AI 추론 능력을 둘러싼 학술 논쟁

Apple vs Lawsen

Apple의 "생각의 환상" 논문과 Alex Lawsen의 반박 논문 "생각의 환상의 환상"을 중심으로
AI 추론 평가의 진실을 조명합니다.

> *우리는 AI의 사고를 올바르게 평가하고 있는가?*

발표자: 인공지능 연구팀

2025년 6월 15일

목차

발표 내용

"우리는 AI의 사고를 올바르게 평가하고 있는가?"

서막: 두 개의 시선이 만나는 곳

학술적 논쟁의 시작

2025년 6월 초

Apple 연구팀의 ***"생각의 환상 (Illusion of Thinking)"*** 논문 발표

대규모 추론 모델(LRM)은 복잡도가 증가하면 성능이 급격히 하락하며 '생각'을 중단한다고 주장

2025년 6월 10일

Alex Lawsen과 Claude Opus 4의 ***"The Illusion of the Illusion of Thinking"*** 반박 논문 발표

실험 설계의 문제점을 지적하며 AI의 실제 추론 능력을 다르게 해석

현재진행형

AI 추론 능력의 본질과 평가 방법에 대한 학계 논쟁 확산

논쟁의 본질

AI가 과연 '생각'이라는 행위를 할 수 있는가라는 근본적 질문에 대한 철학적 대립

두 가지 시선

- Apple: 퍼즐 실험 결과 AI는 복잡도 증가에 따라 추론 성능이 완전히 붕괴
- Lawsen: AI의 실패는 추론 능력의 한계가 아닌 출력 형식과 평가 방법의 문제

핵심 질문

"우리는 AI의 추론 능력을 어떻게 정의하고 측정해야 하는가?"

"이는 단순한 학술 논쟁을 넘어, AI 기술의 미래와 한계를 이해하는 근본적 질문이다."

Apple의 "생각의 환상" 논문 개요

2025년 6월 발표

저자 및 소속

주요 연구진:

Parshin Shojaee · Maxwell Horton · Iman Mirzadeh · Samy Bengio · Keivan Alizadeh · Mehrdad Farajtabar

Apple 인공지능 연구팀

기존 평가 방식의 한계

- 수학, 코딩 벤치마크 중심 평가
- 데이터 오염(contamination) 문제 발생
- 최종 답변 정확도만 평가
- 내부 추론 과정의 품질 파악 불가
- 복잡도 조절의 어려움

연구 목적

- 대규모 추론 모델(LRMs)의 근본적 능력 검증
- 복잡도 증가에 따른 확장 특성 분석
- 현재 평가 방식의 한계 보완
- 추론 과정 자체의 품질 평가
- 일반 LLM과 LRM의 같은 조건 비교

실험 설계: 제어 가능한 퍼즐 환경

핵심 특징:

- 복잡도 정밀 조절 가능
- 일관된 논리 구조 유지
- 데이터 오염 없음
- 명시적 규칙만 필요

평가 방법:

- 최종 답변 정확도
- 내부 추론 과정 분석
- 시뮬레이터 기반 검증
- 단계별 사고 검증

연구에 사용된 4가지 퍼즐 환경

하노이의 탑
복잡도: 디스크 수 (N)

체커 점프
복잡도: 체커 수 (N)

강 건너기
복잡도: 건너는 인원 (N)

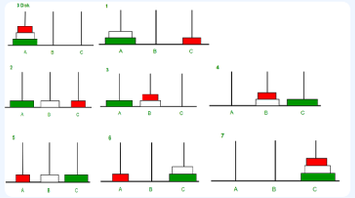
블록 쌓기
복잡도: 블록 수 (N)

Apple 논문의 실험 환경과 방법론

퍼즐 환경과 복잡도 조절

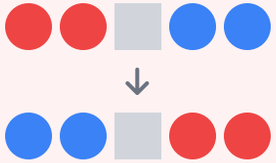
하노이의 탑 (Tower of Hanoi)

- 3개의 기둥과 크기 순서로 쌓인 n개의 원판
- 한 번에 하나의 원판만 이동 가능
- 작은 원판 위에 큰 원판을 올릴 수 없음
- **복잡도 조절:** 원판 수 n을 늘릴수록 증가
- **최소 이동 횟수:** $2^n - 1$



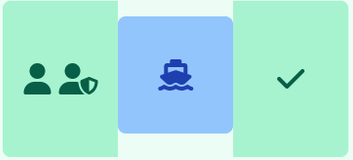
체커 점프 (Checker Jumping)

- 1차원으로 배열된 빨간색/파란색 체커와 빈 공간
- 인접 빈칸으로 이동 또는 반대색 체커를 건너뛰기 가능
- 체커는 후진할 수 없음
- **복잡도 조절:** 체커 개수 n을 증가
- **최소 이동 횟수:** $(n+1)^2 - 1$ (2n개 체커)



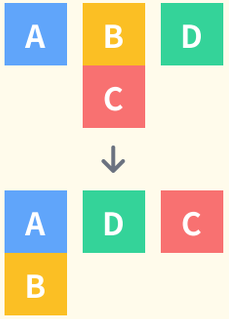
강 건너기 (River Crossing)

- n명의 액터와 n명의 에이전트가 보트로 강을 건너기
- 보트는 최대 k명 탑승 가능, 빈 보트는 운행 불가
- 에이전트 없이 액터는 타인의 에이전트와 동석 불가
- **복잡도 조절:** 액터/에이전트 쌍의 수 n
- **보트 수용량:** n=2, 3인 경우 k=2, n≥4인 경우 k=3



블록 쌓기 (Blocks World)

- 초기 배열에서 목표 배열로 블록을 재배치
- 한 번에 하나의 블록만 이동 가능
- 제약: 스택의 최상단 블록만 움직일 수 있음
- **복잡도 조절:** 블록의 수 n 증가
- **문제 유형:** PSPACE 계산 복잡도 클래스



복잡도 조절의 핵심

Apple 연구팀은 각 퍼즐의 요소 수(n)를 조절하며 복잡도를 체계적으로 증가시켰습니다. 이를 통해 일관된 논리 구조를 유지하면서도 모델의 추론 능력을 단계적으로 평가할 수 있었습니다.

Apple 논문의 주요 결과: 3단계 성능 구간

복잡도에 따른 성능 변화

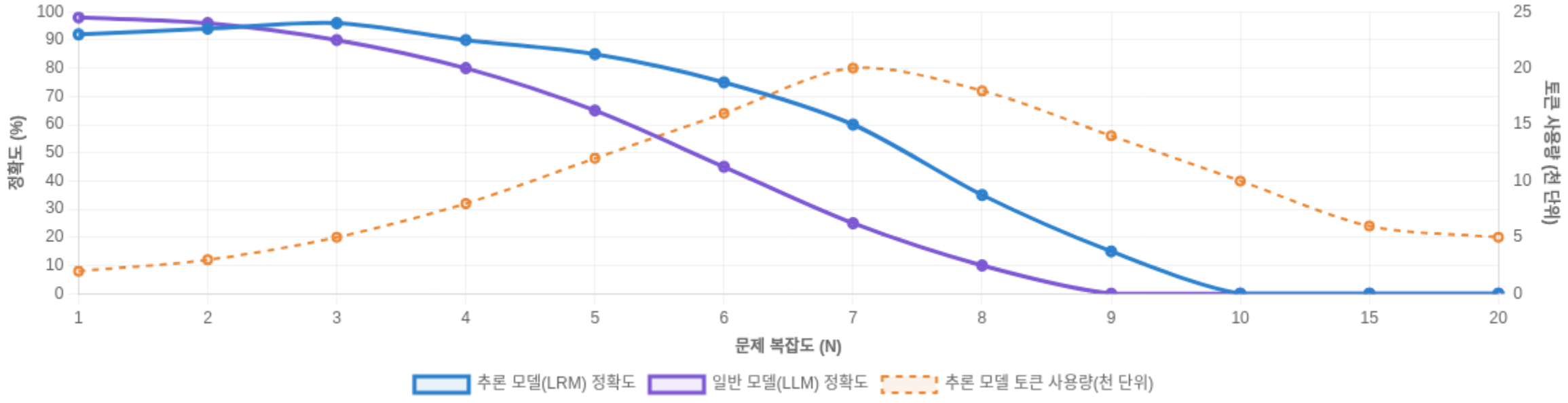


그림 1: 퍼즐 복잡도 증가에 따른 LRM과 일반 LLM 성능 및 추론 토큰 사용량 변화

구간 1: 낮은 복잡도

- 일반 LLM이 오히려 더 우수한 성능
- 추론 모델은 불필요한 사고 단계 생성
- 간단한 문제에서는 '오버씽킹' 현상 발생
- 토큰 효율성 측면에서 일반 모델 우세

구간 2: 중간 복잡도

- 추론 모델(LRM)의 성능이 뚜렷이 앞서기 시작
- 사고 과정이 문제 해결에 결정적 기여
- 복잡도 증가에 비례해 토큰 사용량 증가
- 자기 수정 메커니즘이 가장 효과적으로 작동

구간 3: 높은 복잡도

- 두 모델 모두 성능이 완전히 붕괴
- 역설적으로 사고 토큰 사용량 감소
- 모델이 문제 해결을 '포기'하는 현상 관찰
- 복잡도 임계점 이상에서 0%의 정확도

'사고 포기' 현상

문제가 어려워질수록 모델이 투입하는 사고 노력이 **줄어드는** 역설적 현상이 관찰됨. 마치 어려운 문제 앞에서 포기하는 학생처럼, AI도 한계에 부딪히면 사고를 중단.

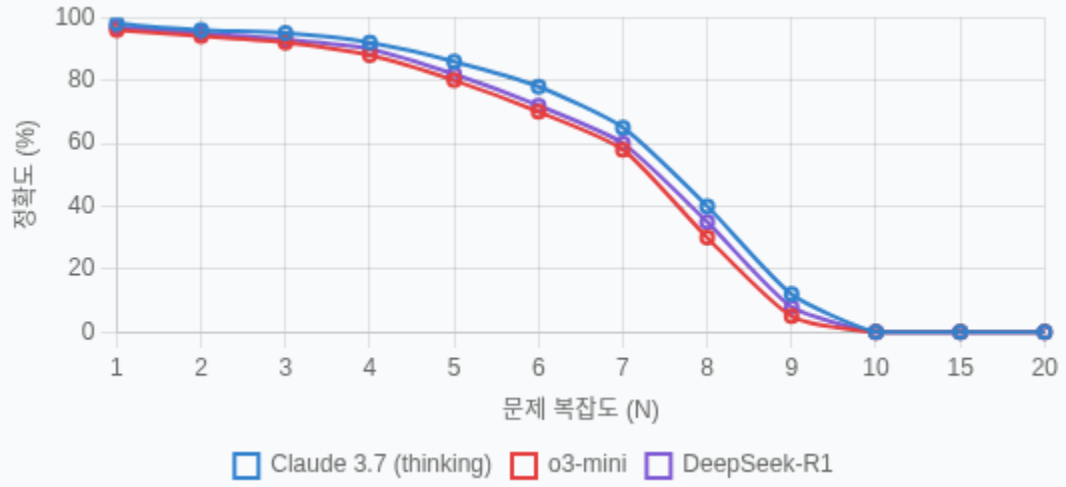
Apple의 해석

이런 붕괴는 현재 LLM의 추론 능력에 근본적 한계가 있음을 시사. '생각'의 환상을 만들어낼 뿐 진정한 일반화 가능한 추론 능력 부재.

Apple 논문의 그래프 및 데이터

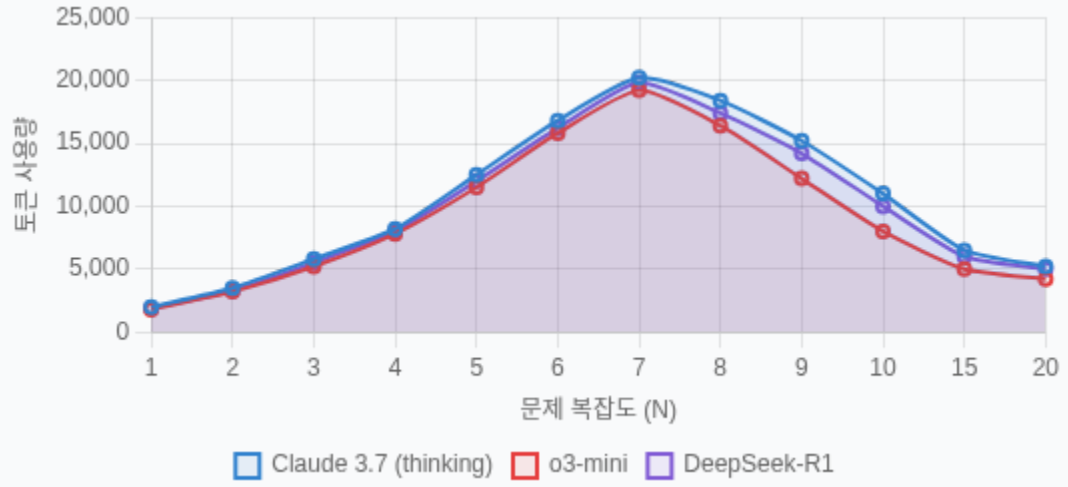
모델별 성능 및 토큰 사용량

Figure 4: 모델별 정확도



문제 복잡도 증가에 따른 주요 LRM 모델들의 정확도 변화

Figure 6: 토큰 사용량



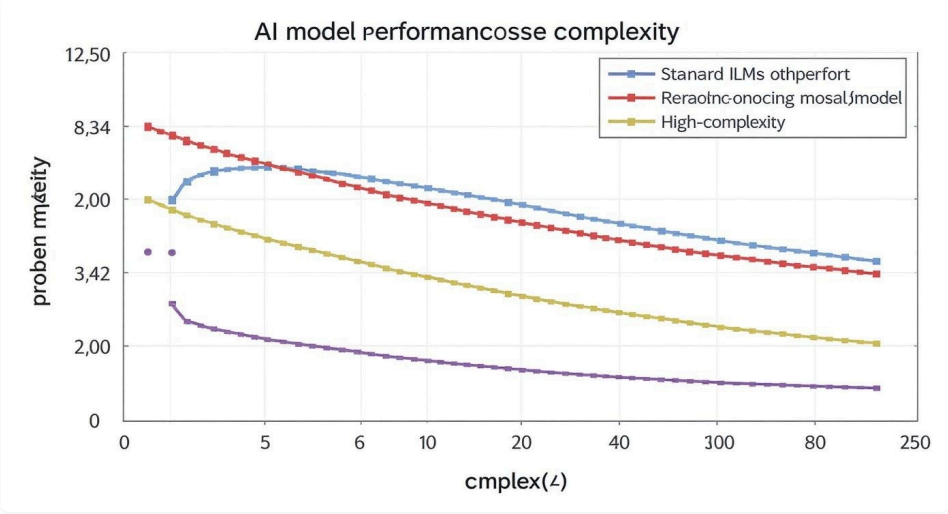
복잡도 증가에 따른 추론 토큰 사용량 변화

핵심 관찰

성능 붕괴 임계점
복잡도 N=8부터 모든 모델의 성능이 급격히 하락하고 N=10 이후 완전히 붕괴

역설적 토큰 사용 패턴
중간 복잡도까지는 토큰 사용량 증가, 이후 복잡도가 높아질수록 오히려 감소

모델 간 차이점
Claude 3.7이 임계점 직전에서 가장 오래 성능 유지, o3 시리즈는 급격한 토큰 감소 보임



Apple 논문의 결론

- 현재 대규모 추론 모델(LRM)은 일정 복잡도 이상에서 **추론 능력의 근본적 한계**를 보여줌
- 중요한 발견: 복잡도가 증가하면 모델이 투입하는 **사고 노력(thinking effort)이 감소**
- 이는 단순한 토큰 제한 문제가 아닌 **내재적 추론 능력의 한계**를 시사
- 테스트 환경의 통제된 특성으로 인해 결과의 신뢰성이 높다고 주장

출처: Shojaee et al. (2025). "The Illusion of Thinking: Understanding the Strengths and Limitations of Reasoning Models via the Lens of Problem Complexity"

Lawsen의 반박: 실험 설계 비판

⚠ 실험 방법론의 결함

1. 토큰 제한의 영향

모델의 출력 토큰 한계가 성능 측정에 직접적 영향을 미침

- 하노이의 탑 N=8부터 표준 출력 형식으로는 64K 토큰 초과
- 모델이 "이 답변은 너무 길어 여기서 중단합니다"라고 명시함
- 모델의 추론 능력 부재가 아닌 **물리적 출력 한계**

2. 불가능한 퍼즐 문제

일부 테스트 케이스는 수학적으로 해결 불가능한 문제

- 강 건너기 퍼즐은 N≥6 Actor/Agent 쌍에서 보트 용량 b=3으로는 해결 불가능
- Efimova (2018)의 수학적 증명에 따르면 이러한 제약 조건에서는 해가 존재하지 않음
- 불가능한 문제를 풀지 못한 것을 추론 실패로 잘못 분류

3. 경직된 평가 방법론

평가 프레임워크가 모델의 실제 이해도를 포착하지 못함

- 모든 이동을 일일이 열거해야만 정답으로 인정
- 알고리즘적 이해를 보여주는 압축된 표현 방식을 허용하지 않음
- 문자 단위 정확도 측정은 장문에서 통계적으로 오류 필연적

토큰 요구량 분석

하노이의 탑 N개 원반의 토큰 요구량:

$$T(N) \approx 5(2^N - 1)^2 + C$$

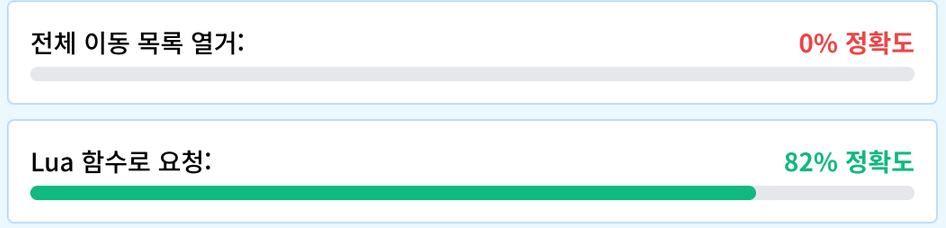
각 모델별 최대 해결 가능 크기:

- Claude-3.7, DeepSeek-R1 (64K 토큰): N ≈ 7-8
- o3-mini (100K 토큰): N ≈ 8

* Apple 논문에서 관찰된 성능 붕괴 지점과 정확히 일치

대안적 표현 방식의 영향

하노이의 탑 N=15에서 출력 형식 변경 시 성능:



❝ Lawsen의 결론

"이 실험이 보여주는 것은 모델이 주어진 토큰 한계보다 더 많은 토큰을 출력할 수 없다는 공학적 통찰이지, 추론 능력의 근본적 한계가 아니다."

핵심 질문은 "LRM이 추론할 수 있는가"가 아니라 "우리의 평가 방식이 추론과 타이핑을 구별할 수 있는가"이다.

Lawsen의 실험 결과: 출력 형식의 중요성

동일 문제, 다른 접근법

그림: 하노이의 탑 N=15 문제에서 출력 형식에 따른 정확도 비교 (Lawsen의 실험 결과)

Apple 논문 방식: 모든 단계 나열

215 − 1 = 32,767개의 개별 이동 단계를 모두 나열해야 함

```
1. 디스크 1을 기둥 1에서 기둥 3으로 이동
2. 디스크 2를 기둥 1에서 기둥 2로 이동
3. 디스크 1을 기둥 3에서 기둥 2로 이동
...
32,765. 디스크 2를 기둥 3에서 기둥 2로 이동
32,766. 디스크 1을 기둥 1에서 기둥 2로 이동
32,767. 디스크 1을 기둥 2에서 기둥 3으로 이동
```

→ 64K 토큰 제한 초과 (약 5 토큰/이동)

Lawsen 논문 방식: 알고리즘 코드 생성

재귀적 알고리즘을 Lua 함수로 표현 (~30 라인)

```
function hanoi(n, source, target, auxiliary)
    if n > 0 then
        -- 먼저 n-1개 원반을 source → auxiliary
        hanoi(n-1, source, auxiliary, target)

        -- n번째 원반을 source → target
        print("디스크 "..n.." 이동: "..
              source.." → "..target)

        -- n-1개 원반을 auxiliary → target
        hanoi(n-1, auxiliary, target, source)
    end
end

-- 메인 호출: 15개 디스크, 1번에서 3번 기둥으로
hanoi(15, 1, 3, 2)
```

→ 5,000 토큰 이내 완성 가능

Lawsen의 핵심 발견

- 모델들은 **알고리즘적 이해**를 갖추고 있으나, 출력 제약에 의해 성능이 제한됨
- 출력 형식 변경만으로 **동일한 15단계 하노이의 탑** 문제에서 정확도가 **0%에서 82%**로 급상승
- 모델들은 자신의 토큰 제한을 인식하여 "더 길어지는 것을 피하기 위해 여기서 멈추겠다"와 같은 메시지 출력
- 실험 평가 방식이 모델의 **알고리즘 이해 능력**이 아닌 **출력 생성 제약**을 측정

두 논문의 핵심 주장 구조도

상이한 철학적 접근

Apple: 엄격한 기준주의

완전한 추론 요구
진정한 추론은 어떤 제약 조건에서도 완벽한 답을 제시할 수 있어야 함

복잡도 붕괴 현상
복잡도가 증가하면 LRM의 성능이 급격히 저하되고 결국 완전히 붕괴

'사고 포기' 해석
높은 복잡도에서 모델이 문제 해결을 포기하고 추론 토큰이 감소하는 현상

거짓된 일반화
현재 LRM은 진정한 추론 능력이 아니라 '생각의 환상'만 제공

VS

Lawsen: 실용적 평가주의

토큰 제한의 영향
모델의 출력 토큰 제한이 성능 저하의 실제 원인 (N=8부터 64K 토큰 초과)

불가능한 퍼즐 문제
일부 평가(특히 N≥6 강 건너기)가 수학적으로 해결 불가능한 사례 포함

출력 형식의 중요성
하노이 탑 N=15도 Lua 함수 형태로 요청 시 정확도 0%→82% 상승

올바른 평가 방식
완전한 열거가 아닌 알고리즘적 이해를 평가해야 모델의 진정한 능력 측정 가능

> "현재 LRM은 복잡도가 증가하면 추론 능력이 완전히 실패한다"

> "실험 설계의 제약이 성능 붕괴의 실제 원인이다"

핵심 쟁점: 평가의 철학

수학 시험 비유

"모든 계산 과정을 보이시오"

프로그래밍 비유

"핵심 알고리즘만 명확하면 된다"

토큰 제한의 영향 분석

하노이의 탑 토큰 소모량

하노이의 탑 토큰 요구량

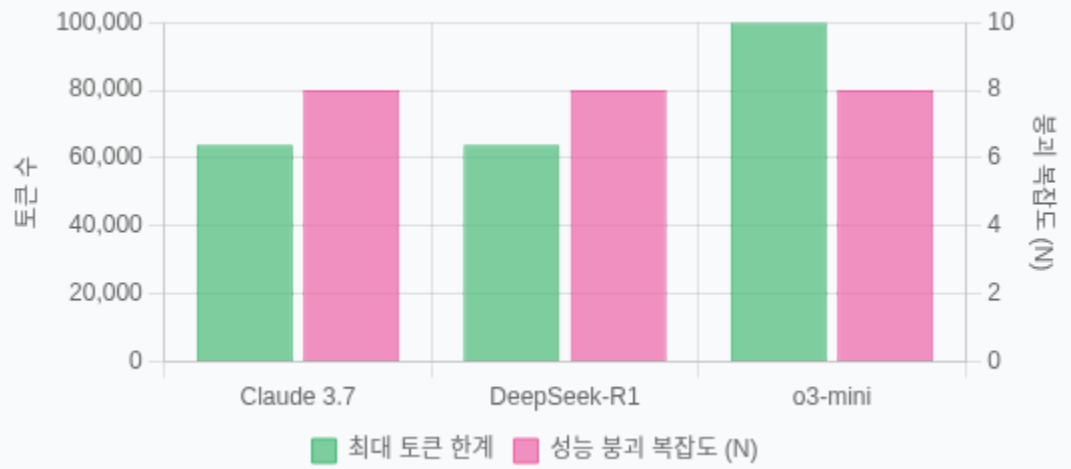
$$T(N) \approx 5(2^N - 1)^2 + C$$

여기서:

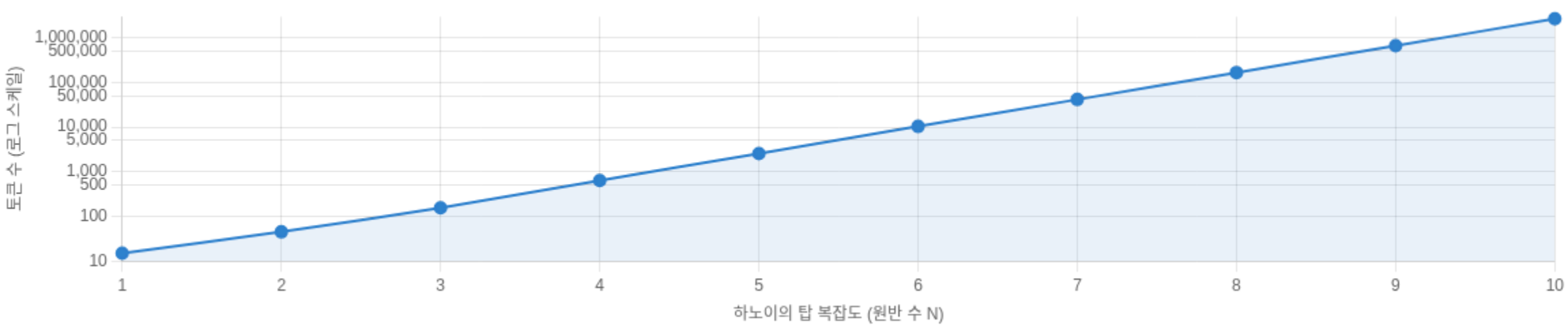
- **T(N)**: N개 원반을 해결하는 데 필요한 총 토큰 수
- **5**: 각 이동당 평균 토큰 수 (경험적 추정치)
- **$2^N - 1$**: 하노이의 탑 최소 이동 횟수
- **C**: 고정 오버헤드 (문제 설명, 초기화 등)

예시: N=10인 경우
10개 원반 풀이에 필요한 이동 횟수: $2^{10}-1 = 1,023$회
예상 토큰 수: $5 \times 1,023^2 + C \approx 5,231,445 + C$

모델별 토큰 한계와 성능 붕괴



각 모델의 최대 토큰 한계와 실제 성능 붕괴 복잡도 비교



하노이의 탑 문제 복잡도(N)에 따른 필요 토큰 수의 기하급수적 증가

핵심 결론

토큰 한계 지점 ≈ 성능 붕괴 지점: Apple 논문에서 관찰된 성능 붕괴는 모델의 최대 토큰 한계와 정확히 일치

최대 가능 복잡도 추정: $N_{max} \approx \lfloor \log_2(\sqrt{(L_{max}/5)}) \rfloor$ 공식으로 계산 가능

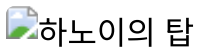
출처: Lawsen & Opus (2025). "The Illusion of the Illusion of Thinking: A Comment on Shojaee et al. (2025)"

퍼즐 난이도와 AI 추론의 본질

문제 길이 vs 탐색 복잡도

복잡도 측정의 두 가지 관점

Apple과 Lawsen의 논쟁에서 핵심 쟁점 중 하나는 AI 추론 평가에 적합한 '복잡도' 개념입니다. 단순히 해결 단계 수를 측정하는 것과 실제 계산적 복잡성은 다른 차원의 문제입니다.

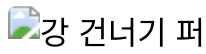


하노이의 탑
O(1) 결정 프로세스, 지수적 이동 수

해결 단계 수
N개 원반: 2^N-1 단계 필요 (N=15: 32,767단계)

탐색 복잡도
단순 재귀 알고리즘으로 해결 가능, 각 단계에서 결정 과정은 자명함

알고리즘 표현
~10줄의 재귀 함수로 표현 가능, 패턴 발견이 용이함



강 건너기 (선교사-식인종 퍼즐)
NP-hard 문제, 복잡한 상태 공간 탐색 요구

해결 단계 수
N=3 쌍: 약 11단계 (하노이의 탑보다 훨씬 적음)

탐색 복잡도
복잡한 제약 조건 고려, 많은 후보 경로 탐색, 막다른 상태 인식 필요

알고리즘 표현
단순 재귀로는 해결 불가능, 상태 탐색 및 백트래킹 알고리즘 필요

Lawsen 논문의 핵심 통찰

퍼즐 유형	해결 단계 길이	분기 계수	탐색 요구 수준
하노이의 탑	2^N-1	1	No
강 건너기	~4N	>4	Yes (NP-hard)
블록 세계	~2N	$O(N^2)$	Yes (PSPACE)

이는 모델이 하노이의 탑에서 100+ 단계를 수행하면서도 5단계 강 건너기 문제에서 실패하는 이유를 설명합니다. 단순히 문제 길이가 아닌 **복잡한 제약 조건 처리와 상태 공간 탐색 능력**이 AI 추론의 진정한 시험대입니다.

출처: Lawsen & Opus (2025). "The Illusion of the Illusion of Thinking: A Comment on Shojaee et al. (2025)"

AI 평가 방법론에 대한 철학적 논쟁

형식적 완전성 vs 알고리즘적 이해

형식적 완전성

- 모든 문제 풀이 단계와 추론 과정을 명시적으로 표현해야 함
- 전체 추론 과정을 완벽하게 드러내는 것이 필수
- 바이먼 페인만처럼 완전한 증명 과정을 중시
- 결과보다 과정의 질적 평가에 중점

> *Apple의 관점: "진정한 추론 모델은 복잡도와 상관없이 전체 풀이 과정을 단계별로 완전하게 생성할 수 있어야 한다"*

VS

알고리즘적 이해

- 문제 해결에 필요한 핵심 알고리즘과 원리를 파악하는 것이 중요
- 모든 단계를 수작업으로 나열하지 않아도 이해가 가능
- 전문가들은 보통 압축된 방식으로 문제를 해결함
- 실용적 관점에서의 문제 해결 능력 평가

> *Lawsen의 관점: "효율적 문제 해결을 위한 알고리즘 이해 능력이 기계적 열거 과정보다 진정한 지능의 척도이다"*

인간 평가자의 역할과 딜레마

학문적 맥락의 영향

수학, 물리학 같은 분야에서는 형식적 증명이 중시되지만, 컴퓨터 과학에서는 알고리즘적 효율성과 추상화 능력이 더 중요할 수 있습니다.

평가 맥락의 중요성

동일한 AI 시스템도 평가 목적과 맥락에 따라 서로 다른 평가 기준이 적용될 필요가 있습니다.

인간도 복잡한 문제를 풀 때 모든 과정을 상세히 설명하지 않습니다.
AI의 사고를 평가할 때, 우리는 무엇을 보고 있으며, 무엇을 놓치고 있을까요?

논쟁의 의미: 한계를 드러낸 두 시선

시사점과 향후 과제

과도한 기대와 맹신 경계

Apple의 경고는 소중합니다. AI의 한계를 정확히 파악하는 것은 과도한 기대와 맹신을 방지하는 첫걸음입니다. 현재의 AI가 보여주는 '사고'는 인간이 기대하는 수준과 여전히 거리가 있습니다.

평가 방법론의 재고

Lawsen의 지적대로 평가 방법의 한계가 모델의 한계로 오해될 위험이 있습니다. 평가 방식이 기술의 진정한 가능성을 제한하는 인위적 장벽이 될 수 있음을 인식해야 합니다.

실험 설계의 중요성

모델의 물리적 한계(토큰 제한)와 문제의 수학적 성질(불가능성)을 고려한 실험 설계가 필수적입니다. 실험이 의도하지 않은 제약을 측정하고 있지는 않은지 항상 검증해야 합니다.

질문의 재정의

"AI가 생각할 수 있는가?"라는 물음 대신, "우리는 AI의 사고를 올바르게 측정하고 있는가?"라는 새로운 화두가 더 생산적일 수 있습니다.

향후 연구 방향

1. **추론 능력과 출력 제약 분리:** 모델의 내부 추론 능력과 출력 형식의 제약을 명확히 분리하는 평가 방법론 개발
2. **창의적 대안 탐색:** 토큰 한계를 고려한 압축된 표현 방식과 알고리즘적 이해 측정 방법 연구
3. **복잡성 메트릭 개선:** 단순 길이가 아닌 문제의 계산 복잡성을 더 정확히 반영하는 평가 지표 개발
4. **인간-AI 비교 방법론:** 동일한 제약 조건에서 인간과 AI의 추론 능력을 공정하게 비교할 수 있는 방법 설계

> 어쩌면 진정한 '환상'은 완벽한 평가 방법이 존재한다고 믿는 우리의 착각일지도 모릅니다. 중요한 것은 어느 한쪽의 시각이 절대적으로 옳다고 단정하는 것이 아니라, 서로 다른 관점들이 서로를 보완하며 진실에 한 걸음씩 다가가는 과정 자체일 것입니다.
>
> *- 인공지능 평가 철학에 관한 새로운 시각*

에필로그: AI와 인간, 사고라는 미궁 속으로

> **우리는 AI의 사고를 올바르게 평가하고 있는가?**
>
> 완벽한 평가 방법이 존재한다고 믿는 것 자체가 우리의 착각인지도 모릅니다

풀리지 않는 미스터리

애플과 로슨의 논쟁이 보여주듯, 같은 현상도 바라보는 시각에 따라 전혀 다른 의미를 갖습니다. 추론 능력의 실패인가, 평가 방법의 한계인가?

순환하는 질문들

AI가 "생각"한다는 것의 의미는 무엇인가? 주어진 작업을 정확히 수행하는 것인가, 인간처럼 모든 맥락을 이해하는 것인가?

진정한 통찰을 찾아서

어느 한쪽의 시각이 절대적으로 옳다고 단정하기보다, 서로 다른 관점들이 서로를 보완하며 진실에 한 걸음씩 다가가는 과정 자체가 중요합니다. AI 추론 능력의 진정한 한계와 가능성을 이해하기 위해서는 평가 방법론의 철학적 기반에 대한 더 깊은 성찰이 필요합니다.

다가올 미래를 위한 질문

- 인간의 사고와 AI의 사고는 근본적으로 다른 것인가, 아니면 같은 스펙트럼의 다른 지점인가?
- 우리는 무엇을 "사고"라고 정의하며, 그 정의가 AI 평가에 어떤 영향을 미치는가?

그렇게 AI와 인간은, 각자의 방식으로, 생각이라는 미궁 속에서 길을 찾아갑니다.