathbf{f}$ (10) 오른쪽 첫 항 $(\boldsymbol{\omega} \cdot \nabla) \mathbf{u}$ 는 와도 신장이다. 소용돌이관이 늘어나면 단면이 줄고 회전이 강해질 수 있다. 2차원에서는 이 신장 메커니즘이 사라지므로 전역 정칙성이 훨씬 잘 통제된다. 3차원에서는 신장에 의한 증폭과 점성 확산의 경쟁이 정칙성 문제의 중심에 놓인다. 오일러의 경우 Beale-Kato-Majda 조건은 유한시간 특이점이 생

긴다면 와도의 L^∞ 노름을 시간 적분한 값이 발산해야 한다는 필요조건을 준다. 이는 blowup을 특정 관측량의 폭주와 연결하지만, 실제 blowup의 존재를 스스로 증명하지는 않는다. 23

6.7 특이점과 무한 발산의 뜻

blowup은 유체 한 방울이 물리적으로 무한 속도에 도달한다는 관측 예측과 동일하지 않다. 수학적으로는 유한한 시간 T 에 접근하면서 속도, 와도, 혹은 특정 정칙성 노름이 유계로 유지되지 않아 고전적 매끄러운 해를 계속 연장할 수 없다는 뜻이다. 어떤 노름이 발산해야 하는지는 해의 개념과 정칙성 기준에 따라 달라진다. 특이점 뒤에도 약한 해가 존재할 수 있지만, 유일성과 물리적 선택이 보장되지 않을 수 있다. 연속체 모형은 분자 규모 이전까지만 물리적으로 유효하므로 실제 무한대가 나타나기 전에 압축성, 열, 상변화, 분자성 같은 빠진 물리가 개입할 수 있다. 그럼에도 blowup 질문은 중요하다. 첫째, 방정식이 매끄러운 초기 조건에서 예측을 유일하게 계속할 수 있는지 묻는다. 둘째, 수치해석의 격자 세분화가 참 해로 수렴하는지에 대한 이론적 기반과 연결된다. 셋째, 난류의 작은 규모 생성과 에너지 전달을 통제하는 부등식의 한계를 드러낸다. 넷째, 특이점이 있다면 나비에-스토크스 연속체 설명이 스스로 수정되어야 하는 경계가 존재함을 뜻한다.

6.8 파동과 유체의 관계

사용자가 언급한 ‘파동’은 구분이 필요하다. 비압축성 나비에-스토크스 식은 밀도 변화가 없는 이상화를 사용하므로 공기 중 음파처럼 압축과 팽창이 전파되는 음향파를 직접 포함하지 않는다. 자유수면이 있으면 중력과 와 모세관파가 경계조건을 통해 나타나고, 회전·성층 유체에서는 관성파·내부중력파가 생긴다. 충격파는 주로 압축성 오일러 또는 압축성 나비에-스토크스 방정식에서 다룬다. 따라서 ‘파동이 깨져 무한대가 된다’는 직관을 비압축성 3차원 정칙성 문제에 그대로 대입해서는 안 된다. 공통점은 비선형성이 에너지를 집중시키고 작은 규모를 만들 수

있다는 것이며, 구체적 특이점 메커니즘과 해 개념은 다르다.

6.9 2026년 OpenAI 논문의 정확한 범위

2026년 9월 공개된 166쪽 원고 Finite Time Blowup for Navier – Stokes는 모든 양의 점성 ν 에 대해 3차 원 비압축성 나비에 – 스토크스 방정식에서 처음 정지 상태로 시작하고, 시공간에서 매끄럽고 콤팩트하게 지지된 외력을 받으며, 운동에너지는 균일하게 유계지만 유한시간에 속도의 L^∞ 노름이 무한대로 가는 해를 구성한다고 주장한다. 원고의 정리 1.1은 전 공간 $\mathbb{R}^3$ 의 결과를 명시한다. 24 23 Jorma Rissanen, “Modeling by Shortest Data Description,” Automatica 14, no. 5 (1978): 465 – 471, [https://doi.org/10.1016/0005-1098\(78\)90005-5](https://doi.org/10.1016/0005-1098(78)90005-5). 최소 기술 길이는 적합도와 복잡도 사이 비용 균형의 고전적 배경이며, 본 연구는 이를 수정 · 폐기 비용까지 확장하는 비유적 출발점으로부터만 사용한다. 24 Charles L. Fefferman, “Existence and Smoothness of the Navier – Stokes Equation,” Clay Mathematics Institute official problem description, <https://www.claymath.org/wp-content/uploads/2022/06/navierstokes.pdf>. 공식 설명은 전 공간과 3차원 토러스에서 존재 · 매끄러움 또는 붕괴를 증명하는 네 선택지 A – D를 제시한다.

여기서 중요한 정정이 있다. ‘외력이 있으므로 클레이 문제와 무관하고, 외력이 없는 경우만이 공식 문제다’라고 단정하면 공식 문제 설명을 지나치게 좁게 읽는 것이다. Fefferman의 네 선택지 가운데 존재 · 매끄러움 선택지 A와 B는 $f \equiv 0$ 을 요구하지만, 붕괴 선택지 C와 D는 조건을 만족하는 매끄러운 외력 f 의 존재를 허용한다. OpenAI 공개 저장소는 해당 정형화가 C와 D를 목표로 한다고 명시한다. 따라서 원고가 정확하고 공식 조건을 충족하며 공동체가 받아들인다면 밀레니엄 문제의 부정적 방향 해법 후보가 될 수 있다. 동시에 ‘외력 없이 유체 스스로 blowup 하는가’라는 더 강한 물리 · 수학 질문은 별개로 남을 수 있다. 25 이 구분은 본 논문의 핵심을 그대로 재현한다. ‘외력 있음’이라는 한 조건은 국소 기술사항이 아니라 어떤 목표 G 를 해결했는지 판단하는 기준이다. 영상 요약의 잔차 – 외력이 있으니 아직 안 풀렸다는 판단 – 만으로는 충분하지 않다. 공식 목표 문서를 판별 시험으로 읽어야 한다. 반대로 공식 선택지에 포함된다는 사실만으로 증명 의 옳음과 물리적 일반성이 자동 확정되는 것도 아니다.

6.10 정형 검증이 보장하는 것과 보장하지 않는 것

Lean 검증은 형식화된 명제가 사용한 공리와 정의에서 논리적으로 따라오는지 기계적으로 확인한다. 긴 증명에서 국소 추론 오류와 누락된 의존성을 줄이는 매우 강한 장치다. 그러나 세 층을 구분해야 한다. 첫째, 커널이 검사한 형식 증명의 정확성. 둘째, 형식 명제가 논문의 자연어 정리와 정확히 대응하는지에 대한 형식화 충실성. 셋째, 그 정리가 클레이 공식 조건과 물리적 질문에 어떤 의미를 갖는지에 대한 해석이다. 첫째가 성립해도 둘째와 셋째는 독립 검토가 필요하다. 26

6.11 인정 절차와 현재의 합리적 표현

클레이 수학연구소 규칙은 제안 해법을 직접 제출받지 않으며, 적격 출판처에 발표되고 최소 2년이 지난 뒤 전 세계 수학 공동체의 일반적 승인을 얻어야 검토할 수 있다고 명시한다. 2026년 9월 10일 현재 공식 나비에 – 스토크스 페이지는 문제를 Active로 표시한다. 따라서 가장 정확한 문장은 ‘OpenAI가 공식 붕괴 선택지를 겨냥한 유한시간 blowup 원고와 Lean 인증서를 공개했으며, 수학적 · 제도적 검증이 진행될 매우 초기 단계다’이다. ‘해결 확정’과 ‘외력이 있어 전혀 관련 없음’은 모두 현재 증거보다 강한 표현이다. 27 25 Fefferman의 공식 설명은 Euler에서 유한 blowup 시간이 있다면 와도의 L^∞ 시간 적분이 발산해야 한다는 Beale – Kato – Majda 기준을 요약한다. 원전: J. T. Beale, T. Kato, and A. Majda, “Remarks on the Breakdown of Smooth Solutions for the 3 - D Euler Equations,” Communications in Mathematical Physics 94 (1984): 61 – 66, <https://doi.org/10.1007/BF01212349>. 26 OpenAI, “Finite Time Blowup for Navier – Stokes,” 166 - page manuscript, 2026, <https://cdn.openai.com/pdf/32d9f210-8b73-45e0-91bc-82a30aef8a9a/navier-stokes.pdf>. Abstract and Theorem 1.1 specify zero initial velocity, smooth compactly supported forcing, bounded kinetic energy, and finite - time unbounded velocity. 27 OpenAI, NavierStokesAndEuler public Lean repository, <https://github.com/openai/NavierStokesAndEuler>. README는 전 공간과 토러스 결과를 공식 설명의 대안 C와 D로 연결하고 Lean 4 정형화를 제공한다.

6.12 현실적 응용과 과장 방지

증명이 검증되더라도 항공기 설계나 일기예보가 다음 날 갑자기 바뀌는 것은 아니다. 공학은 이미 점성 유체를 유한한 격자와 난류모형으로 계산하며, 실제 유체는 압축성·열·분자 규모의 절단을 가진다. 단기적 가치는 특 정 blowup 메커니즘, 정착성 추정 한계, 수치 경고 지표, 정형 수학 워크플로를 연구하는 데 있다. 장기적으로 는 해의 존재·유일성·안정성에 관한 이론이 계산 신뢰도와 모델 선택의 근거를 강화할 수 있다. 응용 가능성은 ‘증명→즉시 제품’의 직선보다 ‘증명→새 구조 이해→새 추정·알고리즘→검증→응용’의 긴 경로로 보는 편이 정확하다.

7 대표 수학 난제 열 개와 해결의 여러 의미

7.1 선정 원칙

‘수학 10대 난제’는 국제적으로 고정된 단일 공식 목록이 아니다. 본 연구는 클레이 밀레니엄 문제 일곱 개에 연속체 가설, 페르마의 마지막 정리, 힐베르트의 열 번째 문제를 더한 교육·분석용 열 개를 사용한다. 열린 문제, 해결된 문제, 독립성으로 귀결된 문제, 일반 알고리즘의 불가능성이 증명된 문제를 함께 놓으면 ‘해를 구한다’는 말의 여러 뜻을 비교할 수 있기 때문이다. 클레이는 일곱 문제에 각 100만 달러를 배정했고, 2026년 9월 10일 현재 공식 목록에서 푸앵카레 추측만 해결로 분류하며 나머지 여섯은 미해결로 제시한다. 나비에-스토크스 최선 원고는 아직 공식 상태를 바꾸지 않았다. 28 문제 직관적 질문 해결 형식 핵심 연결 리만 가설 소수의 불규칙성이 얼마나 통제되는가 증명 또는 임계선 밖 영점 반례 수론·해석·알고리즘 경계 P 대 NP 검산이 빠르면 찾기도 빠른가 복잡도류 동일성 또는 분리 증명 최적화·암호·AI 탐색 나비에-스토크스 3차원 점성 흐름은 영원히 매끄러 운가 전역 정착성 또는 공식 붕괴 사례 PDE·난류·수치 신뢰성 양-밀스 질량 간극 양자 게이지장이 엄밀히 존재하고 최소 에너지가 양수인가 공리적 구성과 질량 간극 증명 수리물리·장론 호지 추측 위상적 흔적이 대수적 조각으로 오는가 유리 호지류의 대수주기 표현 증명/반례 대수기하·위상 버치-스위너턴다이어 타원곡선의 유리점 수를 L함수가 알려 주는가 해석적 영점 차수와 대수적 랭크 연결 수론·암호 간접 기반 푸앵카레 추측 구멍 없는 닫힌 3공간은 3구면인가 리치 흐름과 수술을 통한 증명 위상·기하·PDE 페르마 마지막 정리 $n > 2$ 에서 두 n제곱의 합이 n제곱인가 모듈러성 경우 증명 표현 전환의 전형 28 Jeremy Avigad et al., “Anatomy of a Formal Proof,” arXiv:2411.11885, 2024, <https://arxiv.org/abs/2411.11885>. 정형 증명의 구성요소와 검증 경계를 이해하기 위한 배경 문헌으로 사용한다.

문제 직관적 질문 해결 형식 핵심 연결 연속체 가설 정수 무한과 실수 무한 사이 크기 가 있는가 ZFC로부터 독립성 증명 공리 선택·수리논리 힐베르트 10번 정수 다항식 해 존재를 항상 판정 할 알고리즘이 있는가 그런 알고리즘의 불가능성 증명 계산가능성·정수론 표 11 분석용 대표 난제 열 개

7.2 리만 가설

리만 제타함수 $\zeta(s)$ 의 비자명 영점이 모두 실수부 1/2인 직선 위에 놓인다는 주장이다. 직관적으로 소수는 불규칙하게 나타나지만 완전한 무질서는 아니며, 리만 가설은 소수 개수의 평균적 근사에서 벗어나는 오차가 어느 정도까지 통제되는지를 말한다. 증명은 영점 전체를 지배하는 구조를 보여야 하고, 반례는 임계선 밖의 영점 하나면 충분하다. 해결되면 수많은 조건부 정리가 무조건부가 되고 소수 관련 알고리즘의 오차 경계가 날카로워진다. 그러나 현대 공개키 암호가 즉시 붕괴한다고 말하는 것은 과장이다. 대부분의 암호 안전성은 큰 수 분해·이산로그의 계산 난이도에 의존하며, 리만 가설 자체가 그 문제를 다항시간에 푸는 알고리즘을 주지는 않는다. 수정 관점의 교훈은 광범위한 수치 검증이 보편 명제의 증명이 아니며, 예외 하나를 찾는 검색과 전체 구조를 증명하는 전략의 비용이 비대칭이라는 점이다.

7.3 P 대 NP

P는 입력 크기의 다항시간 안에 풀 수 있는 결정문제의 집합이고, NP는 후보 해가 주어졌을 때 다항시간 안에 검증할 수 있는 문제의 집합이다. 질문은 $P=NP$ 인가이다. 퍼즐의 완성본을 빨리 검사할 수 있다면 완성본을 찾는 일반적 빠른 방법도 반드시 존재하는지를 묻는다. $P \neq NP$ 를 보이려면 가능한 모든 다항시간 알고리즘의 한계를 증명해야 하므로 몇 개 알고리즘이 실패했다는 관찰로는 부족하다. $P=NP$ 가 구성적으로 증명되고 실제 상수가 감당 가능하

면 조합 최적화, 자동 증명, 설계, 일정 계획에 큰 변화가 올 수 있고 많은 암호 가정이 흔들린다. 그러나 존재 증명이 비실용적 알고리즘만 주거나 차수가 지나치게 크면 단기 효과는 제한될 수 있다. $P \neq NP$ 가 증명되면 ‘더 좋은 탐색’의 한계를 명료하게 하고 근사, 무작위화, 매개변수화, 특수구조 활용으로 목표를 바꾸게 한다. 여기서 해결은 제품 하나보다 가능한 전략 공간의 지도를 다시 그리는 일이다. 29

29 Clay Mathematics Institute, “Rules for the Millennium Prize Problems,”

<https://www.claymath.org/millennium-problems/rules/>. The rules require publication in a qualifying outlet, a two-year interval, and general acceptance; the official Navier – Stokes page remained marked Active on 2026 - 09 - 10, <https://www.claymath.org/millennium/navier-stokes-equation/>.

7.4 나비에 – 스토크스 존재와 매끄러움

직관적 질문은 매끄럽게 시작한 3차원 점성 흐름이 유한시간 안에 통제 불가능한 작은 규모를 만들 수 있는 가이다. 해결은 모든 허용 초기자료에 대한 전역 매끄러움을 증명하거나, 공식 조건을 만족하면서 전역 매끄러운 해가 존재하지 않는 사례를 구성하는 방식이 가능하다. 제6절에서 보았듯 외력, 공간, 에너지, 해의 정칙성은 문제의 일부이며 부수 조건이 아니다. 해결의 즉각적 효과는 CFD 소프트웨어의 대체가 아니라, 어떤 추정이 원리상 가능한지와 특이점 근처에서 계산을 어떻게 해석할지에 대한 이론적 기준이다. 2026년 원고가 검증된다면 공식 붕괴 선택지를 해결하는 동시에 ‘외력 없는 blowup’과 ‘실제 유체의 분자적 거동’은 남겨 둘 수 있다. 하나의 정리 해결과 인접한 과학 질문의 종결은 같지 않다.

7.5 양 – 밀스 이론과 질량 간극

양 – 밀스 이론은 입자물리학의 게이지 상호작용을 기술하지만, 4차원 양자 이론을 수학적으로 엄밀하게 구성하고 진공 위 첫 들뜸 에너지 사이에 양의 간극이 있음을 증명하는 일은 남아 있다. 직관적으로 장의 방정식에는 명시적 질량이 없어 보이는데, 양자 상호작용은 유한한 에너지 문턱을 만들어 내는가를 묻는다. 해결은 물리학자가 쓰는 계산이 왜 일관된 수학적 대상을 이루는지, 비섭동 현상을 어떤 구조가 지탱하는지를 보여 줄 것이다. 표준모형 예측이 하루아침에 바뀌는 것보다 수리물리의 엄밀한 기반, 격자 게이지 이론과 연속극한 사이의 연결, 새로운 분석 도구가 주요 산출물일 가능성이 높다. 수정 관점에서는 ‘계산 결과가 맞는다’와 ‘대상이 존재한다’를 구별하는 사례다.

7.6 호지 추측

복소 대수다양체의 위상에는 코호몰로지류라는 전역적 흔적이 있고, 대수기하에는 다항방정식으로 정의되는 부분 공간인 대수주기가 있다. 호지 추측은 특정 유리 호지류가 대수주기들의 유리 선형결합에서 오는지를 묻는다. 쉽게 말해 위상적으로 감지한 모양의 일부를 실제 대수적 조각으로 만들 수 있는가이다. 해결은 위상 · 복소해석 · 대수기하 사이의 번역 가능 범위를 확장한다. 직접 산업 응용을 약속하기보다, 고차원 기하를 계산하고 분류하는 언어와 도구를 바꿀 가능성이 핵심이다. 이 문제는 표현 변경이 단순 표기 변환이 아니라 서로 다른 불변량 체계 사이의 의미 보존 사상이라는 점을 보여 준다.

7.7 버치 – 스위너턴다이어 추측

타원곡선의 유리점들은 덧셈이 가능한 군을 이룬다. 버치 – 스위너턴다이어 추측은 이 군의 자유도인 랭크가

곡선의 L함수가 $s=1$ 에서 몇 차로 0이 되는지와 같다고 말한다. 유리해의 대수적 풍부함을 해석함수의 국소 거동으로 읽어 내는 놀라운 사전이다. 30 해결은 타원곡선의 유리점을 이해하는 구조를 크게 확장하고 수론의 여러 조건부 결과를 정리할 것이다. 타원곡선 암호와 관련된 수학적 대상이라는 이유만으로 기존 암호가 즉시 깨진다고 결론 내릴 수는 없다. 직접 효과는 곡선 분류와 계산의 이론적 보증, 간접 효과는 수론 알고리즘과 더 넓은 산술기하의 새 기법이다. 이 사례의 수정 교훈은 한 표현에서 어려운 ‘해의 개수’를 다른 표현의 ‘함수 영점’으로 옮기는 전략이다.

7.8 푸앵카레 추측

모든 닫힌 3차원 다양체 가운데 단일연결인 것은 3차원 구면과 위상적으로 같은가라는 질문이다. 끈을 어느 고리 위에 걸어도 끊지 않고 한 점으로 줄일 수 있는 닫힌 3공간은 본질적으로 구면인가라는 직관이다. 페렐만은 해밀턴의 리치 흐름을 발전시켜 곡률을 열처럼 퍼뜨리 특이점이 생기면 통제된 수술을 시행하는 방식으로 더 강한 기하화 추측을 증명했다. 31 이 해결의 가장 큰 통찰은 위상적 분류 문제를 기하학적 편미분방정식의 시간 진화로 재표현한 데 있다. 흐름 도중 특이점은 증명의 실패가 아니라 잘라내고 추적해야 할 구조적 사건이 되었다. 이는 본 연구의 핵심 사례이기도 하다. 특이점을 무조건 없애려 하지 않고, 어떤 정보를 보존한 채 어떤 수술을 해야 하는지 설계하면 문제 구조가 바뀐다.

7.9 페르마의 마지막 정리

$n > 2$ 인 정수 n 에 대해 $x^n + y^n = z^n$ 을 만족하는 양의 정수해가 없다는 명제다. 문제 문장은 초등적이지만 와일스의 증명은 가상의 반례에서 프레이 타원곡선을 만들고, 타원곡선과 모듈러 형식의 연결을 이용해 모순을 이끌어냈다. 정수 방정식을 직접 공격하는 대신 전혀 다른 대상들의 대응으로 옮긴 것이다. 1993년 발표 뒤 증명에 틈이 발견되었고, 와일스와 테일러는 서로 따르는 부족했던 아이디어를 결합해 수리했다. 이 역사는 실패가 어느 수준의 수정을 요구했는지 보여 준다. 정리 전체나 모듈러성 전략을 버리지 않고 핵심 접합부를 재구성했다. 해결의 유산은 한 문장의 참·거짓을 넘어 모듈러성 올림 기법과 산술기하의 연결망이다. 32 30 Clay Mathematics Institute, “The Millennium Prize Problems,” <https://www.claymath.org/millennium-problems/>. 공식 페이지는 일곱 문제의 목적, 상금, 현재 분류와 짧은 설명을 제공한다. 31 Stephen Cook, “The Complexity of Theorem - Proving Procedures,” Proceedings of STOC 1971, pp. 151 – 158, <https://doi.org/10.1145/800157.805047>; Clay Mathematics Institute, “P vs NP,” <https://www.claymath.org/millennium/p-vs-np/>. 32 Clay Mathematics Institute, “Birch and Swinnerton - Dyer Conjecture,” <https://www.claymath.org/millennium/birch-and-swinnerton-dyer-conjecture/>.

7.10 연속체 가설

자연수의 가산 무한 와 실수의 크기 $\aleph_0, 2^{\aleph_0}$

$\aleph_0$ 사이에 중간 크기의 무한집합이 없는가를 묻는다. 괴델은 표준 집합론이 모순이 없다면 연속체 가설을 더해도 모순이 생기지 않음을 보였고, 코언은 forcing을 발명해 그 부정도 표준 공리와 양립할 수 있음을 보였다. 따라서 ZFC가 일관적이라는 조건 아래 CH는 ZFC로 증명도 반증도 되지 않는다. 33 이 결과는 ‘답을 못 찾았다’가 아니다. 현재 공리 체계가 질문을 결정하기에 충분하지 않다는 정확한 메타수학적 해결이다. 이후의 방향은 더 강한 공리를 선택하고 그 정당성과 수학적 생산성을 평가하는 것이다. 목표 수정 R^* 를 기술적 에스컬레이션과 분리해야 하는 이유가 여기서 분명해진다. 새 공리를 채택하는 일은 오류 수리만 이 아니라 어떤 수학적 세계를 연구할지 선택하는 규범적 결정이다.

7.11 힐베르트의 열 번째 문제

정수 계수 다항방정식이 정수해를 갖는지 모든 입력에 대해 판정하는 일반 알고리즘을 요구한 문제다. 데이 비스, 퍼트넘, 로빈슨의 작업을 마티야세비치가 완성하여 모든 재귀적으로 열거 가능한 집합이 디오판토스적으로 표현될 수 있음을 보였고, 그런 보편 판정 알고리즘은 존재하지 않는다는 부정적 답을 얻었다. 34 이 해결은 원하는 알고리즘을 제공하지 않았지만 어떤 프로그램도 할 수 없는 일을 증명했다. 현실적 효과는 무모한 일반 해법 탐색을 중단하고, 변수 수·차수·계수 영역·수체 같은 제한된 하위문제로 연구 목표를 전환하게 하는 데 있다. STOP과 목표 명료화가 실패가 아니라 지식 산출물이 될 수 있음을 가장 선명하게 보여 준다.

7.12 열 문제에서 얻는 공통 통찰

해결의 형식 대표 사례 산출되는 지식 연구 수정 보편 증명 페르마·푸앵카레 모든 허용 사례에 대한 필연성 표현·구조의 대전환 반례·붕괴 구성 나비에-스토크스의 가능한 부정 방향 한 보편 주장의 한계 조건·해 개념 재검토 독립성 연속체 가설 현재 공리의 결정 한계 공리·목표의 규범 모듈 알고리즘 불가능성 힐베르트 10번 보편 절차의

부재 제한된 하위문제로 전환 복잡도 분리 · 동일성 P 대 NP 효율적 계산의 구조적 경계 정확해 · 근사 · 특수구조선택 존재와 구조의 엄밀화 양-밀스 · 호지 · BSD 서로 다른 표현 사이의 연결 새 언어 · 불변량 개발 33

Grigori Perelman, “The Entropy Formula for the Ricci Flow and Its Geometric Applications,” arXiv:math/0211159, 2002, <https://arxiv.org/abs/math/0211159>; “Ricci Flow with Surgery on Three-Manifolds,” 2003, <https://arxiv.org/abs/math/0303109>; “Finite Extinction Time,” 2003, <https://arxiv.org/abs/math/0307245>. 34 Andrew Wiles, “Modular Elliptic Curves and Fermat’s Last Theorem,” Annals of Mathematics 141, no. 3 (1995): 443–551, <https://doi.org/10.2307/2118559>; Richard Taylor and Andrew Wiles, “Ring-Theoretic Properties of Certain Hecke Algebras,” Annals of Mathematics 141, no. 3 (1995): 553–572, <https://doi.org/10.2307/2118560>.

표 12 ‘푼다’는 말의 여섯 형식 공통점은 정답이 기존 문제문 안에서만 나오지 않는다는 점이다. 새로운 흐름, 대응, 불변량, 공리, 계산 불가능성의 언어가 문제의 의미를 다시 조직한다. 그러나 이 관찰을 ‘큰 실패는 항상 표현을 바꾸라’는 규칙으로 일 반화하면 안 된다. 와일스의 마지막 틈은 핵심 구조를 보존한 수리였고, 많은 정형 증명 오류는 한 줄 수정으로 해결된다. 사례가 주는 올바른 메타규칙은 수정 깊이를 숭배하는 것이 아니라, 경쟁 가설을 구별하고 보존 가치와 비용을 계산하라는 것이다.

8 수학 추론과 AI 벤치마크의 측정 공백

8.1 최종 정확도 중심 평가

MATH 같은 자연어 문제 벤치마크는 정답 정확도와 풀이 능력을 비교하는 데 큰 역할을 했지만, 같은 정답에 도달한 두 시스템이 실패 뒤에 어떤 경로를 택했는지는 거의 드러내지 않는다. 여러 후보를 무차별 재생성해 우연히 맞힌 시스템과, 오류를 진단하고 검증된 부분을 보존하며 수정한 시스템이 같은 점수를 받을 수 있다. pass@k 역시 예산을 늘려 성공 후보가 하나라도 나오면 높아지므로 실패 이후의 판단 품질과 비용을 분리하지 않는다. 정형 증명 벤치마크는 최종 증명이 커널에서 검증된다는 강점을 가진다. miniF2F, LeanDojo, PutnamBench, miniCTX는 난도 · 검색 · 검색 증강 · 긴 문맥 사용을 확장했다. 그러나 ‘증명됨’은 마지막 상태의 인 증이다. 실패한 시도의 원인이 잘못된 보조정리인지, 표현 불일치인지, 단지 전술 탐색 부족인지, 에이전트가 그 것을 어떻게 구분했는지는 별도로 주석이 없으면 알 수 없다. 35

8.2 과정 평가는 이미 존재한다

측정 공백을 과장해서는 안 된다. PRM800K는 단계 수준 사람 피드백을 대규모로 제공했고 ProcessBench는 최초 오류 단계를 찾는 능력을 직접 평가한다. AgentBoard의 진진률은 최종 실패 과제에서도 얼마나 목표에 접근했는지 보여 준다. TheoremBench는 보조정리 계층을 이용해 정리 내부의 부분 진전을 측정하고, FormalProofBench는 도구 사용 · 실패 유형 · 비용 · 지연을 보고한다. 36 35 Kurt Gödel, The Consistency of the Axiom of Choice and of the Generalized Continuum-Hypothesis with the Axioms of Set Theory, Princeton University Press, 1940, https://books.google.com/books?id=_1-YDwAAQBAJ; Paul J. Cohen, “The Independence of the Continuum Hypothesis,” Proceedings of the National Academy of Sciences 50 (1963): 1143–1148, <https://pmc.ncbi.nlm.nih.gov/articles/PMC221287/>. 36 Yuri V. Matiyasevich, “The Diophantineness of Enumerable Sets,” Doklady Akademii Nauk SSSR 191, no. 2 (1970): 279–282, <https://www.mathnet.ru/eng/dan35274>; Yuri Matiyasevich, Hilbert’s Tenth Problem, MIT Press, 1993, <https://books.google.com/books?id=42h5VRwG5rcC>.

Theory, Princeton University Press, 1940, https://books.google.com/books?id=_1-YDwAAQBAJ; Paul J. Cohen, “The Independence of the Continuum Hypothesis,” Proceedings of the National Academy of Sciences 50 (1963): 1143–1148, <https://pmc.ncbi.nlm.nih.gov/articles/PMC221287/>. 36 Yuri V. Matiyasevich, “The Diophantineness of Enumerable Sets,” Doklady Akademii Nauk SSSR 191, no. 2 (1970): 279–282, <https://www.mathnet.ru/eng/dan35274>; Yuri Matiyasevich, Hilbert’s Tenth Problem, MIT Press, 1993, <https://books.google.com/books?id=42h5VRwG5rcC>.

수리도 이미 평가된다. Baldur는 실패한 전체 증명과 오류 메시지를 입력으로 받아 증명을 수리한다. OProver는 실패-피드백-수리 궤적을 훈련 데이터로 만든다. 최신 적응형 검색은 부분적으로 맞는 증명이 후속 수정에서 망가질 수 있음을 인식하고 현재 최신 후보를 보존하며, 정제 시 새 시작점을 탐색한다. 일반 에이전트 연구의 ToolMaze와 CostBench는 동적 차단과 도구 실패 아래 재계획 · 비용 적응을 시험한다. 37

8.3 무엇이 아직 결합되어 있지 않은가

남은 공백은 과정 평가의 부재가 아니라 다음 다섯 조건의 결합이다. 첫째, 최초 오류 표면은 같지만 정답 수정 범위가 다른 쌍. 둘째, 에이전트가 추가 판별 시험을 선택할 수 있는 상호작용. 셋째, 최종 성공과 별도로 수정 범위 선택을 채점하는 레이블. 넷째, 검증된 부분 결과를 얼마나 보존했는지와 불필요한 재시작 비용. 다섯째, 계속 · 전환 · 명료화 · 중단의 보정된 선택이다. 벤치마크 · 방법 최종 성공 오류 위치 반복 수리 부분 진전 · 보존 수정 범위 정답 별칭 실패쌍 MATH ● - - - - miniF2F · LeanDojo ● 검사기 일부 시스템 검색 상태 - - ProcessBench - ● - - - - AgentBoard ● 분석 - ● 진전 - - FormalProofBench ● 분석 도구 루프 일부 - - OProver · Baldur ● 컴파일러 ● 귀적 암묵적 - 적응형 증명 검색 ● 컴파일러 ● ● 보존 정제 규칙 - ToolMaze · CostBench ● 도구 이상 재계획 비용 경로 적응 도구 도메인 RevisionScope - Math ● ● ● ● ● ● 표 13 검토 집합의 측정 차원 비교 · ●는 명시적 핵심 측정, ‘ 일부/암묵적 ’ 은 보조 분석 표 13은 논문 전문과 공개 설명을 바탕으로 한 설계 수준 비교이며, 모든 구현 세부사항을 완전히 재현한 체계적 감사는 아니다. 특히 2026년 연구는 빠르게 갱신되고 있으므로 논문 투고 직전 재검색이 필요하다. 이 표가 반증 되는 가장 간단한 길은 기존 수학 벤치마크가 이미 별칭 실패쌍, 수정 범위 레이블, 보존 비용, 중단 보정을 함께 제공함을 보이는 것이다. 그 경우 본 제안은 새 벤치마크가 아니라 그 자원의 확장 또는 재분석으로 위치를 바꾸어야 한다. 37 George Tsoukalas et al., “ PutnamBench: Evaluating Neural Theorem - Provers on the Putnam Mathematical Competition, ” NeurIPS 2024, <https://arxiv.org/abs/2407.11214>; Jiewen Hu et al., “ miniCTX, ” 2024, <https://arxiv.org/abs/2408.03350>.

8.4 왜 별칭 실패쌍이 필요한가

서로 다른 오류가 서로 다른 메시지를 내면 분류기는 표면 패턴만 외울 수 있다. 예를 들어 ‘ unknown identifier ’ 는 가져오지 않은 라이브러리, 잘못 번역한 개념, 오타자에서 각각 생길 수 있다. 첫 메시지를 같게 만든 쌍에서 한 사례는 import 추가가 정답이고 다른 사례는 표현 자체를 바꾸어야 한다면, 모델은 추가 시험을 통해 원인을 가려야 한다. 이런 쌍이 수정 범위 식별을 오류 유형 분류와 분리한다. 반사실적 쌍은 전체 문제를 거의 동일하게 유지하되 하나의 잠재 원인만 바꾼다. 관찰 이력을 제한해 초기에는 구별되지 않게 하고, 특정 시험을 요청하면 차이가 드러나게 한다. 핵심 채점점은 첫 수정이 아니라 ‘ 정보가 부족함을 인식하고 판별 시험을 선택했는가 ’ 가 될 수 있다. 이는 성급한 자신감보다 교정 가능성을 평가한다.

9 RevisionScope-Math 벤치마크 설계

9.1 평가 목표

RevisionScope - Math의 목표는 모델이 실패한 뒤 최종 답을 고치는 능력만이 아니라, 현재 증거가 허용하는 불확실성을 표현하고, 필요한 판별 시험을 선택하고, 적절한 수정 범위를 결정하고, 검증된 부분 결과를 보존하는 능력을 평가하는 것이다. 벤치마크는 자연어 수학, 프로그램 보조 계산, Lean 정형 증명의 세 트랙으로 구성 하되 처음에는 Lean 트랙과 소규모 자연어 트랙으로 시작하는 것이 현실적이다.

9.2 한 에피소드의 구조

단계 환경이 제공하는 것 에이전트 행동 기록

0 과제

명제 · 정의 · 허용 도구 · 예산 초기 계획 문제 표현

1 실패

부분 풀이 · 검사기 피드백 · 잔차 실패 여부 판단 검출

2 가설

제한된 궤적 원인 후보와 확률/순위 국소화

3 시험

사용 가능한 질문 · 실행 · 반례 생성 판별 시험 선택 정보 이득 · 비용

4 수정

시험 결과 수정 연산자와 보존 목록 범위 선택

5 재검사

보류 · 새 문맥 · 형식 검사 채택 · 추가 수정 · 중단 일반화 · 자연 표 14 벤치마크 에피소드

9.3 수정 범위 레이블

레이블은 $R, R, R, R, R, STOP$ 을 기본으로 한다. 단일 정답을 강제하지 않고 ‘ 허용 가능한 최소 수정 집합 ’ 을 함께 제공한다. 예를 들어 국소 수리와 탐색 정책 전환이 모두 합리적이면 $\{R, R\}$ 을 정답 집합으로 두고 비용에 따라 순위를 정한다. 레이블에는 수정 대상, 보존해야 할 부분, 금지되는 의미 변경, 판별 시험, 시험 결과별 권장 행동을 포함한다.

필드 정의 예시 surface_failure 처음 보이는 오류 Lean type mismatch latent_cause 제작자가 통제된 근본 원인 정의역 번역 오류 admissible_scopes 비용상 허용되는 수정 범위 집합 $\{R\}$ discriminating_tests 원인 후보 예측을 가르는 시험 경계값 $0 \cdot 1 \cdot 2$ 평가 preserve_set 수정 뒤에도 남겨야 할 검증 결과 보조정리 L1 - L4 invalid_shortcuts 문제 의미를 훼손하는 쉬운 변경 가정을 결론에 추가 stop_condition 추가 탐색의 가치가 낮은 조건 예산 · 독립성 증거 표 15 사례 주식 스키마

9.4 실패 주입 방식

골드 풀이 · 증명 · 형식화가 있는 과제에 통제된 실패를 주입한다.

국소 실패: 부호, 계수, 바인더, 전술 인자, 누락된 전제 하나를 바꾼다.

구조 실패: 불충분한 귀납가정, 잘못된 분해, 사용할 수 없는 보조정리 그래프를 넣는다.

표현 실패: 정의역, 좌표, 단위, 자연어 - 형식어 대응, 양화사 순서를 바꾼다.

탐색 실패: 성공 경로는 남아 있지만 에이전트가 같은 전술 가족을 반복하도록 유도하는 이력을 준다.

목표 실패: 조건 누락, 서로 모순된 요구, 공식 문제와 대중적 변형의 차이를 넣는다.

환경 실패: 도구가 일시적 · 영구적 · 암묵적으로 잘못된 결과를 주도록 한다. 주입된 오류가 인위적 장난감에 머물지 않도록 실제 실패 궤적에서 패턴을 수집한 뒤 통제 사례로 재구성한다. 원본 문제와 변형 문제 사이의 의미 차이를 전문가가 기록하고, Lean 트랙은 컴파일러와 커널 검증을 사용한다. 자연어 트랙은 두 명 이상의 수학자 독립 주석과 합의 절차가 필요하다.

9.5 관찰적 별칭쌍 제작

각 쌍은 동일한 문제 표면, 동일한 첫 실패 메시지, 비슷한 궤적 길이를 갖지만 잠재 원인이 다르다. 예를 들어 쌍 A는 실제로 import 누락이어서 R이 맞고, 쌍 B는 이름은 같지만 정의가 다른 라이브러리를 가져와 표현의 의미가 틀려 R가 맞다. 두 사례 모두 첫 메시지는 unknown constant로 시작한다. 에이전트가 라이브러리 선언을 조회하면 차이가 드러난다. 표면 메시지만 분류하는 모델은 우연 수준을 넘기 어렵고, 능동 조회 모델은 정보 이득을 얻는다. 또 다른 쌍은 부분 증명 네 개가 모두 맞지만 전체 연결만 실패한 사례와, 보조정리 중 하나가 더 강한 숨은 가정을 사용해 전체가 실패한 사례다. 전자는 구조 연결만 수정하면 되고 V를 보존해야 한다. 후자는 해당 보조 정

리를 재검증해야 한다. 같은 최종 실패만 본다면 전면 재시작이 둘 다 해결할 수 있지만, 보존 비용 지표는 차이를 드러낸다.

9.6 데이터 분할과 누출 방지

무작위 문제 분할만으로는 템플릿 누출을 막기 어렵다. 별칭쌍의 두 구성원이 서로 다른 분할에 들어가면 잠재 원인 패턴이 유출될 수 있다. 따라서 쌍 · 원형 문제 · 정리 가족을 그룹으로 묶어 분할한다. 권장 분할은 (1) 동일 분야 새 문제, (2) 새 수학 분야, (3) 새 형식 라이브러리, (4) 새 오류 생성기, (5) 사람 연구 궤적 외부 검증이다. 공개 훈련 세트에는 레이블과 판별 시험을 제공하되 비공개 테스트 세트의 잠재 원인과 별칭 관계는 숨긴다.

9.7 상호작용 프로토콜

에이전트는 제한된 호출 예산 안에서 세 종류의 행동을 할 수 있다. 관찰 행동은 정의 조회, 경계 사례 계산, 보조 정리 재검사, 다른 표현으로 재형식화한 샌드박스 시험이다. 수정 행동은 연산자와 구체적 변경을 제출한다. 메타 행동은 불확실성 보고, 사용자에게 목표 명료화 요청, 중단이다. 환경은 미리 정의된 결과를 반환해 모델별 도구 사용 차이를 통제한다. 자유 도구 트랙은 별도로 운영한다.

9.8 기저선

기저선 규칙 무엇을 검증하는가 항상 국소 수리 첫 오류 주변만 변경 과소 수정 비용 k회 실패 후 재시작 고정 임계값 에스컬레이션 반복 횟수 규칙의 한계 무작위 수정 범위 비용 무시 균등 선택 분류 난이도 하한 오류 국소화 모델 ProcessBench형 비평 후 국소 수정 국소화와 선택의 차이 Self - Refine · Reflexion 언어 피드백 반복 자기 피드백 효과 컴파일러 수리 Baldur · OProver형 오류 기반 재생성 도구 피드백 수리 정제 감지 검색 최선 후보 보존 · 재표집 강한 최신 기저선 정보 이득 정책 식 (4)로 시험 선택 본 이론의 증가분 표 16 권장 기저선

10 평가 지표와 실증 계획

10.1 핵심 지표

지표 정의 좋은 방향 최종 정확도와 다른 점 RSA 허용 수정 범위 집합에 대한 선택 정확도 높음 첫 의사결정을 직접 평가 DTS 선택한 시험이 경쟁 가설을 실제로 분리한 비율 높음 정보 획득 행동 평가 UER 필요보다 큰 수정 · 전면 재시작 비율 낮음 낭비와 지식 손실 측정

지표 정의 좋은 방향 최종 정확도와 다른 점 IRR 필요보다 작은 수정을 반복한 비율 낮음 고착과 과소 수정 측정 PSR 검증된 부분 결과 중 수정 뒤 유지된 비율 높음 성공 과정의 보존성 측정 STL 충분한 판별 증거 뒤 올바른 전환까지 단계 수 낮음 전환 지연 측정 RFR 같은 실패 원인을 같은 경로로 다시 만든 비율 낮음 반복 실패 측정 SCR 중단 · 계속 결정의 선택적 위험과 보정 보정 양호 무리한 답변 억제 CHR 오류 문맥에서 수정 후 성공률 높음 국소 과적합 분리 CNR 총 비용당 회복 성공 또는 효용 높음 예산으로 덮은 성공 제어 표 17

RevisionScope - Math 지표 RSA는 단일 레이블 정확도가 아니라 허용 집합과 비용 순위를 반영하는 적절성 점수로 구현한다. 예측 분포 $q(d)$ 에 대해 정답 분포 또는 허용 효용 $u(d)$ 를 사용한 로그점수 · Brier 점수를 함께 보고하면 불확실성 보정을 평가할 수 있다. 모델이 증거 부족을 정확히 인식해 두 후보에 확률을 나누는 행동을, 근거 없이 하나를 확신하는 행동보다 높게 평가해야 한다. $RSA_soft = \sum_d q(d) \cdot u(d)$, $Brier = \sum_d (q(d) - p^*(d))^2$ (11)

10.2 판별 시험 점수

DTS는 시험을 실행한 뒤 잠재 원인에 대한 엔트로피가 얼마나 감소했는지, 비용 대비 얼마나 효율적인지를 함께 측정한다. 벤치마크 제작자가 모든 시험 결과의 조건부분포를 알면 실제 정보 이득을 계산할 수 있고, 결정 론적 사례에서는 후보 집합의 분할 크기로 근사할 수 있다. $DTS(a) = [H(D|H) - E H(D|H, Y=y)] / [1 + \lambda c_test(a)]$ (12) 가장 많은 정보를 주는 시험이 최종 목표에 가장 유용한 것은 아닐 수 있다. 예를 들어 원인을 완벽히 구별하지만 증명을 새로 형식화해야 하는 시험보다, 두 비싼 수정만 배제하는 값싼 경계 사례가 더 나을 수 있다. 따라서 정보 이득 점수와 결정 효용 점수를 따로 보고, 하나의 종합점수로 모든 모델 순위를 고정하

지 않는다.

10.3 부분 성과 보존

검증된 부분 결과 집합 V 의 각 원소에 중요도 가중치 w_i 를 부여하고, 수정 뒤에도 유효하게 재사용된 집합 V_{i-1} 과의 가중 교집합으로 PSR을 계산한다. 단순히 텍스트를 복사한 것은 보존이 아니다. 새 문제와 수정된 정의 아래에서도 검사에 통과해야 한다. $PSR = [\sum_{v \in V \cap V_{i-1}} w_i] / [\sum_{v \in V} w_i]$ (13)

전면 재시작이 때로는 정당일 수 있으므로 PSR을 무조건 최대화하면 안 된다. 잘못된 공리나 표현에 의존한 부분 결과를 보존하면 위험하다. 그래서 제작자는 ‘보존 가능’, ‘재검증 필요’, ‘폐기 필요’를 구분해 주석한다. 모델은 보존 이유를 제출하고 환경은 의존 그래프를 따라 재검사한다.

10.4 전환과 중단

STL은 첫 실패부터가 아니라 올바른 전환을 정당화할 판별 증거가 제공된 시점부터 센다. 그렇지 않으면 신중하게 시험한 모델이 성급히 전환한 모델보다 불리해진다. STOP은 실패로 일괄 처리하지 않고, 추가 행동의 기대 가치가 0 이하인 사례에서 정당 행동으로 포함한다. 반대로 풀 수 있는 문제에서 조기 중단하면 선택적 위험이 증가한다. Continue iff $EVSI(\text{next test or revision}) > \text{cost}(\text{next action})$ (14)

10.5 파일럿 실험

최초 논문이 감당할 수 있는 파일럿은 다음 규모가 적절하다. 항목 최소안 권장안 성공 기준 원형 문제 50 100 대수 · 해석 · 조합 · 수론 균형 별칭쌍

100 에피소드

300 에피소드

각 범위 최소 30개 전문가 주석 2인+합의 3인+심판 Krippendorff α 또는 ≥ 0.67 κ 탐색 기준 모델 4

8 이상

폐쇄 · 개방 · 전문화 모델 포함 기저선 4 8 고정 임계값과 적응형 검색 필수 반복 실행

3 seed

5 seed

분산 · 비용 보고 외부 검증 없음 새 기관 1곳 레이블 재현성 확인 표 18 파일럿 규모 효과 분석은 문제 난도와 도메인을 고정효과, 원형 문제와 모델을 무작위효과로 둔 계층 모형을 사용할 수 있다. 핵심 비교는 정보 이득 정책이 강한 정체 감지 기저선보다 RSA와 CNR을 높이는지, 최종 성공률이 같을 때 UER · RFR · STL이 낮은지다. 모든 프롬프트, 도구 결과, 토큰 · 시간 비용, 실패 궤적, 주석 변경 이력을 공개하는 것이 바람직하다.

10.6 사전등록 가설

13. H1. 별칭 실패쌍에서 첫 오류 메시지만 본 모델의 수정 범위 정확도는 비별칭 사례보다 유의하게 낮다. 14. H2. 판별 시험을 허용하면 최종 성공률을 크게 해치지 않으면서 UER와 IRR이 감소한다. 15. H3. 강한 최신 증명 수리 시스템도 최종 pass@k 대비 RSA · PSR에서 서로 다른 순위를 보인다. 16. H4. 고정 k회 실패 후 전환 규칙은 과제별 비용을 반영한 정보 이득 정책보다 CNR이 낮다.

17. H5. PSR이 높은 시스템은 문맥이 바뀐 보류 세트에서 CHR이 높다. 18. H6. 불확실성 보고와 STOP을 허용하면 강제 응답 조건보다 선택적 위험이 낮다.

10.7 반증 조건

이 연구 프로그램은 다음 결과가 나오면 수정되어야 한다. 첫째, 전문가들이 수정 범위 레이블에 합의하지 못하고 허용 집합으로도 안정화되지 않는다. 둘째, RSA가 최종 성공 · 비용 · 보류 일반화를 전혀 예측하지 못한다. 셋째, 별칭쌍이 실제 연구 실패와 동떨어진 인공 패턴만 측정한다. 넷째, 강한 단순 규칙이 정보 이득 정책과 동등하거나 더 낫다. 다섯째, 기존 벤치마크의 재분석만으로 모든 제안 지표를 계산할 수 있다. 이 경우 새 데이터셋 대신 기존 자료의 주석 확장이 더 정당하다.

11 논문으로서의 장점과 한계

11.1 방어 가능한 장점

장점 왜 중요한가 원고에서의 구현 정직한 축소 중복된 거대 주장을 스스로 폐기 DROP · KEEP · DEFER 판정과 선행연구 충돌 공개 과업 분해 자기수정이라는 모호어를 검출 · 국소화 · 선택 · 검증으로 나눔 표 6과 독립 지표 형식 반례 직관이 아니라 비식별성을 구성적으로 보임 $x \in \{0,1\}$ 세 세계 능동적 해법 실패 횟수 규칙 대신 판별 시험 제시 정보 이득 - 비용 식 근접 선행 수용 OProver · 적응형 검색을 경쟁자로 인정 강한 기저선 포함 검증 가능성 독창성을 데이터 · 지표 · 반증으로 바꿈 RevisionScope - Math 프로토콜 사례의 깊이 나비에 - 스토크스의 조건 차이를 이론과 연결 공식 C/D와 unforced 질문 분리 교양과 연구의 연결 열 난제를 해결 유형별 비교 본문과 사례집의 기능 분리 출판 확장성 한 초고에서 두 논문 · 단행본으로 분리 가능 부록의 발전 계획 표 19 논문으로서의 차별점 가장 큰 장점은 주제의 크기가 아니라 실패를 다루는 방식이다. 이 원고는 ‘모든 지식은 수정된다’는 일반론에 머물지 않고, 언제 현재 증거가 수정 범위를 식별하지 못하는지, 그때 어떤 시험을 선택해야 하는지, 잘못된 전 환을 어떻게 수치화할지를 제안한다. 또한 최신 선행연구가 침범한 부분을 숨기지 않아 독창성의 실제 경계를 독자가 확인할 수 있다.

11.2 현재 한계

현재 초고에는 분명한 한계가 있다.

문헌 검토가 범위 지정 서사 검토이며 데이터베이스별 완전한 검색식 · 중복 제거 · PRISMA 흐름을 갖춘 체계적 문헌고찰은 아니다.

수정 연산자 집합은 후보 체계다. 도메인별로 합치거나 세분해야 하며 전역적 깊이 순서를 가정하지 않는다.

비식별성 명제 자체는 식별 이론의 일반 원리를 수학 추론에 적용한 것이므로, 순수 수학적 새 정리로 과장할 수 없다.

벤치마크는 설계 단계다. 데이터셋, 주석 일치도, 모델 결과, 통계적 검정이 생기기 전에는 실증 논문이 아니다.

사람의 연구 과정은 비선형이고 암묵적이어서 텍스트 궤적이 실제 인지 과정을 충실히 나타내지 않을 수 있다.

나비에 - 스토크스 2026년 원고는 매우 최근이다. 논문과 정형화의 독립 검증, 저자성 · 데이터 사용 논쟁은 본 초고가 판정할 수 없다.

열 난제의 현실 응용은 대부분 간접적이다. 새로운 증명이 언제 어떤 기술로 이어질지 예측하는 데 큰 불확실성이 있다.

11.3 과장해서는 안 되는 문장

피해야 할 주장 권장 표현 실패 패턴만 보면 수정 깊이를 알 수 있다. 실패 패턴은 후보 우선순위를 바꾸지만, 별칭 가설에는 판별 시험이 필요하다. 기존 벤치마크는 과정이나 복구를 전혀 보지 않는다. 과정 · 복구 평가는 빠르게 발전했지만 수정 범위 · 별칭쌍 · 보존 · 중단의 결합 측정이 제한적이다. RevisionScope - Math는 최초의 실패 벤치

마크다. 검토 범위에서 확인된 결합 공백을 겨냥하는 신규 설계다. OpenAI가 나비에-스토크스를 확정 해결했다. 공식 붕괴 선택지를 겨냥한 원고와 정형 인증서가 공개되었고 검증 초기 단계다. 외력이 있으므로 밀레니엄 문제와 무관하다. 공식 붕괴 선택지 C·D는 외력을 허용하지만 unforced blowup은 별도 질문이다. 수학 난제 해결은 곧산업 혁신이다. 직접 효과보다 구조·경제·도구를 통한 장기 간접 효과가 일반적이다. 표 20 투고 전 주장 강도 조정

11.4 이론 논문과 벤치마크 논문의 분리

통합 초고는 사고와 기획에는 유리하지만 한 편의 학술지 논문으로는 범위가 넓다. 첫 논문은 비식별성, 네 과업 분리, 판별 실험, 수학사 사례 두세 개에 집중하는 개념·철학 논문으로 만들 수 있다. 둘째 논문은 수정 범위 코드북, 별칭쌍, 데이터 구축, 모델 실험, 비용 지표를 중심으로 한 벤치마크 논문이 된다. 나비에-스토크스 상세 해설과 열 난제 비교는 단행본 또는 보충자료에 남긴다. 이 분리는 내용을 버리는 것이 아니라 독자와 검증 방식을 맞추는 편집이다.

12 결론

실패는 무엇인가 잘못되었다는 신호일 수 있지만 무엇을 바꾸어야 하는지까지 말해 주지는 않는다. 동일한 잔차는 국소 계수, 구조, 표현, 데이터, 탐색 정책의 문제에서 생길 수 있고, 현재 관찰 영역에서는 서로 구별되지 않을 수 있다. 그러므로 반복 실패를 수정 깊이의 자동 계단으로 읽는 보편 규칙은 폐기해야 한다. 살아남는 원리는 조건부다. 실패를 기록하고, 경쟁 수정 가설을 세우고, 가설이 다른 예측을 내놓게 하는 시험을 선택하고, 정보 이득과 수정·폐기·자연 비용을 함께 계산한 뒤 수정하고, 새로운 문맥에서 재검증한다. 이 과정에서 실패 검출, 원인 국소화, 수정 선택, 수정 검증은 각기 다른 능력이다. 최종 정답률은 이 능력들을 한 숫자에 숨긴다. 나비에-스토크스 사례는 문제 조건의 한 단어가 해결의 지위를 바꿀 수 있음을 보여 준다. 외력 있는 blowup은 물리적으로 더 강한 unforced 질문과 다르지만 클레이 공식 붕괴 선택지와 무관하지 않다. 정형 검증은 논리적 정확성을 강화하지만 형식화 충실성과 과학적 의미를 대신하지 않는다. 열 개 난제는 해결이 증명, 반례, 독립성, 불가능성, 표현의 발명으로 나타날 수 있음을 보여 준다. 본 프로젝트의 현재 핵심은 거대한 이해 이론이 아니라 작고 측정 가능한 질문이다. 동일한 실패 표면 아래에서 에이전트는 어떤 추가 정보를 구하고, 어느 범위를 수정하며, 무엇을 보존하고, 언제 멈추는가. RevisionScope - Math가 이 질문을 재현 가능하게 만들고 기존 강한 수리 시스템과 구별되는 결과를 낸다면 독창성 후보는 논문 기여가 된다. 그렇지 않다면 설계를 다시 수정해야 한다. 바로 그 반증 가능성이 이 프로젝트가 주장하는 교정 가능성을 스스로 충족하는 조건이다.

참고문헌 및 출처

웹 자료의 최종 확인일은 2026년 9월 10일이다. arXiv 원고는 동료평가 여부와 버전을 개별적으로 확인해야 한다. 아래 목록은 본문 각주에 인용한 핵심 자료와 벤치마크 비교에 사용한 원출처를 함께 제시한다.

- [1] Alchourr n, Carlos E., Peter G rdenfors, and David Makinson. 1985. “On the Logic of Theory Change: Partial Meet ó ä Contraction and Revision Functions.” *Journal of Symbolic Logic* 50(2): 510–530. <https://doi.org/10.2307/2274239>
- [2] Argyris, Chris, and Donald A. Sch n. 1978. *Organizational Learning: A Theory of Action Perspective*. Addison-ö Wesley. <https://books.google.com/books?id=2aYOAQAAMAAJ>
- [3] Avigad, Jeremy, et al. 2024. “Anatomy of a Formal Proof.” arXiv:2411.11885. <https://arxiv.org/abs/2411.11885>
- [4] Batchelor, G. K. 1967. *An Introduction to Fluid Dynamics*. Cambridge University Press. <https://doi.org/10.1017/CBO9780511800955>
- [5] Beale, J. Thomas, Tosio Kato, and Andrew Majda. 1984. “Remarks on the Breakdown of Smooth Solutions for the 3-D Euler Equations.” *Communications in Mathematical Physics* 94: 61–66. <https://doi.org/10.1007/BF01212349>
- [6] Bertozzi, Andrea L., and Andrew J. Majda. 2002. *Vorticity and Incompressible Flow*. Cambridge University Press. <https://doi.org/10.1017/CBO9780511613203>
- [7] Chaloner, Kathryn, and Isabella Verdinelli. 1995. “Bayesian Experimental Design: A Review.” *Statistical Science* 10(3): 273–304. <https://doi.org/10.1214/ss/1177009939>
- [8] Clay Mathematics Institute. “The Millennium Prize Problems.” <https://www.claymath.org/millennium-problems/>
- [9] Clay Mathematics Institute. “Navier–Stokes Equation.” <https://www.claymath.org/millennium/navier->

stokes - equation/

- [10] Clay Mathematics Institute. “Rules for the Millennium Prize Problems.” <https://www.claymath.org/millennium-problems/rules/>
- [11] Cohen, Paul J. 1963. “The Independence of the Continuum Hypothesis.” *Proceedings of the National Academy of Sciences* 50(6): 1143 – 1148. <https://pmc.ncbi.nlm.nih.gov/articles/PMC221287/>
- [12] Cook, Stephen A. 1971. “The Complexity of Theorem - Proving Procedures.” *Proceedings of STOC*, 151 – 158. <https://doi.org/10.1145/800157.805047>
- [13] de Regt, Henk W. 2017. *Understanding Scientific Understanding*. Oxford University Press. <https://global.oup.com/academic/product/understanding-scientific-understanding-9780190652913>
- [14] de Regt, Henk W., et al. 2026. “Scientific Understanding across Disciplines: In Varietate Concordia?” [arXiv:2608.14378](https://arxiv.org/abs/2608.14378). <https://arxiv.org/abs/2608.14378>
- [15] Fefferman, Charles L. 2000. “Existence and Smoothness of the Navier – Stokes Equation.” Clay Mathematics Institute official problem description. <https://www.claymath.org/wp-content/uploads/2022/06/navierstokes.pdf>
- [16] First, Emily, Markus N. Rabe, Talia Ringer, and Yuriy Brun. 2023. “Baldur: Whole - Proof Generation and Repair with Large Language Models.” <https://arxiv.org/abs/2303.04910>
- [17] Gabry, Jonah, Daniel Simpson, Aki Vehtari, Michael Betancourt, and Andrew Gelman. 2017. “Visualization in Bayesian Workflow.” <https://arxiv.org/abs/1709.01449>
- [18] Gelman, Andrew, et al. 2020. “Bayesian Workflow.” <https://arxiv.org/abs/2011.01808>
- [19] Gödel, Kurt. 1940. *The Consistency of the Axiom of Choice and of the Generalized Continuum - Hypothesis with the Axioms of Set Theory*. Princeton University Press. https://books.google.com/books?id=_1-YDwAAQBAJ
- [20] Gou, Zhibin, et al. 2024. “CRITIC: Large Language Models Can Self - Correct with Tool - Interactive Critiquing.” *ICLR 2024*. <https://arxiv.org/abs/2305.11738>
- [21] Grastien, Alban, Patrik Haslum, and Sylvie Thi baux. 2023. “A More General Theory of Diagnosis from First é Principles.” <https://arxiv.org/abs/2309.16180>
- [22] Hauser, Alain, and Peter B hlmann. 2012. “Characterization and Greedy Learning of Interventional Markov ü Equivalence Classes of Directed Acyclic Graphs.” *JMLR* 13: 2409 – 2464. <https://jmlr.org/papers/v13/hauser12a.html>
- [23] Hendrycks, Dan, et al. 2021. “Measuring Mathematical Problem Solving with the MATH Dataset.” *NeurIPS Datasets and Benchmarks*. <https://arxiv.org/abs/2103.03874>
- [24] Hu, Jiewen, et al. 2024. “miniCTX: Neural Theorem Proving with Long Contexts.” <https://arxiv.org/abs/2408.03350>
- [25] Huang, Jie, et al. 2024. “Large Language Models Cannot Self - Correct Reasoning Yet.” *ICLR 2024*. <https://arxiv.org/abs/2310.01798>
- [26] Lakatos, Imre. 1976. *Proofs and Refutations: The Logic of Mathematical Discovery*. Cambridge University Press. <https://books.google.com/books?id=HJ2NCgAAQBAJ>
- [27] Lightman, Hunter, et al. 2023. “Let ’ s Verify Step by Step.” <https://arxiv.org/abs/2305.20050>
- [28] Liu, Qi, et al. 2025. “Beyond Theorem Proving: Formulation, Framework and Benchmark for Formal Problem - Solving.” <https://arxiv.org/abs/2505.04528>
- [29] Liu, Zhuo, Ding Yu, and Hangfeng He. 2026. “Compiler - Guided Adaptive Proof Search with Cross - Model Synergy on Context - Dependent Theorem Proving.” *Findings of EMNLP 2026*. <https://arxiv.org/abs/2608.18084>
- [30] Liu, Jiayu, et al. 2025. “CostBench: Evaluating Multi - Turn Cost - Optimal Planning and Adaptation in Dynamic Environments for LLM Tool - Use Agents.” <https://arxiv.org/abs/2511.02734>
- [31] Ma, Chang, et al. 2024. “AgentBoard: An Analytical Evaluation Board of Multi - Turn LLM Agents.” <https://arxiv.org/abs/2401.13178>
- [32] Ma, David, et al. 2026. “OProver: A Unified Framework for Agentic Formal Theorem Proving.” <https://arxiv.org/abs/2605.17283>
- [33] Madaan, Aman, et al. 2023. “Self - Refine: Iterative Refinement with Self - Feedback.” <https://arxiv.org/abs/2303.17651>
- [34] Matiyasevich, Yuri V. 1970. “The Diophantineness of Enumerable Sets.” *Doklady Akademii Nauk SSSR* 191(2): 279 – 282. <https://www.mathnet.ru/eng/dan35274>
- [35] Matiyasevich, Yuri. 1993. *Hilbert ’ s Tenth Problem*. MIT Press. <https://books.google.com/books?id=42h5VRwG5rcC>
- [36] Narendra, Kumpati S., and Jeyendran Balakrishnan. 1997. “Adaptive Control Using Multiple Models.” *IEEE Transactions on Automatic Control* 42(2): 171 – 187. <https://doi.org/10.1109/9.554398>
- [37] OpenAI. 2026. “Finite Time Blowup for Navier – Stokes.” <https://cdn.openai.com/pdf/32d9f210-8b73->

45e0 - 91bc - 82a30aef8a9a/navier - stokes.pdf

- [38] OpenAI. 2026. NavierStokesAndEuler: Lean Certificates Accompanying Navier – Stokes and Euler Results. <https://github.com/openai/NavierStokesAndEuler>
- [39] Perelman, Grigori. 2002. “The Entropy Formula for the Ricci Flow and Its Geometric Applications.” <https://arxiv.org/abs/math/0211159>
- [40] Perelman, Grigori. 2003. “Ricci Flow with Surgery on Three - Manifolds.” <https://arxiv.org/abs/math/0303109>
- [41] Perelman, Grigori. 2003. “Finite Extinction Time for the Solutions to the Ricci Flow on Certain Three - Manifolds.” <https://arxiv.org/abs/math/0307245>
- [42] Pham, QuocViet, et al. 2026. “TheoremBench: Evaluating LLMs on Theorem Proving in Formal Mathematics.” <https://arxiv.org/abs/2606.09450>
- [43] Ravi, Nikil, et al. 2026. “FormalProofBench: Can Models Write Graduate Level Math Proofs That Are Formally Verified?” <https://arxiv.org/abs/2603.26996>
- [44] Reiter, Raymond. 1987. “A Theory of Diagnosis from First Principles.” *Artificial Intelligence* 32(1): 57 – 95. [https://doi.org/10.1016/0004 - 3702\(87\)90062 - 2](https://doi.org/10.1016/0004 - 3702(87)90062 - 2)
- [45] Rissanen, Jorma. 1978. “Modeling by Shortest Data Description.” *Automatica* 14(5): 465 – 471. [https://doi.org/10.1016/0005 - 1098\(78\)90005 - 5](https://doi.org/10.1016/0005 - 1098(78)90005 - 5)
- [46] Shinn, Noah, et al. 2023. “Reflexion: Language Agents with Verbal Reinforcement Learning.” <https://arxiv.org/abs/2303.11366>
- [47] Thurston, William P. 1994. “On Proof and Progress in Mathematics.” *Bulletin of the American Mathematical Society* 30: 161 – 177. <https://arxiv.org/abs/math/9404236>
- [48] Tsoukalas, George, et al. 2024. “PutnamBench: Evaluating Neural Theorem - Provers on the Putnam Mathematical Competition.” <https://arxiv.org/abs/2407.11214>
- [49] Wiles, Andrew. 1995. “Modular Elliptic Curves and Fermat ’ s Last Theorem.” *Annals of Mathematics* 141(3): 443 – 551. <https://doi.org/10.2307/2118559>
- [50] Taylor, Richard, and Andrew Wiles. 1995. “Ring - Theoretic Properties of Certain Hecke Algebras.” *Annals of Mathematics* 141(3): 553 – 572. <https://doi.org/10.2307/2118560>
- [51] Wilkenfeld, Daniel A. 2013. “Understanding as Representation Manipulability.” *Synthese* 190: 997 – 1016. <https://doi.org/10.1007/s11229 - 011 - 0055 - x>
- [52] Woodward, James. 2003. *Making Things Happen: A Theory of Causal Explanation*. Oxford University Press. <https://books.google.com/books?id=LrAbrrj5te8C>
- [53] Yang, Kaiyu, et al. 2023. “LeanDojo: Theorem Proving with Retrieval - Augmented Language Models.” *NeurIPS* 2023. <https://arxiv.org/abs/2306.15626>
- [54] Yao, Shunyu, et al. 2023. “ReAct: Synergizing Reasoning and Acting in Language Models.” *ICLR* 2023. <https://arxiv.org/abs/2210.03629>
- [55] Zheng, Chujie, et al. 2025. “ProcessBench: Identifying Process Errors in Mathematical Reasoning.” *ACL* 2025. <https://arxiv.org/abs/2412.06559>
- [56] Zheng, Kunhao, Jesse Michael Han, and Stanislas Polu. 2022. “miniF2F: A Cross - System Benchmark for Formal Olympiad - Level Mathematics.” *ICLR* 2022. <https://arxiv.org/abs/2109.00110>
- [57] Zhu, Dongsheng, et al. 2026. “When Tools Fail: Benchmarking Dynamic Replanning and Anomaly Recovery in LLM Agents.” <https://arxiv.org/abs/2606.05806>
- [58] Zhu, Kunlun, et al. 2026. “AgentDebugX: An Open - Source Toolkit for Failure Observability, Attribution, and Recovery in LLM Agents.” <https://arxiv.org/abs/2607.18754>