지표 붕괴, Goodhart's Law

점수는 오르는데, 능력은 오르지 않는다

Reviewed Papers

The Leaderboard Illusion

Line Goes Up?

Presenter: 한완규

주제 선언 — "평가를 믿어도 되는가?"

■ **이 주차의 목표:**

벤치마크와 리더보드 **그 자체를 의심하는 것**

× **다루지 않는 것:**

단순 모델 비교나 에이전트 설계 방법론

❓ **핵심 질문:**

"점수 상승이 실제 **능력 향상**을 의미하는가?"

두 논문의 역할 구분

The Leaderboard Illusion : 현상 (Phenomenon)

왜 leaderboard가 현실을 반영하지 못하는가

연결 고리: Goodhart’s Law

Line Goes Up? : 구조적 원인 (Root Cause)

왜 benchmark 점수 상승이 실제 능력 향상이 아닌가

"현상(Leaderboard) + 구조적 원인(Benchmark) = 평가의 위기"

Goodhart's Law: 정의와 적용

❝ "측정이 목표가 되면, 더 이상 좋은 측정이 아니다"
(When a measure becomes a target, it ceases to be a good measure)

📈 LLM 평가의 현주소: Accuracy ↑ Leaderboard 순위 ↑

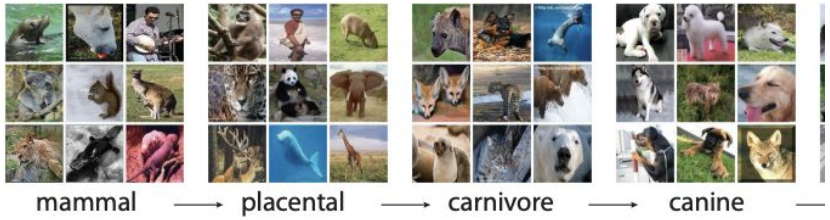
❓ 그러나 실제 문제 해결 능력(Real-world problem solving)은 미지수

⚠️ 결과: 지표 자체가 학습의 목표(Target)가 되면서 평가 시스템 붕괴

ML 역사: 같은 실수의 반복

컴퓨터 비전 (ImageNet 시대)

- ImageNet Top-1 accuracy는 상승했으나, 배경·워터마크 등 바이어스 학습
- 결과: In-distribution 성능 상승 대비 OOD(Out-of-Distribution) 성능 미흡
- 대응: ImageNet-C/A/R 등 강건성(Robustness) 벤치마크 등장



ImageNet

NLP (GLUE → SuperGLUE)

- GLUE의 빠른 포화(Saturation) → SuperGLUE 도입 → 곧바로 포화
- 이후: "NLP 벤치마크 자체가 문제"라는 논의 확산

추천/광고 시스템

- Offline metric(AUC, NDCG) 최적화가 Online CTR/CVR 개선으로 이어지지 않음
- 대응: Simpson's paradox 및 Selection bias 인지 후 온라인 A/B 테스트 중심 전환

논문 ① The Leaderboard Illusion — 배경

LMSYS 챗봇 아레나 (LMArena)

- 출범: 2023년 5월, LMSYS 주도, UC Berkeley, Stanford, UCSD, CMU, MBZUAI 협력. 2024년 9월 비영리 법인 설립
- 현황: 수백만 참여자, 300만+ 투표 수집, 커뮤니티 주도 실시간 LLM 평가 플랫폼

업계, 학계, 미디어에 막대한 영향력 행사

사실상의 SOTA(State-of-the-Art) 표준 지표로 통용

핵심 질문: 이 리더보드가 '현실의 능력'을 반영하는가?

https://openlm.ai/chatbot-arena/

Arena 순위 계산 — Bradley-Terry (BT) 모델

⚖ Pairwise Comparison (쌍 비교) 기반 추정

- ✓ 모든 모델이 서로 대결하지 않아도 순위 산출 가능
- ✓ 무승부(Tie) 처리 가능 및 통계적 신뢰구간 제공

⚠ 모델의 핵심 가정 (The Leaderboard Illusion에서 위반됨):

1. 비편향 샘플링 (Unbiased Sampling)

→ 특정 제공자가 **Best-of-N 전략**을 쓰면 위반됨

2. 전이성 (Transitivity)

→ A > B, B > C 이면 A > C 성립해야 함 (폐기 모델 발생 시 문제)

3. 완전 연결 그래프 (Fully Connected Graph)

→ 모든 모델이 직간접적으로 연결되어야 함

Private Testing: 비공개 다중 제출

일부 제공자의 특권 (Undisclosed Policy)

- 다수의 비공개 모델(variant) 테스트 허용
- 테스트 후 **최고점 모델만 선택 공개** 가능

Private Variants 현황 (2025 Jan-Mar)

> • **Meta**: 한 달간 **27개** 테스트 (Llama 4 출시 전)
> • **Google**: **10개** 테스트
> • **Amazon**: 다수 테스트 확인

문제점

- 공정성 훼손: 자원이 많은 기업만 가능한 전략
- 데이터 접근 비대칭: 테스트 자체가 데이터 확보 수단

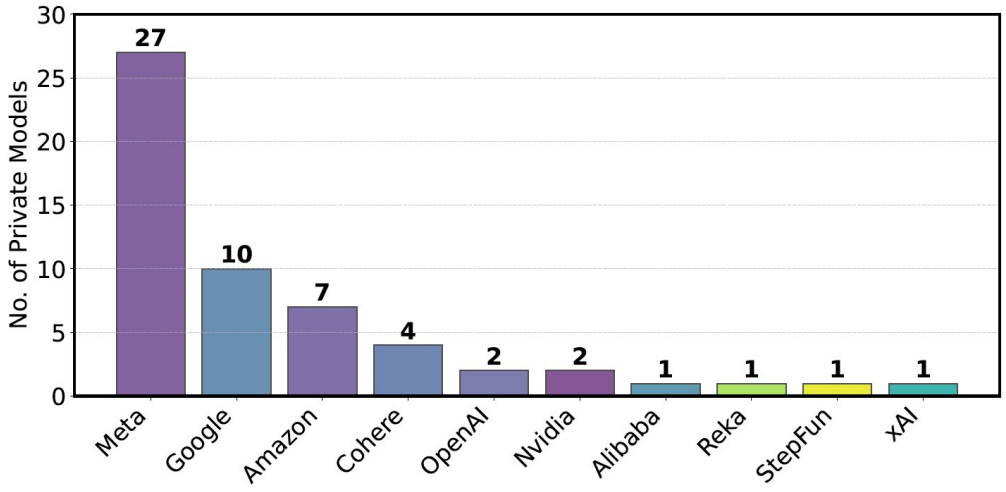


Figure 6: **Number of privately-tested models per provider based on `random-sample-battles` (January–March 2025)**. Meta, Google, and Amazon account for the highest number of

Selective Disclosure: Best-of-N 효과

- **Bradley–Terry 모델 가정 위반**
  - 기본 가정: "비편향 샘플링(Unbiased Sampling)"
  - 현실: N개 모델 중 최고값만 선택하여 공개
- **점수 인플레이션 메커니즘**
  - 극값 편향(Extreme Value Bias) 발생
  - 실제 평균 능력보다 훨씬 높은 점수로 기록됨
- **Cohere 측의 시뮬레이션 결과**
  - 단 20개 변형 테스트만으로 **최대 +50점** 상승
  - 능력 향상 없이 순위 조작 가능

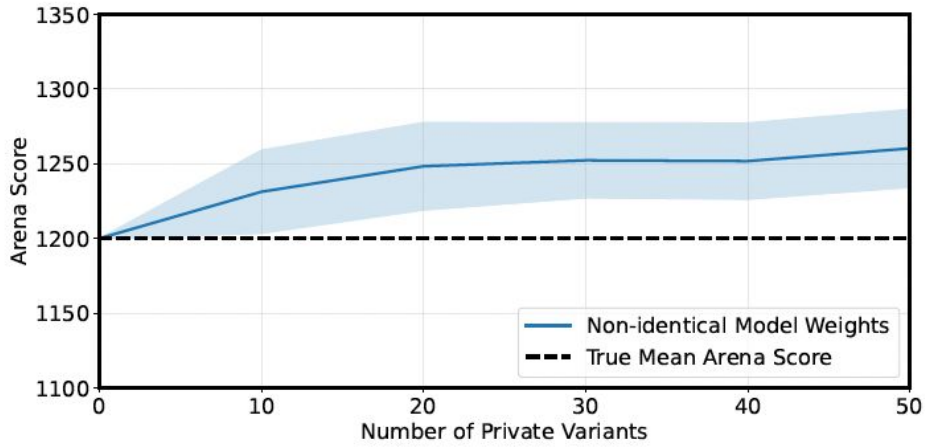


Figure 7: **Impact of the number of private variants tested on the best Expected Arena Score.** We simulate a family of model variants with a latent average Arena Score of 1200. As we

데이터 불균형과 샘플링 격차

Proprietary 모델의 데이터 독식

- 전체 Arena 데이터의 약 **54%~70%**를 차지
- OpenAI (20.4%) + Google (19.2%) ≈ 전체의 40%

오픈 모델의 소외 (Marginalization)

- 83개 오픈 모델 합계 ≈ **29.7%** 불과
- 학계/비영리 연구소 모델은 데이터 접근 기회 희박

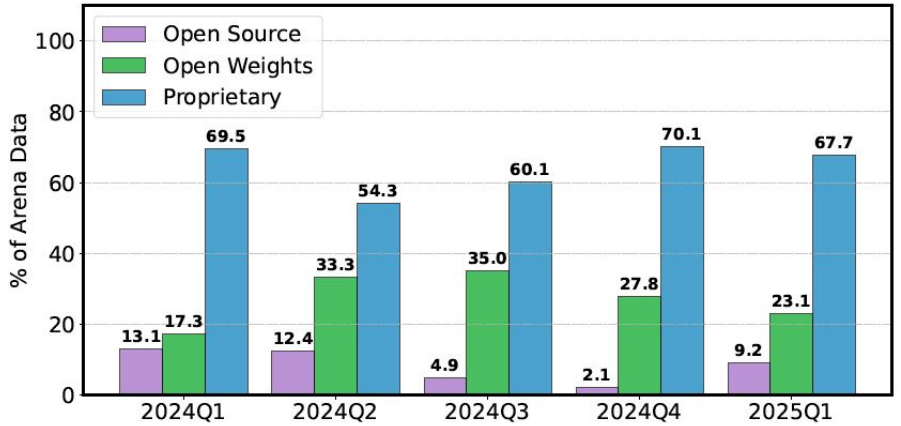


Figure 3: **Volume of Arena battles involving proprietary, open-weight, and fully open-source model providers from January 2024 to March 2025, based on `leaderboard-stats`.**

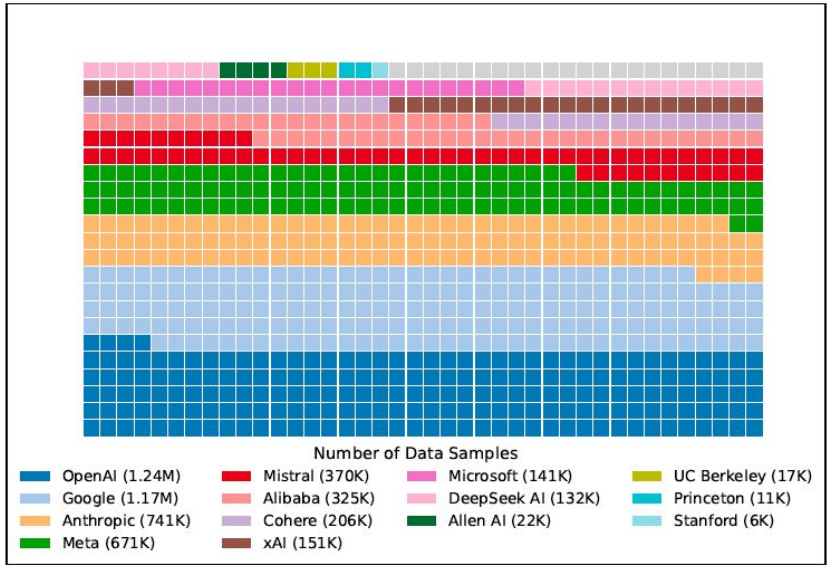


Figure 4: **Data availability to model providers.** We observe large differences in data access

Arena 데이터 과적합 위험 — 실제 실험 결과

실험 설정:

동일한 예산에서 Arena 데이터 비율 증가 (0% → 70%)

ArenaHard 승률 변화:

> 23.5% → **49.9%** (상대적 +112% 상승)

일반화 성능 (MMLU):

개선 없이 소폭 하락 66.5 -> 65.9% (Arena 분포에 특화된 최적화)

Silent Deprecation — 조용한 폐기 정책

정책 : "동일한 시리즈에 더 최근의 모델이 두개 이상 존재하거나 동일하거나 더 저렴한 가격 Arena Score기준 훨씬 우수한 모델을 제공하는 공급자가 3개 이상일 경우 3000표 이상 획득한 모델은 폐기".

그러나 실제로 검증하기 어렵다. 가격이나 품질에 대해 어떤 기준인지 명확하지 않음.

- **투명성 부족 (공식 vs 실제):**
  공식 폐기 **47개** vs 실제 조용한 폐기 **205개**
- **Bradley-Terry 가정 위반:**
  연결 그래프 붕괴 및 전이성 추정 실패 **순위 왜곡**
- **불균등한 영향:**
  오픈/오픈웨이트 모델이 더 많이 폐기됨 **데이터 접근 격차 심화**

Silent Deprecation — 조용한 폐기 정책

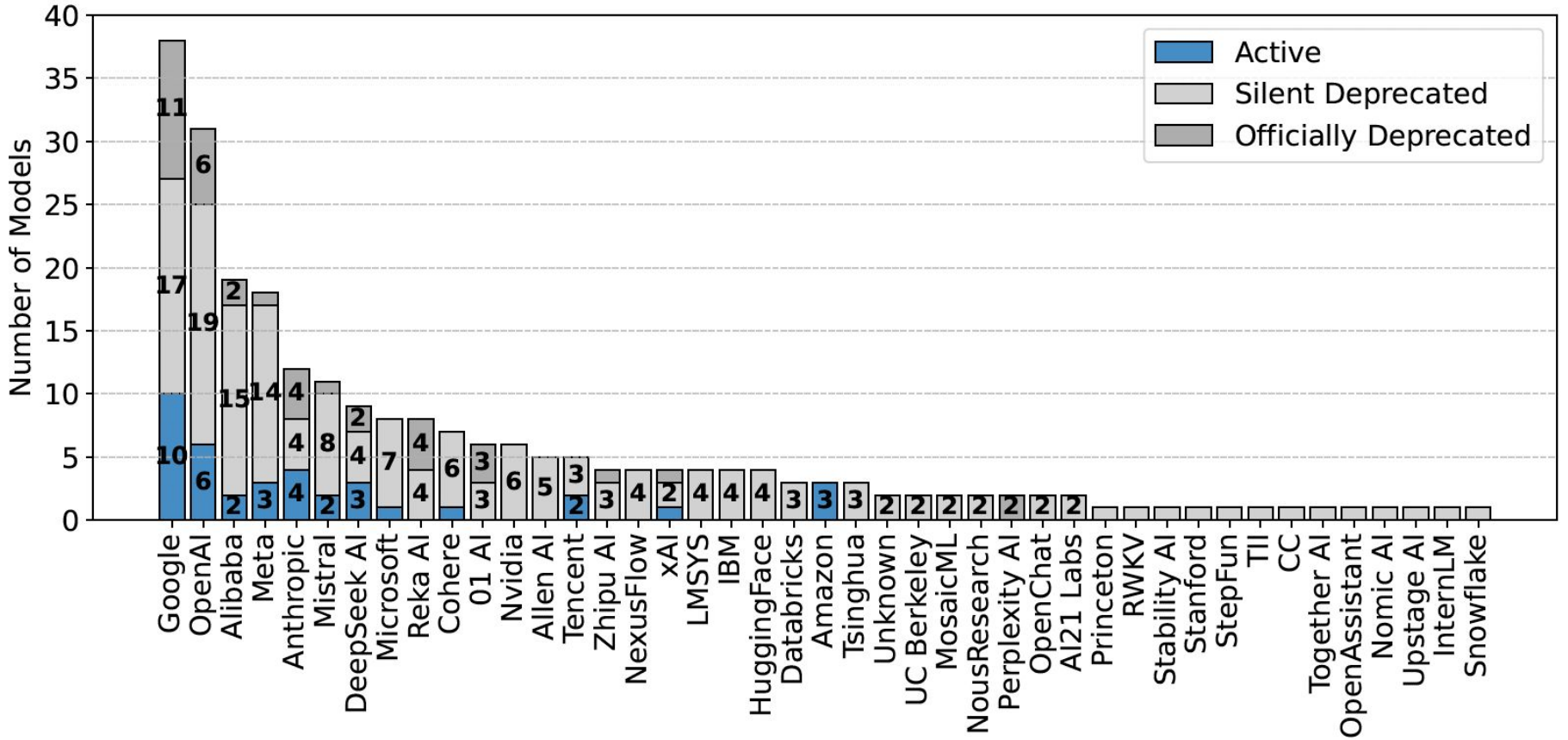


Figure 18: Share of active and deprecated models by provider including official and silent deprecations based on model activity between March 3-April 23, 2025.

핵심 메시지

■ Leaderboard는 모델의 능력을 측정하지 않습니다.

- 대신 **'Leaderboard를 잘 오르는 능력'**을 측정합니다.

■ Arena 특화 최적화 ≠ 일반적 모델 품질

- 제한된 데이터셋과 비공개 테스트로 인한 과적합 현상

■ Leaderboard의 본질적 한계

> **"Leaderboard is more of a MARKETING ARTIFACT than a SCIENTIFIC INSTRUMENT."**
>
> 리더보드는 과학적 계기라기보다 마케팅 산물에 가깝다.

개선을 위한 5가지 제안 (논문 권고)

🚫 1. 점수 철회 금지 (Score Retraction Ban)

비공개 포함 **모든 평가 결과 영구 공개** (선택적 공개 불가)

🔒 2. 비공개 테스트 제한 (Private Testing Limits)

투명하고 공개된 제한 설정 (예: 제공자당 최대 3개 동시 변형)

🗑️ 3. 공정한 폐기 정책 (Fair Deprecation)

라이선스 그룹별 **동일 기준 적용** (예: 각 그룹 하위 30% 폐기)

⚖️ 4. 샘플링 공정성 (Fair Sampling)

불확실성 높은 쌍 우선 평가하는 **Active Sampling** 방식 복원

👁️ 5. 투명성 강화 (Transparency)

모든 비공개 모델, 폐기 모델 목록 및 샘플링 비율 전면 공개

Line Goes Up?: 핵심 주장과 구조적 문제

핵심 주장: **점수 상승 ≠ 일반적 능력 향상**

벤치마크의 3가지 구조적 한계

Saturation (포화)

쉬운 문제는 이미 해결.
남은 것은 Trick이나 Artifact 위주의 문제들.

Contamination (오염)

Pre-training 데이터나 Instruction Tuning에 정답/패턴 노출.

Low Diversity (다양성 부족)

실제 세상은 Open-ended이나,
벤치마크는 Closed-form.

> *"Benchmarks are not measuring progress,*
> *they are measuring optimization pressure."*

벤치마크 분석 (1): 데이터 오염 & 과적합

SciEval (과학 문제 18,000문항)

> 목적: 동적 데이터 생성으로 오염 방지
>
> GPT-4 물리 성능 격차:
> 정적 데이터 **65%** → 동적 데이터 **26%** (**-39%p**)
>
> 결론: 문제의 '구조'가 아닌 '답'을 기억하거나 형식에 과적합됨

MMLU (다분야 지식 평가)

> Llama 3 사례: 포맷(Format)만 변경해도 성능 **-25%p** 하락
>
> 품질 관리 부실: 무관한 질문 포함 및 정답 오류 다수 보고

시사점: 벤치마크 점수 상승이 실제 '이해력(Understanding)' 향상을 의미하지 않음

벤치마크 분석 (2): 새 벤치마크의 한계

GPQA (전문가급, Google-proof)

목적: 검색으로 못 푸는 고난도 문제

한계: 인간은 난이도별 성능 차이가 뚜렷하나(전문가>>비전문가), **LLM은 난이도 변별력이 거의 없음** (표면적 단서 추측)

FrontierMath (신규 수학)

목적: 전문가 개발, 비공개 데이터셋

한계: **수치 정답만 요구**(추론 과정 검증 불가), OpenAI 자금 지원 및 일부 해법 선 보유 의혹

RE-Bench (ML 코드 생성)

목적: 실행 가능한 코드 평가

한계: 실시간 피드백으로 **무차별 대입(Brute-force)** 가능, 자동 평가는 통과했으나 수동 채점 시 오답인 사례 다수

벤치마크 분석 (3): 일반화 실패

GSM-Symbolic (산술 추론): 문제 구조 이해 실패

변수명만 변경해도 점수 하락: **-1 ~ -9%p**

무관한 정보 추가 시 성능 급락: **-65%** (일반 모델), **-17%** (o1 모델)

Theory of Mind (사회적 추론): 일관성 전무

GPT-4 대상 5가지 변형 실험 결과 극단적 편차 발생

2개 변형에서는 **0%**, 나머지에서는 **~100%** 정확도 기록

논리 추론 (Logical Reasoning): 취약한 강건성

규칙 순서만 변경해도 정확도 추락: **95% → 40%**

단순 Paraphrasing에도 답변 내용 **20~40%** 변경됨

벤치마크 분석 (4): 품질 & 자동평가 문제

벤치마크 데이터 자체의 **품질 결함**

Google Emotions: 데이터셋의 **30%**가 잘못 라벨링됨 (오류)

NLI & MMLU: 정답 오류, 모호한 질문 다수 발견

모델이 'all', 'some' 같은 **표면적 단어**에만 의존하여 정답 추측

자동 평가(Automated Evaluation)의 취약성

Null Model 실험: 항상 같은 답을 내지만 포맷만 조작하여 **GPT-4 평가자를 80% 속임**

Reversed Text: 문자 역순 데이터로 학습해도 정상 데이터와 **동등/이상 성능**

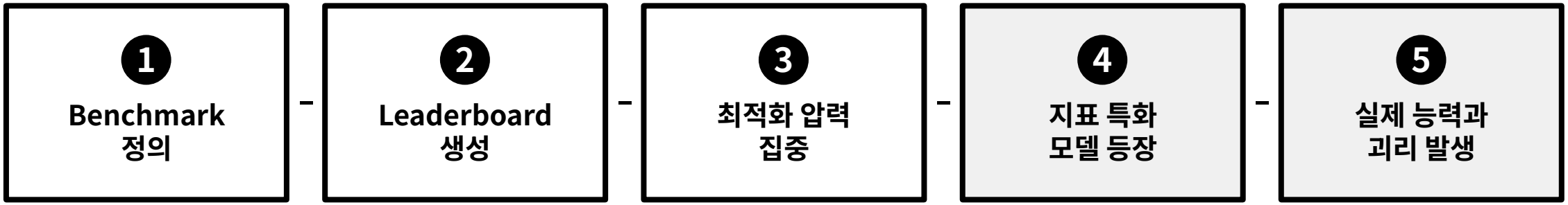
Chain-of-Thought (CoT)의 환상

초반에 의도적 오류를 주입해도 **60%+** 동일한 정답 도출 (사후 합리화)

프롬프트를 **무의미한 토큰**으로 대체해도 점수 유지 (95% vs 94%)

지표 붕괴 메커니즘 (5단계)

Goodhart's Law가 실제로 작동하는 과정: 벤치마크 정의부터 능력 괴리까지의 흐름



> 결과: **점수는 오르지만**(Line goes up), **일반화 능력과 강건성**(Generalization & Robustness)은 **저하됨**

기존 Evaluation의 실패 원인

Model-centric (모델 중심적 평가)

사용자나 실제 업무(Task) 맥락이 결여됨.
단순히 모델의 '기억력 테스트'에 치중하는 경향.

Static (정적 데이터셋)

고정된 데이터셋으로 반복 평가하여 오염 및 과적합 유발.
변화하는 실제 세상의 분포(Distribution Shift)를 반영 못함.

Scalar Metric (단일 스칼라 지표)

복잡한 성능을 숫자 하나(Accuracy, Pass@k)로 압축.
Failure mode, Robustness, Generalization 등 핵심 요소를 놓침.

Enterprise Evaluation ≠ Leaderboard Evaluation

리더보드 평가 (Leaderboard)

목적: 마케팅 & 커뮤니티
모델 간 순위 경쟁 및 가시성 확보 중심

메트릭: 단일 점수 중심
평균 정확도/승률 등 스칼라 지표 (Elo score)

데이터: 정적 & 공개 데이터셋
벤치마크 오염(Leakage) 및 과적합 취약

절차: 단순 선형 프로세스
제출 → 자동 평가/투표 → 점수 산출 → 순위

기업 내부 평가 (Enterprise)

목적: 의사결정 & 리스크 관리
실제 배포 여부(Go/No-Go) 판단 및 책임성

메트릭: 다중 복합 지표
품질, 안전, 비용, 지연시간, UX 등 동시 최적화

데이터: 동적 & 프로덕션 로그
실제 사용자 데이터, 적대적 공격, 도메인 특화

절차: 다단계 검증 파이프라인
Gate(공개) → Private → Shadow/Canary → A/B → Monitor

⚠ 핵심 메시지: Public Benchmarks는 단순한 'Gate'일 뿐, 최종 배포 의사결정은 '내부 프로덕션 평가'가 좌우합니다.

References:
[1] OpenAI, "GPT-4 System Card" (2023). https://cdn.openai.com/papers/gpt-4-system-card.pdf
[2] OpenAI, "GPT-4o System Card" (2024). https://openai.com/index/gpt-4o-system-card/
[3] Anthropic, "Claude 4 System Card" (2024). https://www.anthropic.com/claude-4-system-card

Enterprise Evaluation Pipeline

How production decisions actually happen

OFFLINE FILTER

1

Baseline Gate

Public Benchmarks

MMLU, GSM8K

2

Domain Offline

Internal Dataset

Log-based Tests

3

Stress & Red Team

Adversarial Tests

Jailbreak / Injection

ONLINE DECISION

4

Shadow / Canary

Safe Deployment

Traffic Mirroring

5

Online A/B

Production Metrics

Real KPI / Conversion

6

Continuous Monitor

Post-deployment

Drift / Rollback

"Offline is just a filter. Online is the final decision."

References:

• OpenAI. (2024). GPT-4o System Card. https://openai.com/index/gpt-4o-system-card/

• Google Vertex AI GenAI Evaluation. https://docs.cloud.google.com/vertex-ai/generative-ai/docs/models/evaluation-overview

• AWS SageMaker Shadow Tests. https://docs.aws.amazon.com/sagemaker/latest/dg/shadow-tests.html

Stage 1: Domain-Specific Offline Evaluation

Public Benchmarks are Gate Checks Only → Real Evaluation Happens with User Logs

게이트 통과 후, 도메인 오프라인 평가로 전환

실제 사용자 로그에서 샘플링한 태스크 및 실패 케이스 기반 배치 평가

제품 목표 지표로 재측정 (작업 완수율, 사실성, 정책 위반 회피 등)

Google Search

전통적 **정확도(Accuracy)만으로는 검색 품질 설명 불가**. 행동 지표(Click-through, Dwell time)와 체감 품질 지표를 결합하여 평가.

Microsoft Bing (Chat)

공개 벤치마크 + 검색 품질 평가 + **제한적 프리뷰**를 통한 피드백 수집 병행. "Real-world use"에서의 학습 강조.

OpenAI

내부 정량 평가 세트(Internal Quantitative Evals) 구축. 정책 위반(Hate speech 등) 및 환각 여부를 체크포인트별로 자동 비교.

Meta

출시 전 기능별 **세이프가드(Safeguards)** 구축 및 사전 테스트. 기능 단위의 책임있는 롤아웃 원칙 적용.

References:

• Google AI Blog (2021). *Evaluating Search Quality Beyond Accuracy Metrics.* https://ai.googleblog.com/2021/07/evaluating-search-quality-beyond.html - "Traditional accuracy metrics are insufficient for real-world search quality"

• Microsoft (2023). *Reinventing Search with a New AI-Powered Bing.* https://blogs.microsoft.com/blog/2023/02/07/reinventing-search-with-a-new-ai-powered-bing-and-edge/ - "We will continue to learn from real-world use and feedback during the preview"

• OpenAI (2023). *GPT-4 System Card.* https://cdn.openai.com/papers/gpt-4-system-card.pdf - "We built internal quantitative evaluations ... to automate and accelerate evaluations"

• Meta (2023). *Building Generative AI Features Responsibly.* https://about.fb.com/news/2023/09/building-generative-ai-features-responsibly/

Stage 2: Stress Testing & Adversarial Evaluation

Finding Failures Beyond Benchmarks through Red Teaming & RSP

⚡ 벤치마크 밖의 실패를 찾는 스트레스 테스트

테스트 범위 확장: 길이·포맷 변화, 입력 잡음/교란(Noise), 정책 경계(Policy Boundaries), 복합 툴 체인 등 벤치마크가 놓치는 영역 집중 공략.

운영 포인트: 발견된 실패 케이스를 즉시 카탈로그화하여 **회귀 방지(Regression) 스위트**에 필수 편입, 모델 업데이트 시 재검증.

An Anthropic (RSP & Safety)

RSP(Responsible Scaling Policy)에 따라 사전 안전성 테스트 의무화. 단순 성능이 아닌 **적대적 지시(Adversarial instructions)**, 프롬프트 인젝션, 에이전틱 코딩 위험(CBRN 등)을 중점 평가.

OA OpenAI (Red Teaming)

50명 이상의 전문가 레드팀(Expert Red Teamers) 운영. 도메인별 경계·스트레스 테스트를 반복하며 **완화책(Mitigations)**을 지속적으로 튜닝하고 실패율 감소 검증.

References:

• Anthropic (2025). *Claude 4 System Card*. https://www.anthropic.com/claude-4-system-card - "A wide range of pre-deployment safety tests conducted in line with the commitments in our Responsible Scaling Policy"

• OpenAI (2023). *GPT-4 System Card*. https://cdn.openai.com/papers/gpt-4-system-card.pdf - "We refer to these adversarial testing processes informally as 'red teaming' ... 'a structured effort to find flaws and vulnerabilities'"

Stage 3: Shadow Deployment & Canary Testing

Detect Hidden Failures with Real Traffic Before Full Launch

실제 트래픽 그림자·소량 노출로 숨은 실패 탐지

Shadow Deployment: 동일 요청 복제하여 모델에 전송하되 응답은 비노출, 성능/에러 모니터링

Canary Testing: 1–10% 소량 트래픽에만 점진적 롤아웃, 문제 발생 시 즉시 중단

Meta (Responsible Rollout)

출시 전 책임있는 롤아웃 원칙 하에 **제한적 사용자군**에서 모델의 거동 및 세이프가드 작동 검증. 잠재적 위험 최소화를 위해 단계적 접근 필수.

Uber Engineering (Michelangelo)

Michelangelo 플랫폼으로 대규모 모델 배포 관리. Shadow 모드 검증 후 Canary 배포, 최종 Full Rollout의 파이프라인 정착. 실서비스 데이터 기반 안정성 확보.

Industry Standard (Rollback)

Rollback Triggers 사전 정의: 안전성 지표(Safety violations), 지연 시간(Latency), 에러율(Error rates) 급증 시 자동 롤백 시스템 구축.

Technical Debt Perspective

Sculley et al. (NIPS 2015)에서 강조한 **"Hidden Technical Debt"**: 프로덕션 ML 시스템은 오프라인 메트릭을 넘어 지속적인 모니터링과 검증이 필수적임.

References:

- Meta (2023). *Building Generative AI Features Responsibly*. https://about.fb.com/news/2023/09/building-generative-ai-features-responsibly/ - "Before we launch ... safeguards ... testing"
- Uber Engineering. *Michelangelo Machine Learning Platform*. https://eng.uber.com/michelangelo/ - "Allows teams to deploy and monitor models in production at scale"
- Sculley et al. (2015). *Hidden Technical Debt in Machine Learning Systems* (NIPS 2015). https://research.google/pubs/hidden-technical-debt-in-ml-systems/ - "Production ML requires ongoing monitoring and validation beyond offline metrics"
- Oracle (2026). *Canary Deployments for Securing LLMs*. https://medium.com/@oracle_43885

Stage 4: Production A/B Testing

Real-World Verification: Measuring Actual Usage Effects Online

온라인 실험으로 '실사용 효과'를 계량

목표: 사용자 가치(Value)와 리스크(Risk) 지표의 실질적 변화 검증

핵심 지표: 작업 완수율(Task Completion), 참여도(Engagement), 이탈률, 신고 수, 비용 및 지연시간

Meta

대규모 온라인 실험 프레임워크 (PlanOut) 활용: 기능, 랭킹 알고리즘, 생성 품질 평가

Online Metrics 우선: **CTR(클릭률), Dwell-time(체류시간), Engagement** 변화 측정

통계적 유의성 검증을 통한 배포 의사결정

OpenAI

외부 파트너(Independent Experts)와 협력하여 배포 전 사전 테스트 수행
출시 후 지속적인 모니터링 강화로 **실사용 신호(Real-world signals)** 수집 및 피드백 통합

Preparedness Framework에 따른 배포 후 지속 평가

References:

• Facebook Research (2014). *PlanOut: A Framework for Online Field Experiments.* https://research.facebook.com/publications/planout-a-deployment-framework-for-online-field-experiments/ - "A framework for designing, deploying, and analyzing online experiments at scale"

• Meta (2023). *Responsible AI at Meta.* https://ai.facebook.com/blog/responsible-ai-at-meta/ - Online A/B testing, CTR monitoring emphasized

• OpenAI (2025). *Strengthening Safety with External Testing.* https://openai.com/index/strengthening-safety-with-external-testing/ - "Work with independent experts to evaluate frontier AI systems"

• Sculley et al. (NIPS 2015). *Hidden Technical Debt in ML Systems.* https://research.google/pubs/hidden-technical-debt-in-ml-systems/ - Emphasizes online evaluation loops

Stage 5: Human Evaluation as Risk Sensor

Humans Detect Risks (Hallucination, Harm), Not Just Accuracy Scores

인간 평가는 '정확도 점수'가 아닌 '리스크 감지 센서'

환각(Hallucinations), 불일치 추론, 정책 경계 위반, 에이전트 오용 시나리오 포착

정량적 지표가 놓치는 **맥락적 뉘앙스**와 **잠재적 위험** 식별에 집중

OA OpenAI

정책 카테고리별 정량 평가 + **인간 분석(Human Analysis)**으로 체크포인트 비교 및 완화책 효과 검증.
"Evaluators look for hallucinations, inconsistent reasoning, harmful outputs."

Microsoft

Responsible AI 표준에 따라 **고위험 사용사례(High-risk use cases)**에 인적 감독(Human Oversight) 및 휴먼 재량 도입.

An Anthropic

무해성(Harmlessness), 정직성, 에이전틱 안전 평가 결과를 투명하게 공개.
Claude Opus 4: 98.43% harmless response rate 달성.

운영 포인트 (Operation)

휴먼 라운드트립 샘플링(고위험 분포 가중) → 라벨 품질·합의도 관리

실패 사례를 회귀 스위트(Regression Suite) 및 적대 세트로 전파

References:

- OpenAI (2023). *GPT-4 System Card*. https://cdn.openai.com/papers/gpt-4-system-card.pdf - "Generated text ... classified ... using classifiers and human analysis"
- Microsoft. *Responsible AI Standard*. https://www.microsoft.com/en-us/ai/responsible-ai - "Human oversight and impact assessments for higher-risk use cases"
- Anthropic (2025). *Claude Opus 4 System Card*. https://www.anthropic.com/claude-4-system-card - "Tests of model safeguards, honesty, and agentic safety"

OpenAI: Risk detection, not accuracy measurement

> **"OpenAI는 모델을 하나의 점수로 보지 않는다"**

Preparedness Framework의 다중 위험 축

Persuasion (설득)	Autonomy (자율성)	Cybersecurity (사이버 악용)
Deception (기만)	CBRN (화생방방사능)	

평가 설계의 핵심

- ✕ 정적 문제풀이 (Static Q&A)
- ✓ **레드팀 대화, 실제 공격 시도**
- ✓ **정책 경계 상황 테스트**

평가 결과의 라벨링 (이산적 실패 이벤트)

위험한 행동을 끝까지 수행했는가?

거절이 필요한 상황에서 거절했는가?

안전 가이드를 우회했는가?

"OpenAI의 평가는 정확도 측정이 아니라 위험 행동 감지용 분류 시스템"

• OpenAI (2024). Preparedness Framework v2: Multiple capability axes (Persuasion, Autonomy, Cyber, CBRN, Deception). https://openai.com/index/preparedness-framework/
• OpenAI (2024). GPT-4o System Card: Risk-based evaluation categories, red teaming. https://openai.com/index/gpt-4o-system-card/
• OpenAI (2023). GPT-4 System Card, p.2-3: "engaged more than 50 experts" for domain-specific risk evaluation. https://cdn.openai.com/papers/gpt-4-system-card.pdf

OpenAI Operator: From tokens to actions

> **"평가 대상이 '토큰'에서 '행동'으로 바뀌는 순간"**

Agent 모델의 출력 변화

기존 LLM		Operator
텍스트 토큰 시퀀스	→	**행동 시퀀스**
(Text Generation)		(Click, Input, Scroll)

기술적 평가 방법의 전환:

단순 텍스트 매칭으로는 평가 불가능.
사용자 컴퓨터 제어 환경에서의 **Action Trace** 분석 필수.

행동 레벨(Action-Level) 평가 방법

1) 행동 로깅 & 그래프화 (Action Trace)

각 행동을 로그로 기록하고 순서를 그래프로 표현하여 분석

2) 행동 레벨의 실패 조건

- "의도하지 않은 행동을 했는가?"
- "사람 개입 없이 위험 작업을 완료했는가?"

3) 평가 구조의 적응 (Adaptive Eval)

기존 텍스트 평가 ≠ 행동 평가 → 새 기능 맞춤형 구조 설계

"새 기능이 생기면 평가 구조도 반드시 바뀌어야 함"

- OpenAI (2025). Operator System Card: Computer use, action-level evaluation, agentic safety. https://openai.com/index/operator-system-card/
- OpenAI (2025). "Strengthening Safety with External Testing": Agent evaluation, behavioral monitoring. https://openai.com/index/strengthening-safety-with-external-testing/

Anthropic Responsible Scaling Policy (RSP): "Better model = Harder eval"

"모델 성능 ↑ → 평가를 더 세게 해야 한다"

RSP : 모델의 능력이 커질수록, 평가 강도·안전 장치·배포 제약을 의무적으로 강화하도록 설계된 정책

RSP 단계별 평가 강도

모델 능력 (Capability)	평가 & 보호조치
기본 수준	표준 평가 + ASL-1
임계치 근접 ⚠	강화 평가 + ASL-2
위험 수준 ⚠⚠	엄격 평가 + **ASL-3**
심각 위험 🚫	최대 평가 + ASL-4 or 배포 제한

* RSP: Responsible Scaling Policy
* ASL: AI Safety Level

실제 적용 사례

Claude Opus 4:
ASL-3 Standard (강화된 Safeguards)

Claude Sonnet 4:
ASL-2 Standard

설계 철학

✗ 성능 좋음 → 평가 덜 해도 됨

✓ **성능 좋음 → 평가 더 세게 해야 함**

"Capability ↑ → Evaluation intensity ↑ → Safeguards ↑"

• Anthropic (2024). Responsible Scaling Policy (RSP): "Capability thresholds trigger increased evaluation intensity". https://www.anthropic.com/news/responsible-scaling-policy
• Anthropic (2025). Claude Opus 4 & Sonnet 4 System Card, Section 1.2.3: "Claude Opus 4 under ASL-3 Standard, Claude Sonnet 4 under ASL-2 Standard". https://www.anthropic.com/claude-4-system-card
• Anthropic (2025). "Activating ASL-3 Protections for Claude Opus 4". https://www.anthropic.com/news/activating-asl-3-protections

Goodhart's Law in LLM evaluation

Why metrics stop being good measures

The Mechanism

1 Metric becomes target

벤치마크 점수가 개발의 목표가 됨

→

2 Optimization pressure ↑

포맷, 규칙, 데이터 누수에 과적합

→

3 Metric ≠ Goal

점수는 올라도 실제 능력은 정체

Manifestation in LLMs

Leaderboard overfitting

특정 벤치마크의 프롬프트 형식이나 평가 데이터셋 누수에 모델을 과도하게 최적화함

"Judge" gaming

LLM-as-a-judge 자동 평가 모델이 선호하는 패턴(길이, 특정 단어 등)을 학습하여 점수를 부풀림

Distribution shift

실제 사용 환경이나 조건이 조금만 바뀌어도 벤치마크 점수와 무관하게 성능이 급락함

"Enterprise evaluation pipelines are essentially Goodhart countermeasures."

(기업의 평가 파이프라인은 사실상 Goodhart 방지장치다)

References:

• The Leaderboard Illusion. arXiv:2504.20879. https://arxiv.org/abs/2504.20879 • Line Goes Up? Inherent Limitations of Benchmarks. arXiv:2502.14318. https://arxiv.org/abs/2502.14318

How enterprises design against Goodhart

4 Defense Mechanisms for Evaluation Integrity

Multi-metric & Category-based Eval

단일 점수(Single Score)가 목표가 되는 것을 원천 차단. 안전성, 유용성, 윤리성 등 다차원 평가 지표를 복합적으로 활용.

Adversarial / Red Teaming

모델의 취약점, 꼼수(Gaming), 우회 방법을 인간 전문가와 자동화된 툴이 공격적으로 찾아내어 조기 노출시킴.

Adaptive Eval

고정된 테스트셋 대신, 기능 추가나 환경 변화에 맞춰 평가 기준과 데이터셋을 지속적으로 갱신(Update)함.

Online Decision + Monitoring

배포 후 실제 사용자 데이터(Real Traffic)에서의 KPI와 사고율로 최종 검증하고, 문제 시 즉각 롤백(Rollback) 준비.

"Leaderboards rank models. Enterprises manage failure."

(리더보드는 순위를 매기지만, 기업은 실패를 관리한다)

References:

• OpenAI. System Cards (GPT-4, GPT-4o, Operator). https://openai.com/index/gpt-4o-system-card/

• Meta. Responsible AI at Meta. https://ai.facebook.com/blog/responsible-ai-at-meta/

• Sculley et al. (2015). Hidden Technical Debt in Machine Learning Systems. NIPS 2015.