# 수학, 인공지능, 그리고 형식화 1 – 수학을 위한 인공지능

이시우 (UC Berkeley)

요즘에는 어디를 가도 모두가 인공지능(AI) 이야기를 하고 있습니다. 저는 주로 카페에서 일하는 편인데, 주변을 둘러보면 ChatGPT 로 과제를 해결하거나, Claude 와 Codex 의 도움을 받아 코드를 작성하는 학생들을 어렵지 않게 볼 수 있습니다. 더 나아가서, 인공지능은 의료 진단이나 기후 모델링, 신약 개발 등 사회 전반에 걸쳐서 실질적인 변화를 만들어내고 있고, 단순한 기술의 발전이 아닌 실제로 인류가 풀어야 하는 어려운 문제들에 대해서 조금씩 답을 내놓기 시작하고 있습니다. 물론 AI 활용에 우려되는 면이 없는 것은 아니지만, 전반적으로는 업무와 학습의 생산성을 크게 높이고 있다고 생각합니다.

이처럼 AI 가 다양한 분야에 스며들고 있는데, 수학 연구도 예외일 필요는 없을 것 같습니다. 이 글을 읽는 분이라면, AI 가 에르되시(Erdős) 문제를 풀었다거나 난제를 해결했다는 소식을 접한 적이 있을지도 모릅니다. 이 글에서는 AI 를 적극적으로 활용한 최신 수학 연구들을 좀 더 자세히 살펴보고자 합니다.

## 인공지능으로 증명하기

인공지능으로 수학 연구를 한다고 했을 때 가장 쉽게 떠올릴 수 있는 방향은, ChatGPT 나 Gemini 같은 LLM 에게 수학 명제의 증명을 물어보는 것입니다. 불과 2 년 전까지만 해도 9.9 와 9.11 중 무엇이 더 큰지 판단하지 못하고, "strawberry"에서 r 이 몇 개인지도 세지 못하는 인공지능에게 최신 수학 연구에 기여하기를 기대하는 것은 무리였습니다. 하지만 빠르게 성장해온 인공지능이 그 수준에 머물러 있을 리는 없었습니다. 시간이 흐르면서 ChatGPT 3 은 ChatGPT 5.2 가 되었고, 대소 비교도 하지 못하던 인공지능들은 이제 국제 수학 올림피아드(IMO)에서 금메달을 받거나[1] 미국의 대학생 수학 경시대회인 Putnam 에서 만점을 받는 수준으로 성장했습니다[2]. 이러한 대회에서 좋은 성적을 거두는 것은 9.9 가 9.11 보다 크다는 것을 아는 것보다 훨씬 어렵고, 단순히 답을 이끌어내는 것이 아닌 엄밀한 증명을 작성할 수 있어야 합니다.

물론 경시대회에서 좋은 성적을 거두는 것과 수학 연구를 잘 하는 것 사이에는 어느 정도 양의 상관관계가 있지만, 절대적인 것은 아닙니다. 2022 년 필즈상 수상자인 Hugo Duminil-Copin 은 본인이 경시대회에서 좋지

[1] Google DeepMind 공식 블로그: https://deepmind.google/blog/advanced-version-of-gemini-with-deep-think-officially-achieves-gold-medal-standard-at-the-international-mathematical-olympiad/

[2] Axiom Math 공식 블로그: https://axiommath.ai/territory/from-seeing-why-to-checking-everything

않은 성적을 거두었다고 밝힌 바 있습니다[3]. 그렇다면, IMO 에서 금메달을 받은 인공지능들은 아직 수학 연구를 하기에는 부족한 것일까요? 놀랍게도, 전혀 그렇지 않습니다. 이를 뒷받침하는 몇 가지 연구 사례를 소개합니다.

먼저 ChatGPT Pro 를 활용해 다양한 수학 연구 문제를 해결한 사례들이 있습니다. 볼록 최적화(convex optimization)에서 널리 쓰이는 네스테로프 가속 경사 하강 알고리즘(Nesterov Accelerated Gradient, NAG)은 1983 년 Yurii Nesterov 가 제안한 방법으로, 볼록함수 $f : \mathbb{R}^n \to \mathbb{R}$ 의 최솟값을 빠르게 찾는 알고리즘입니다. 이 알고리즘이 생성하는 수열 $x_1, x_2, x_3, \dots$에 대해 함숫값의 수열 $f(x_1), f(x_2), f(x_3), \dots$가 함수의 최솟값 $\min f$ 에 수렴한다는 사실은 잘 알려져 있었습니다. 그러나 수열 $x_k$ 자체가 최솟값을 주는 점 $x_\infty \in \operatorname{argmin} f$ 에 수렴하는지는 오랫동안 미해결 추측으로 남아 있었습니다. UCLA 의 류경석 교수와 장의정 박사과정 학생은 ChatGPT-5 Pro 와의 반복적인 대화를 통해 핵심 아이디어를 얻어 이 추측을 증명하는 데 성공했습니다[4]. 주목할 점은 한 번의 질문으로 완성된 증명을 얻어낸 것이 아니라, 여러 차례의 대화를 거쳐 틀린 증명들 사이에서 가능성 있는 아이디어를 발견하고 이를 발전시켜 나갔다는 것입니다. 이 외에도 이론물리 연구에서 ChatGPT 가 복잡한 공식을 예측하는 데 크게 기여한 사례도 있습니다[5].

구글 딥마인드는 최근 Gemini Deep Think 를 기반으로 한 수학 에이전트 Aletheia[6]를 이용한 연구 결과들을 발표했습니다. 에르되시 문제 웹사이트[7]에 공개된 700 여 개의 미해결 문제에 Aletheia 를 시도한 결과, Aletheia 의 자체 검증 메커니즘을 통해 700 개의 문제 중 212 개의 응답이 잠재적으로 정답일 가능성이 있다고 판단되었고, 인간 수학자들의 검토를 거쳐 63 개가 수학적으로 올바른 것으로 확인되었습니다. 그러나 이 중 문제의 의도에도 부합하는 것은 13 개였으며, 그 중 5 개는 기존 문헌에서 찾을 수 없는 독자적인 풀이로, 나머지 8 개는 이미 풀렸지만 데이터베이스에 반영되지 않은 문제들이었습니다. 연구팀은 스스로 이 결과를 겸손하게 평가하는데, 수십 년간 미해결로 남아 있었음에도 불구하고 실제로 해결된 에르되시 문제들은 수학적으로는 그리 어렵지 않은 수준이었다고 밝혔습니다. 더 나아가 Aletheia 프로젝트를 이끌고 있는 UC Berkeley 의 Tony Feng 교수는 자신의 연구 문제에도 Aletheia 를 활용했는데, 사람의 개입이 거의 없이 Arithmetic Hirzebruch Proportionality 에서 등장하는 특정 미분 연산자의 고유

[3] Quanta Magazine: https://www.quantamagazine.org/hugo-duminil-copin-wins-the-fields-medal-20220705/
[4] arXiv: https://arxiv.org/abs/2510.23513
[5] arXiv: https://arxiv.org/abs/2602.12176
[6] Google DeepMind 공식 블로그: https://deepmind.google/blog/accelerating-mathematical-and-scientific-discovery-with-gemini-deep-think/
[7] https://www.erdosproblems.com/

가중치(eigenweight)를 여러 타입의 대수적 군(algebraic group)에 대해 계산해냈다고 하고[8]. 그래프 이론[9]과 확률론 연구[10]에서도 주요 정리의 증명에 Aletheia 가 핵심적인 역할을 했음이 보고되었습니다.

Anthropic 의 Claude 역시 빠르게 성장하고 있습니다. 1974 년 튜링상 수상자이자 컴퓨터과학도에게 잘 알려진 *The Art of Computer Programming* (TAOCP)을 집필한 Stanford 대학교의 Donald Knuth 교수는, 자신이 몇 주간 고민하던 문제를 Claude Opus 4.6 이 해결해주었다고 밝혔습니다[11]. 문제는 $m > 2$ 일 때 $m^3$ 개의 꼭짓점을 가지고 각 꼭짓점마다 3 개의 나가는 간선(edge)을 가진 특정 방향 그래프(digraph)를 길이 $m^3$ 인 세 개의 사이클(cycle)로 분해하는 것으로, 곧 출간될 TAOCP 새 판에 미해결 문제로 수록될 예정이었다고 합니다. 논문의 첫 줄을 "Shock! Shock!"으로 시작할 만큼 Claude 의 결과에 크게 놀란 Knuth 의 반응은 큰 화제가 되었고, 논문이 공개된 후 얼마 지나지 않아 다른 연구자들이 LLM 을 활용해 새로운 증명을 찾아내거나 Claude 를 통해 Lean 형식 증명을 자동으로 생성하기도 했습니다[12].

X(구 트위터)가 개발 중인 Grok 역시 수학 연구에 조금씩 기여하고 있습니다. UC Irvine 의 해석학자 Paata Ivanisvili 교수는 이진 함수(Boolean function), 즉 ${-1, 1}^n$위에서 정의된 함수를 주로 연구하며, X 와의 협업을 통해 Grok 을 연구에 적극적으로 활용하고 있습니다. 그의 연구에서 중요한 미해결 문제 중 하나는 ${-1, 1}^n$에서 정의된 차수가 $d$ 이하인 모든 다항식 $f$ 에 대해서

$$ \|f\|_2 \le C^d \|f\|_1 $$

을 만족하는 최적의 상수 $C >0$를 찾는 것 입니다. $d = 1$인 경우 $\sqrt{2}$가 최적이라는 것은 이미 Kintchine 부등식에 의해서 알려져 있고 $d = 2$인 동차다항식(homogenous polynomial)의 경우에도 $\|f\|_2 \le 2 \|f\|_1$이 성립한다는 것이 Pelczyński 의 추측입니다. 이를 바탕으로 일반적인 차수의 경우에도 $C = \sqrt{2}$가 최적일 것이라는 추측이 있었는데, Ivanisvili 교수가 Grok 에게 이 문제에 대해서 물어보았더니 $C$가 충분히 큰 차원에 대해서 최소한 $\sqrt{3}$이상이어야 한다는 반례를 제시해주었다고 합니다[13].

---

[8] arXiv: https://arxiv.org/abs/2601.23245
[9] arXiv: https://arxiv.org/abs/2602.02450
[10] arXiv: https://arxiv.org/abs/2601.23229
[11] 논문: https://www-cs-faculty.stanford.edu/~knuth/papers/claude-cycles.pdf
[12] GitHub: https://github.com/kim-em/KnuthClaudeLean/. Lean 은 쉽게 말해 수학 증명을 컴퓨터의 언어로 옮겨적어서 증명이 맞다는 것을 컴퓨터로 검증할 수 있게 도와주는 언어 중 하나인데, 다음 편에서 더 자세히 알아볼 예정입니다.
[13] Grok 대화 로그: https://grok.com/share/c2hhcmQtNA_826731f3-afbc-42a5-aa61-acca3ec1324a?rid=de4a202d-357d-4ece-9954-16c0dc5951d5

## 인공지능으로 발견하기

단순히 LLM 과의 대화를 통해 원하는 증명을 얻어내는 것 이외에도, AI 를 좀 더 흥미롭게 활용하는 방법들이 있습니다. 원하는 정리의 증명을 직접적으로 묻는 대신, AI 를 이용해 흥미로운 수학적 발견을 하고, 이를 바탕으로 수학자가 동기를 얻어 새로운 결과를 이끌어내는 경우들도 많이 있습니다.

대표적인 예시 중 하나가 바로 구글 딥마인드의 AlphaEvolve 입니다[14]. AlphaEvolve 는 LLM 기반의 진화 모델로, 주어진 문제 상황과 최적화할 목표 점수가 주어졌을 때 이를 달성하는 코드를 작성하고 점차 개선해 나가는 방식으로 작동합니다. 이 구조는 조합 최적화(combinatorial optimization) 문제를 비롯한 상당히 많은 수학 문제에 자연스럽게 들어맞습니다. 예를 들어, 고차원 구의 표면에 같은 크기의 구를 겹치지 않게 배치하는 키싱 문제(kissing problem)는 공 쌓기 문제(sphere packing problem)와도 깊은 관련이 있는 고전적인 난제인데, AlphaEvolve 는 11 차원에서의 키싱 수에 대한 기존 기록인 592 를 593 으로 개선하는 데 성공했습니다. 또한 최소한의 곱셈 횟수로 행렬을 곱하는 알고리즘을 찾는 문제 역시 활발히 연구되는 분야인데, AlphaEvolve 는 $4 \times 4$ 행렬의 곱셈에서 1969 년부터 유지되어 오던 49 번이라는 기록을 48 번으로 줄이는 데 성공했습니다. 테렌스 타오는 AlphaEvolve 를 활용해 유한체(finite field) 위의 Nikodym 집합을 구성하는 문제를 연구했는데, 몇몇 작은 경우에 대한 AlphaEvolve 의 제안을 바탕으로 이를 모든 차원과 모든 유한체로 일반화하는 데 성공했습니다[15]. AlphaEvolve 의 두드러지는 강점 중 하나는, 단순히 예시를 제시하는 데 그치지 않고 이를 생성하는 파이썬 코드를 함께 제공하기 때문에 그 알고리즘을 어느 정도 해석할 수 있다는 것입니다. Nikodym 집합 문제의 경우, AlphaEvolve 의 코드가 실제로 하고 있었던 것은 전체 벡터 공간에서 특정 대수 초곡면(algebraic hypersurface) 몇 개를 제외하는 것이었는데, 이 아이디어가 일반화 과정에서 핵심적인 역할을 했다고 합니다.

제 연구도 하나의 예시로 소개하면서 이 절을 마치고자 합니다. 저는 정수론을 전공하고 있는데, 갈루아 군(Galois group)은 정수론에서 가장 핵심적인 대상 중 하나입니다. 체(field) $F$와 그 확장체 $E$가 있을 때, $\mathrm{Gal}(E/F)$는 $E$ 에서 $F$ 를 고정하는 자기동형사상(automorphism)들의 모임으로, 두 체 사이의 대칭성을 기술하는 군입니다. 약간의 과장을 덧붙이자면, 정수론의 궁극적인 목표 중 하나는 유리수 체의 절대적 갈루아 군(absolute Galois group)을 이해하는 것이라고도 할 수 있습니다. 갈루아 군은 그 중요성에 비해 일반적으로 계산하기가 쉽지 않은데, 예를 들어 8 차 정수계수 다항식

$$F(x) = x^8 – 2x^7 + 2x^6 – 2x^5 + 7x^4 – 10x^3 + 8x^2 – 4x + 1$$

---

[14] Google DeepMind 공식 블로그: https://deepmind.google/blog/alphaevolve-a-gemini-powered-coding-agent-for-designing-advanced-algorithms/

[15] arXiv: https://arxiv.org/abs/2511.07721

에 대해서 이 다항식의 모든 근을 포함하는 분해체(splitting field) $K$에 대해, $\mathrm{Gal}(K/\mathbb{Q})$가 크기 8 인 5 종류의 군 중 어느 것인지를 손으로 판별하는 것은 결코 쉬운 문제가 아닙니다.[16]

이 글의 주제를 생각해보면 자연스럽게 이런 질문을 해볼 수 있습니다. AI 를 이용해서 갈루아 군을 계산하는 것이 가능할까요? LLM 에게 직접 물어보는 방법도 있지만, 의사결정트리(decision tree)와 같은 고전적인 기계학습의 관점에서도 접근할 수 있습니다. 예를 들어 차수가 9 인 유리수체의 정규(normal) 확장체의 갈루아 군은 크기가 9 인 군인데, 크기가 9 인 군은 순환군 $C_9$와 두 순환군의 곱 $C_3 \times C_3$ 두가지뿐이므로 이를 이진 분류 문제(binary classification)로 볼 수 있습니다. 이때 입력 특성(feature)으로는 다항식의 계수를 사용할 수도 있지만, $K$의 데데킨드 제타함수(Dedekind zeta function) $\zeta_K(s) = \sum_{n \ge 1} a_n(K) / n^s$의 계수 $a_n(K)$를 사용하는 것이 정수론을 공부하는 사람들에게는 더 자연스럽게 느껴집니다. 여기서 $a_n(K)$는 $K$의 정수환 $\mathcal{O}_K$에서 크기가 $n$인 아이디얼의 개수이며, 실제 실험에서는 고정된 $N$이하의 $n$에 대한 $a_n(K)$만을 입력으로 사용했습니다.

He, 이규환, Oliver 의 선행 연구 Machine Learning Number Fields[17]에서 이미 이러한 설정 하에 고전 기계학습 알고리즘으로 갈루아 군을 포함한 수체의 다양한 불변량을 상당히 높은 정확도로 예측할 수 있음을 보였습니다. 하지만 이러한 알고리즘들이 높은 예측 정확도를 보여야 하는 원인에 대해서는 알려져 있지 않았는데 , 저와 이규환 교수님의 공동 연구[18]에서 이 이유를 좀 더 명확히 밝힐 수 있었습니다. 이 논문에서는 차수가 4, 6, 8, 9, 10 인 정규 확장체에 대해서 실험을 했는데, 특히 차수 9 인 경우, 의사결정 트리(decision tree)가 $n \le 1000$에 대해 $a_n(K)$ 를 이용해 정확도 100%를 달성했습니다[19]. 트리의 훈련에서 아무런 제한을 두지 않았기 때문에 1000 개의 입력을 사용해서 훈련된 트리는 구조가 매우 복잡할 수 있지만, 실제 트리는 놀라울 만큼 단순했습니다.

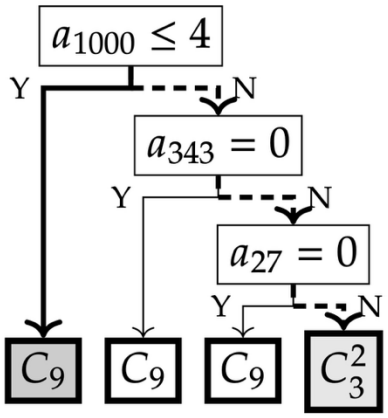
---

[16] 정답은 크기가 8 인 정이면체군 (dihedral group) $D_4$이고, 여기에서 찾아볼 수 있습니다. https://www.lmfdb.org/NumberField/8.0.2985984.1.

[17] arXiv: https://arxiv.org/abs/2011.08958

[18] arXiv: https://arxiv.org/abs/2508.06670

[19] 여기서 데이터셋은 정수론 에서 등장하는 각종 오브젝트들의 데이터를 모아놓은 LMFDB (L-function and Modular Form Database, https://www.lmfdb.org/)를 활용하였고, 총 1266 개의 차수 9 인 정규확장체가 있었습니다




생각보다 너무 간단하지 않나요? 갈루아 군을 계산하는 것 자체가 쉽지 않은 문제라는 것을 고려했을 때 (처음에는) 전혀 예상하지 못한 결과였습니다. 위에서 볼 수 있듯이 이 트리는 1000 개의 주어진 특성 중에서 $a_{27}(K), a_{343}(K), a_{1000}(K)$ 의 3 개만으로 100%의 정확도를 달성할 수 있다는 것 입니다. 더 놀라운 점은, 27, 343, 1000 이 모두 세 제곱수라는 점 입니다. 특히, 중간의 두 노드(node)를 봤을 때, $a_{27}(K)$나 $a_{343}(K)$가 0 이면 갈루아 군이 순환군 $C_9$라고 예측하고 있음을 알 수 있는데, 여기로부터 (다소 성급하게) 생각해 볼 수 있는 추측은 다음과 같습니다.

*(Conjecture) $K$가 차수가 9 인 유리수체의 갈루아 확장체라고 하자. 어떤 소수 $p$에 대해, $a_{p^3}(K) = 0$이라면 $\mathrm{Gal}(K/\mathbb{Q}) \simeq C_9$이다.*

이 추측은 참이며, 더 나아가 $a_n(K)$가 곱셈적(multiplicative)라는 것으로부터 소수가 아닌 일반적인 세제곱수에 대해서도 같은 주장을 할 수 있고, 심지어 위의 조건이 충분조건일 뿐만 아니라 필요조건이라는 것 까지 보일 수 있습니다. 다음은 실제로 논문에서 증명한, 더 일반화된 정리입니다.

*(Corollary 3.3) $\ell$이 소수이고 $K$가 차수가 $\ell^2$인 유리수체의 갈루아 확장체라고 하자. 이 때 $\mathrm{Gal}(K / \mathbb{Q}) \simeq C_{\ell^2}$ 일 필요충분조건은 $a_{p^\ell}(K) = 0$인 소수 $p$가 존재한다는 것이다.*

이 정리의 증명은 대학원 1 학년 수준의 정수론으로 어렵지 않게 할 수 있으며, 차수 6, 8, 10 에 대해서도 비슷하지만 조금 더 복잡한 정리들을 증명할 수 있습니다. 특히, 저희의 증명으로부터 알 수 있는 것은 하나의 zeta coefficient $a_{n}(K)$이 전체 갈루아 군을 결정하는 경우가 많고, 이는 왜 의사결정트리가 좋은

성능을 보이는지에 대한 설명이 될 수 있습니다. 이 연구에서 강조하고 싶은 것은 "얼마나 어려운 정리를 증명했는가"가 아니라, "단순한 결정 트리 하나로부터 어떻게 새로운 수학적 추측을 발견하고 증명으로까지 이어갈 수 있었는가"입니다.

### 그런데...(to be continued)

지금까지 LLM 과 고전 기계학습 방법론이 수학 연구에 활용된 사례들을 살펴보았습니다. 그런데 LLM 이 항상 올바른 답을 내놓는 것은 아닙니다. 환각(hallucination) 문제는 점차 개선되고 있지만 완전히 사라지지는 않았고, 특히 AI 가 새로운 수학을 발견했거나 난제를 풀었다는 주장이 늘어나고 있는 요즘, 그 상당수가 수학적으로 전혀 말이 안 되는 "AI slop"인 경우도 많습니다. 그렇다면 이를 어떻게 해결할 수 있을까요?

다음 시리즈에서는 최근 화두가 되고 있는 형식화(formalization)와, 수학 연구에서 형식화가 AI 의 신뢰성 문제를 어떻게 보완할 수 있는지에 대해 알아보겠습니다.