The Clean Paper

WHAT THE PAPER SAYS · WHAT IT DOESN'T · WHY IT MATTERS

언어 모델은 왜 환각하는가—그리고 우리의 채점 방식은 왜 그 행동을 계속 부추기는가

우리가'환각'이라고 부르는 자신만만한 거짓말은 수수께끼 같은 고장이 아니다. 일부는 훈련 과정에서 생기는 통계적 부산물이고, 주류 벤치마크가 정직한'모르겠습니다'보다 자신 있는 추측을 더 보상하기 때문에 계속 남는다. 네 프런티어 모델을 대상으로 한 사례 연구에서는 프롬프트 안에 채점 규칙을 명시하는'오픈 루브릭'이 이 인센티브를 뒤집었다.

Written by Lucio Vaglio · Cross-checked by Laura Nesso · Edited by Michele Renda · Figures by Laura Nesso and Nova Vela · AI-assisted, human-reviewed

Paper: *Evaluating large language models for accuracy incentivizes hallucinations*, Adam Tauman Kalai, Ofir Nachum, Santosh S. Vempala, Edwin Zhang, Nature 653, 1047–1050 (2026) · DOI: 10.1038/s41586-026-10549-w

모델은 왜 추측하는가, 그리고 우리는 왜 그렇게 하도록 가르쳤는가

대규모 언어 모델에게 낯선 사람의 생일을 물어보면, 마치 카드에서 그대로 읽는 듯 흔들림 없이 "3 월 7 일"이라고 답할 수 있다. 그리고 세 번 물으면 세 번 모두 틀리면서 날짜도 각각 다르게 말할 수 있다. 저자들은 바로 이런 사례를 든다. 최첨단 모델들에게 한 사람의 생일이나 생소한 약어의 뜻처럼 단순한 사실 질문을 했더니, 모델마다 서로 다른 답을 자신 있게 지어냈고 모두 틀렸다. 업계에서는 이를 환각 *(hallucination)*이라고 부른다. 마치 지각 과정에 이상이 생긴 것처럼 들리는 말이다. 이 논문의 첫 번째 일은 여기서 신비감을 걷어내는 것이다.

우선 모델이 어떻게 만들어지는지부터 보자. 가장 처음이자 가장 큰 훈련 단계에서 모델은 방대한 텍스트를 읽으며, 사실상 유창한 언어가 어떤 모습인지 배운다. 이제 그 뒤에 아무런 패턴이 없는 사실 하나를 생각해보자. 특정 한 사람의 생일 같은 것이다. 그 날짜가 훈련 텍스트에 한 번만 등장했거나 아예 등장하지 않았다면, 패턴을 학습하는 모델이 붙잡을 만한 것이 없다. 모델 입장에서는 답이 임의적이다. 저자들은 다른 문제를 다루던 앨런 튜링의 오래된 아이디어를 빌려 이를 정밀하게 표현한다. 데이터에 등장한 생일 가운데 5 분의 1 이 딱 한 번만 나타난다면, 모델은 적어도 5 분의 1 정도를 틀릴 것으로 예상해야 한다. 모델이 망가졌기 때문이 아니라, 애초에 학습할 것이 없었기 때문이다. 반대로 모델이 나라의 수도를 거의 틀리지 않는 것도 같은 논리로 설명된다. 수도 이름은 데이터에 끊임없이 반복해서 나타난다. 저자들은 참인 문장과 그럴듯하지만 거짓인 문장을 구별하는 것 자체가 어려운 문제이며, 오직 참인 문장만 생성하는 일은 적어도 그만큼 어렵다고 신중하게 논증한다. 일정 수준의 오류 하한은 처음부터 내장되어 있다.

밑바탕의 아이디어: 아직 보지 못한 것을 어떻게 셀까

여기에는 정말 영리한 아이디어가 깔려 있다. 언어 모델보다 훨씬 오래된 개념이고, 제대로 알아둘 만하다.

색깔 공이 든 자루부터 시작하자. 자루 안에 몇 가지 색이 들어 있는지는 모른다. 공을 하나씩 100 개 뽑아 세어보니 빨강 40 개, 파랑 25 개, 초록 15 개, 노랑 5 개, 보라 3 개, 주황 2 개가 나왔다. 그리고 **서로 다른 열 가지 색이 각각 딱 한 번씩** 나왔다.

튜링이 실제로 마주했던 질문은 전혀 다른 문제에서 나온 것이지만, 질문의 형태는 이렇다. 다음에 뽑는 공이 지금까지 한 번도 보지 못한 색일 확률은 얼마일까? 아직 뽑아보지 않은 것은 직접 셀 수 없다. 하지만 **딱 한 번만 나온 색**, 즉 "싱글턴 (singleton)" 은 셀 수 있다. **Good–Turing 추정**이라 불리는 방법의 핵심은, 표본에서 싱글턴이 차지하는 비율로 아직 관측하지 못한 색들에 숨어 있는 확률 질량을 추정하는 것이다. 100 번 가운데 10 번이 한 번만 나온 색이었다면, 다음 공이 완전히 새로운 색일 확률은 대략 **10 / 100 = 10%** 다.

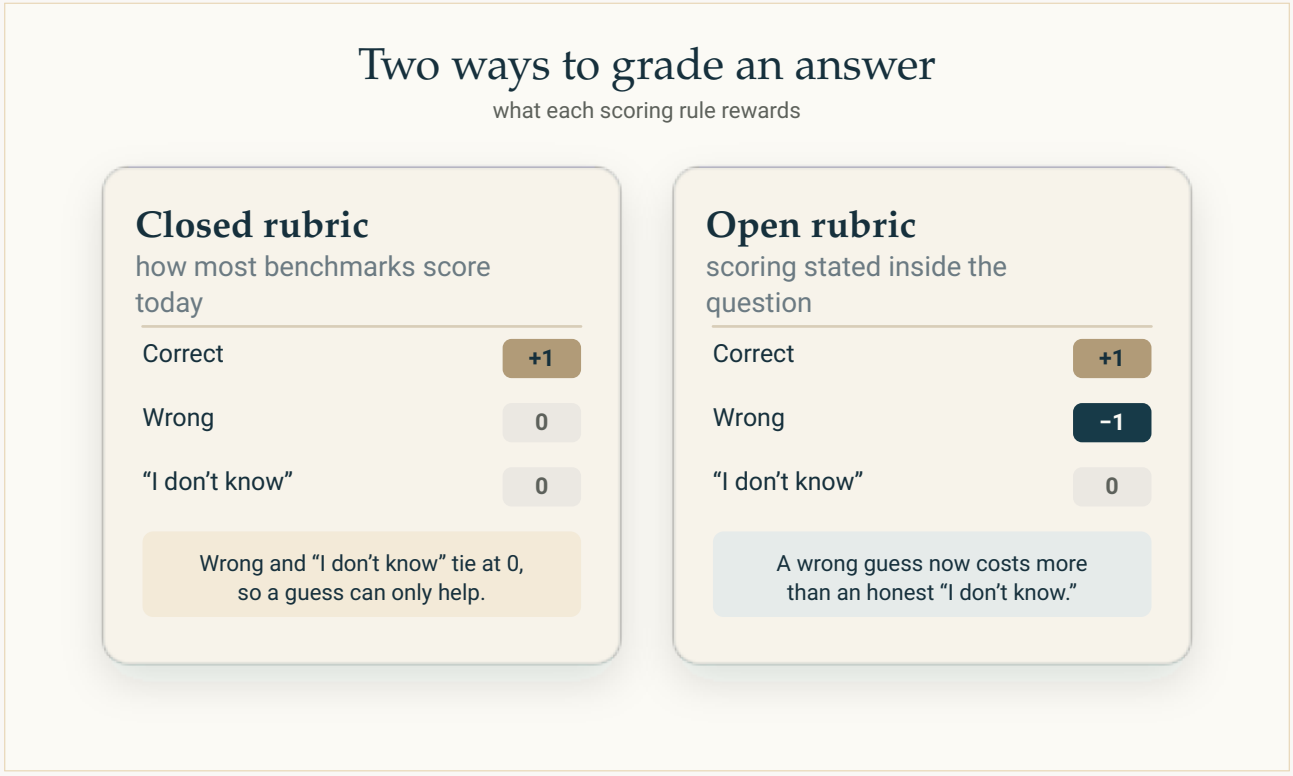
한 번만 나온 색들은 실수가 아니다. 그것들은 우리가 얼마나 모르는지를 측정해준다. 표본에 딱 한 번만 나온 색이 많다는 사실은, 세상에는 아직 우리가 뽑아보지 못한 색이 더 많이 있다는 신호다.

이제 색을 생일로 바꾸고, 자루를 모델의 훈련 텍스트로 바꿔보자. 모델이 본 생일 가운데 **5 분의 1 이 정확히 한 번만 등장했다고** 하자. 같은 논리를 적용하면, 전체 확률의 약 5 분의 1 이 모델이 사실상 한 번도 충분히 보지 못한 생일들에 걸려 있다. 생일에는 대신 의지할 패턴도 없다. 누군가의 생일을 논리적으로 계산해낼 수는 없기 때문이다. 한 번 본 날짜나 한 번도 보지 못한 날짜는 모델이 가중치를 제대로 둘 근거가 없는 동전과 비슷하며, 대략 **5 개 중 1 개**는 틀리게 된다. 아무리 영리해도 이 문제를 없앨 수 없다. 배울 정보가 애초에 없었기 때문이다.

이것이 전체 논증을 작은 모형으로 압축한 모습이다. 싱글턴 비율은 이 데이터로부터는 학습할 수

> 없는 세계의 몫을 측정하고, 그것이 오류의 **하한선**이 된다. 모델이 수도 이름을 거의 틀리지 않는 이유도 같다. *Paris* 같은 말은 반복해서 등장하므로 싱글턴 비율이 거의 0 이고, 배울 정보가 충분하다.

이것은 환각이 어디에서 생기는지는 설명한다. 하지만 왜 환각이 남아 있는지는 설명하지 못한다. 이후의 훈련 단계에서 모델을 더 유용하고 정직하게 만들려고 했는데도, 왜 모델은 여전히 모른다고 말하기보다 허세를 부릴까? 이 지점에서 논문의 비유는 조금 불편할 만큼 정확하다. 시험을 보는 학생이 답을 모른다고 생각해보자. 빈칸은 0 점이고 찍어서 맞히면 1 점을 받을 수 있다면, 점수를 최대화하는 전략은 추측하는 것이다. 자신 있게, 구체적으로, "잘 모르겠습니다" 라고 말하지 않는 쪽이 유리하다. 학생은 이를 배운다. 그리고 모델도 마찬가지다. 우리가 같은 방식으로 채점하기 때문이다. 저자들은 실제로 연구계가 경쟁하는 벤치마크, 즉 모델의 성능 순위를 매기는 평가표들을 살펴봤다. 거의 모두가 "모르겠습니다" 에 틀린 답과 정확히 같은 점수, 즉 0 점을 준다는 사실을 발견했다. 이 규칙 아래에서는 항상 추측하는 모델이, 불확실할 때 정직하게 표시하는 것만 빼면 완전히 같은 모델보다 높은 점수를 받는다. 말 그대로 우리는 채점 방식으로 모델이 추측하도록 밀어붙이고 있다.



벤치마크는 추측을 합리적인 전략으로 만들 수 있다. 대부분의 벤치마크가 쓰는 채점 방식 (왼쪽) 에서는 틀린 답과 정직한 "모르겠습니다" 가 모두 0 점이므로, 추측은 손해가 없고 맞으면 이득이다. 반대로 질문 안에 채점 규칙을 명시하고 틀린 답에 벌점을 주면 (오른쪽), 확신이 없을 때 답을 보류하는 것이 더 나은 전략이 될 수 있다. 이것은 시험이 무엇을 보상하는지를 바꾸는 것이지, 그 자체로 환각 문제를 해결하는 것은 아니다.

이 부분은 특히 기억해둘 만하다. 흔히 접하는 헤드라인과 정반대 방향이기 때문이다. 환각은 종종 기술의 피할 수 없는, 거의 신비적인 한계처럼 이야기된다. 이 논문은 두 주장 모두에 이의를 제기한다. 사전학습 단계의 오류 하한은 미스터리가 아니다. 머신러닝이 수십 년 전부터 다뤄온 평범한 통계적 오류다. 그리고 환각이 계속 남는 것도 필연은 아니다. 적어도 일부는 우리가 만들어놓은 인센티브이며, 바꿀 수 있다. 확

신이 없을 때 단순히 답을 거부하는 시스템이라면 환각하지 않을 수 있다. 실제 배포된 모델들이 그렇게 행동하지 않는 이유 중 하나는 우리의 평가 방식이 답변 거부를 벌주기 때문이다.

연구진이 한 일

논문은 세 부분으로 구성된다. 첫째, 사전학습 동안 어느 정도의 환각이 통계적으로 강제된다는 수학적 논증이다. "유효한 텍스트만 생성하기" 가 최소한 "이 문장이 유효한가?" 라는 이진 분류 문제만큼 어렵다는 점을 보인다. 둘째, 영향력 있는 벤치마크 10 개를 조사해 주류의 정확도 중심 지표가 답변 보류보다 추측을 보상한다는 논증이다. 셋째, 해결책을 제안하고 사례 연구로 시험한다. 바로 **오픈 루브릭 (open-rubric)** 평가다. 질문 안에 채점 규칙 자체를 적는다. 예를 들어 "정답은 1 점, 오답은 −1 점이므로 50% 이상 확신하지 못하면 답하지 말라" 고 명시한다. 그러면 모델은 언제 정직함이 보상받는지 알 수 있다. 연구진은 SimpleQA의 사실 질문 4,326 개를 이용해 Google 의 Gemini 3 Pro, OpenAI 의 GPT-5, xAI 의 Grok 4, Anthropic 의 Claude Opus 4.5 등 네 개의 프런티어 모델에 이를 적용했다. 저자들은 이 사례 연구가 설명용이며 "모델 간 통제된 평가가 아니다" 라고 명시한다. 기본 설정을 사용했고, 별도 튜닝이나 비용 정규화도 하지 않았다.

무엇을 발견했나

- **사전학습은 어느 정도의 오류를 강제한다.** 모델이 자신 있게 거짓 정보를 내놓는 비율에는 하한이 있으며, 대략 그 모델에서 만들 수 있는 최선의 "이 문장이 유효한가?" 분류기의 오류율 두 배와 연결된다. 학습 가능한 패턴이 없는 사실의 경우 이 하한은 적어도 싱글턴 비율, 즉 훈련 데이터에 정확히 한 번만 등장하는 사실의 비율만큼이다. 데이터가 완벽하게 깨끗해도 어느 정도의 환각은 피할 수 없다.
- **채점은 실제로 추측을 보상한다.** 일반적인 정답/오답 채점에서는 절대 답을 보류하지 않는 전략이 최적이며, 저자들이 조사한 인기 벤치마크의 압도적 다수는 "모르겠습니다" 를 그냥 오답으로 취급한다. 저자들 자신이 다루는 모델에서도 생생한 예가 나온다. SimpleQA 에서 원시 정확도만 보면 거의 모든 질문에 답하고 그중 4 분의 3 이상을 틀리는 OpenAI 의 o4-mini 가, 확신이 없을 때 답을 보류해 훨씬 적게 틀리는 GPT-5-mini 보다 조금 더 높은 점수를 받는다. 더 무모한 모델이 평가표에서는 더 좋아 보이는 셈이다.
- **오픈 루브릭은 인센티브를 뒤집는다 (이 사례 연구에서는).** 연구진은 간단한 환각 완화법을 시험한다. 모델에게 두 번 답하게 하고, 두 답이 다르면 답변을 보류하게 하는 방법이다. 표준 정확도에서는 이 방법이 오류를 줄이지만 정확도도 낮추기 때문에 지표가 도입을 억제한다. 반면 오픈 루브릭에서는 여러 벌점 수준에 걸쳐 네 모델 모두 같은 완화법을 쓰는 쪽이 더 높은 점수를 얻는다. 원시 정확도에서 확신이 없을 때 답을 보류한다는 이유로 불리했던 GPT-5-mini 도 채점 규칙을 공개하면 o4-mini 보다 앞선다 (모델당 n = 4,326 문항).

이것이 아마 뜻하는 것

환각을 줄이는 일은 주로 환각 전용 테스트를 더 많이 만드는 문제가 아닐 수 있다. 핵심은 주류 벤치마크가 불확실성을 채점하는 방식을 바꿔, “모르겠습니다” 라고 인정하는 행동이 더 이상 벌을 받지 않게 하는 것이다. 평가 방식이 바뀌지 않는 한 환각을 줄인 모델은 계속 정확도 점수를 잃고, 따라서 그런 행동은 계속 억제될 것이다. 저자들이 이 문제를 “사회기술적 (socio-technical)” 이라고 부르는 이유다. 더 나은 지표를 만드는 기술 문제인 동시에, 영향력 있는 평가 체계가 그것을 채택하도록 만드는 사회적 문제다.

이것이 증명하지 않는 것

- 오픈 루브릭이 실제 현장에서 환각을 해결한다는 것을 **보여주지 않는다**. 이를 뒷받침하는 실험은 의도적으로 **통제하지 않은** 작은 사례 연구다. 기본 설정의 네 모델, 선택된 완화법 하나, 사실 질의응답 테스트 하나를 사용했다. 목적은 모델 순위를 매기거나 일반적인 효과를 증명하는 것이 아니라 인센티브가 뒤집힌다는 점을 보여주는 것이다.
- 채점 방식이 유일한 원인이라고 **주장하지 않는다**. 훈련 데이터의 오류, 정말 어려운 문제, 익숙하지 않은 프롬프트 등은 여전히 별개의 원인이다.
- 환각이 필연이라는 흔한 주장도 **지지하지 않는다**. 오히려 저자들은 반대로 주장한다. 확인 가능한 질문에만 답하고 나머지는 “모르겠습니다” 라고 말하는 시스템이라면 환각하지 않을 수 있다는 것이다.
- 사전학습의 오류 하한을 **없애는 것도 아니다**. 그것을 설명하고 경계를 제시할 뿐이다. 또한 그 경계는 모든 모델 행동이 아니라 자신 있게 내놓는 사실 오류에 관한 것이다.
- 오픈 루브릭만으로 충분하다는 것을 **보여주지 않는다**. 오픈 루브릭은 평가가 무엇을 보상하는지를 바꾼다. 검색, 도구 사용, 더 잘 보정된 모델을 대신하지 않는다.

근거는 얼마나 탄탄한가?

- 핵심은 **수학**이다. 측정값이 아니라 형식적인 하한선이다. 이론적 논증으로서는 제시된 가정 안에서 타당하다.
- 문제를 **의도적으로 단순화한 모델**에 의존한다. 저자들도 모든 응답을 정답, 오답, “모르겠습니다” 셋으로만 나누는 것을 “거짓 삼분법 (false trichotomy)” 이라고 지적하며, 가장 명확한 하한을 얻기 위해 이상화된 “임의 사실 (arbitrary facts)” 설정을 사용한다.
- 벤치마크 조사는 **작게 선별된 표본**이다. 영향력 있는 평가 10 개를 다뤘지만 전수조사는 아니다.
- 사례 연구는 **실제이지만 제한적**이다. 프런티어 모델 네 개, 완화법 하나, SimpleQA 하나, 기본 설정만 사용했으며, 명시적으로 “통제된 평가가 아니다” 라고 한다. 이것은 인센티브 논증을 위한 개념증명이지 벤치마크 결과가 아니다.
- 연구자의 위치도 말해둘 가치가 있다. 저자 네 명 중 세 명은 **OpenAI** 에 재직 중이거나 재직한 적이 있고, 논문은 업계가 모델 평가 방식을 바꿔야 한다고 주장한다. 잘 논증된 입장이지만, 이해관계가 없는 외부 기관의 중립적 검토는 아니다. 무시할 이유는 아니지만 감안할 요소다. 다만 논문은 다른 회사 모델뿐 아니라 자사의 o4-mini 와 GPT-5-mini 에도 같은 비판을 적용한다.

왜 중요한가

과장되기 쉬운 문제의 틀을 다시 짠다. “환각” 은 으스스한 결함이나 넘을 수 없는 벽처럼 이야기되기 쉽다. 이 논문은 그것을 훨씬 평범하고, 일부는 우리가 스스로 만든 문제로 바꿔놓는다. 우리가 실제로 이해할 수 있는 통계적 하한 위에, 우리가 선택한 인센티브가 얹혀 있다는 것이다. 더 넓은 교훈은 조용하지만 더 유용하다. 신뢰성을 더 높이려면 무엇을 만드는가 못지않게 무엇을 측정하는가가 중요할 수 있다.

핵심 요약

언어 모델이 자신 있게 내놓는 거짓 답에는 두 가지 뿌리가 있다. 첫째는 통계적이다. 학습할 패턴이 없는 사실이라면 언어를 모방하도록 훈련된 모델은 때때로 틀릴 수밖에 없고, 그 하한은 예를 들어 훈련 데이터에 사실이 한 번만 등장하는 비율로 추정할 수 있다. 둘째는 인센티브다. 모델 순위를 매기는 거의 모든 벤치마크는 “모르겠습니다” 를 오답과 똑같이 채점하므로, 추측은 항상 유리하다. 심지어 4 분의 3 이상을 틀리는 모델이 더 정직하게 답을 보류하는 모델보다 높은 점수를 받을 수도 있다. 저자들의 제안은 새로운 환각 테스트가 아니라 “오픈 루브릭” 이다. 질문 안에 채점 규칙을 명시하는 것이다. 네 프런티어 모델을 대상으로 한 사례 연구에서는 이것이 인센티브를 뒤집어, 환각을 줄이는 방법이 벌점을 받는 대신 보상을 받게 했다. 이 논문은 이론과 벤치마크 조사에 작고 명시적으로 통제되지 않은 실험을 더한 연구이며, *Nature* 에서 동료평가를 받았다. 해결책은 유망하지만 대규모에서 효과가 입증된 것은 아니며, 저자들은 환각이 신비하지도, 엄밀한 의미에서 필연적이지도 않다고 주장한다.

과장 없이 보면

논문이 보여주는 것: 사전학습 중 어느 정도의 환각이 강제되는 수학적 하한 (패턴 없는 사실에서는 적어도 “싱글턴 비율”); 주요 벤치마크 대부분이 “모르겠습니다” 에 점수를 주지 않는다는 조사; 그리고 프롬프트 안에 채점 기준을 명시하는 “오픈 루브릭” 을 사용하면 보통 정확도 지표에서는 불리했던 환각 완화법이 네 모델 사례 연구에서 더 높은 점수를 받는다는 결과.

그럴듯하지만 입증되지 않은 것: 오픈 루브릭을 주류 벤치마크에 도입하면 실제 배포 모델의 환각이 의미 있게 줄어들 것이라는 주장. 이를 뒷받침하는 실험은 작고 명시적으로 통제되지 않았다.

보여주지 않는 것: 환각이 필연이라는 것 (논문은 반대로 주장한다), 채점이 유일한 원인이라는 것, 환각을 완전히 제거할 수 있다는 것, 사례 연구가 네 모델의 우열을 평가한다는 것.

주요 한계: 의도적으로 단순화한 정답/오답/“모르겠습니다” 모델 (저자들은 이를 “거짓 삼분법” 이라고 부른다), 작게 선별된 벤치마크 조사 (10 개 평가), 통제되지 않은 사례 연구 (네 모델, 완화법 하나, 테스트 하나, 기본 설정), 그리고 모델 평가 방식을 바꿔야 한다고 주장하는 OpenAI 주도 연구라는 점.

일반 독자는 어느 정도 확신해도 되는가? 환각이 신비한 현상도, 엄밀히 피할 수 없는 현상도 아니며, 현재 주류 벤치마크가 추측을 보상한다는 데에는 높은 확신을 가져도 된다. 제안된 해결책이 도움이 된다는 데에는 중간 정도의 확신이 적절하다. 실제 개념증명은 있지만, 광범위하고 대규모로 효과가 있다는 증거는 아직 없다.

출처

Nature 653, 1047–1050 (2026)
https://doi.org/10.1038/s41586-026-10549-w

Why Language Models Hallucinate (arXiv 2509.04664)
https://arxiv.org/abs/2509.04664

hallucinations-paper-experiments (OpenAI)
https://github.com/openai/hallucinations-paper-experiments

SimpleQA
https://github.com/openai/simple-evals