LLM Evaluation: an Overview

2026-01-17

김기범

Reference: 거꾸로 읽는 SSL

[스터디] 거꾸로 읽는 SSL 시즌4: Large Language Models and Alignment Learning

- 신청 Google form (신청 마감)
- 거꾸로 읽는 SSL 이번에는 LLM and AL 분야로 넘어 왔습니다! :)
- 2022년도 이후로 가파르게 발전하고 있는 LLM 논문에 집중하여 의미가 있었던 논문을 살펴봅니다.
- 해당 논문에서 제시하는 메소드의 특징 그리고 역사적으로 평가되는 이유에 대해서 즐겁게 토론하는 시간을 가집니다.

기간 (예정)

- 2024 1/27 ~ 6/15 (예정)

발표 논문 및 순서

1 주차	overview	overview (참고 논문)			강재욱
2 주차	pretraining	Training Compute-Optimal Large Language Models	DeepMind	2022 Mar	김택민
3 주차	pretraining	Will we run out of data? An analysis of the limits of scaling datasets in Machine	MIT	2022 Oct	이인

https://github.com/jwkangmarco/LLM-narratives , https://www.youtube.com/@gokkulearn

프로젝트 소개: 인간 연구자의 리뷰 평가

Do Peer Review Scores Predict Scientific Impact?

- 도메인 전문가(reviewer)는 연구의 향후 impact를 판별할 수 있을까?

  Github: https://github.com/isingmodel/openreview_ratings_vs_citations

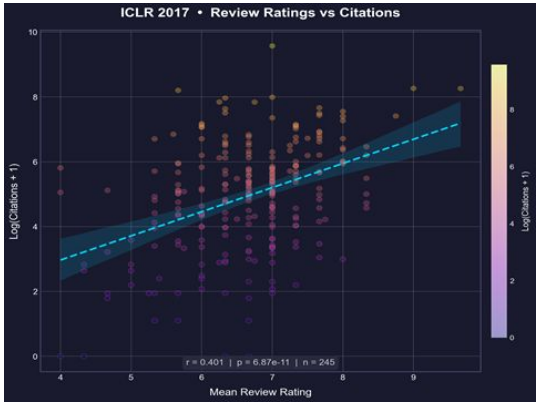
  PyconKR 2022: https://www.youtube.com/watch?v=MqM2ROgWwhU

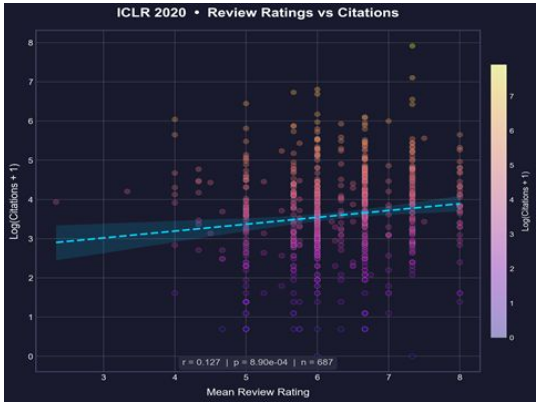
- 2017-2020 ICLR review와 Google citation 데이터의 연관성을 탐색

  correlation(log(citation), review rating)

- 결론: 인간 전문가는 Scientific Impact를 평가하지 못하고 있다.

Year	Papers	Correlation (r)	p-value
2017	245	0.40	6.9e-11
2018	425	0.37	1.6e-15
2019	502	0.19	1.2e-05
2020	687	0.13	8.9e-04





왜 평가해야 할까?

- 평가가 없으면,
  - 전부 **감**으로 굴러간다.
  - 서비스의 구조를 바꿀 때마다 기도해야 한다.

- 정확도, 신뢰도 관리

- 모델 선택:
  - 어떤 모델을 써야 좋은 결과가 나올까?
  - 비용 대비 효과, Pro 모델을 꼭 써야할까?

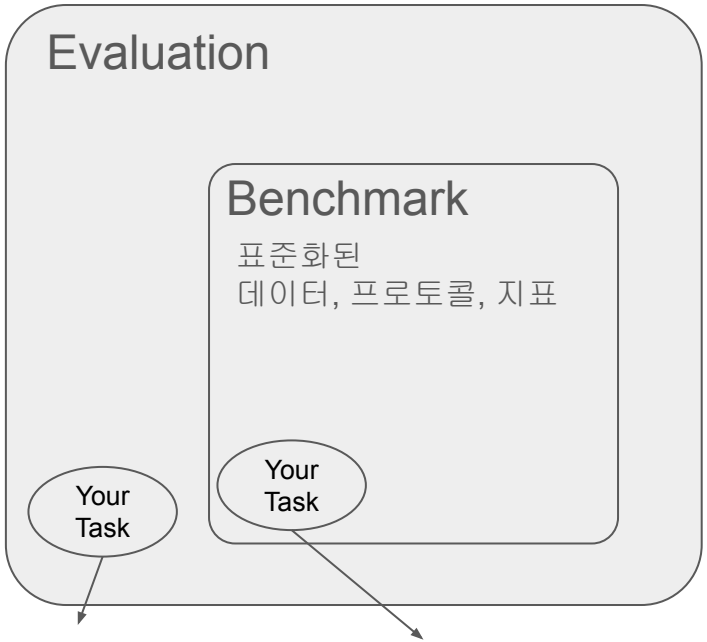
- '품질'을 정량화해야 한다.
  - 품질: 목표 기능에 대한 정확도, 신뢰도, 비용

- Fine-Tuning/RL용 데이터셋

어떻게 평가해야 할까? Benchmark를 따라가자.

- 표준화된 기준
- 모델 선택의 근거
- 평가 방법론 제공: 어떤 데이터를, 어떤 방식으로.

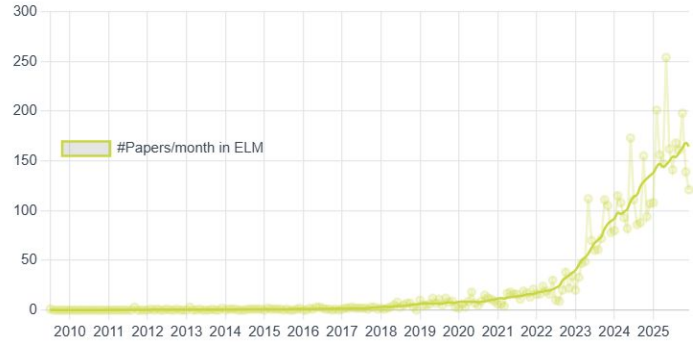
How to Evaluate: Follow Benchmarks



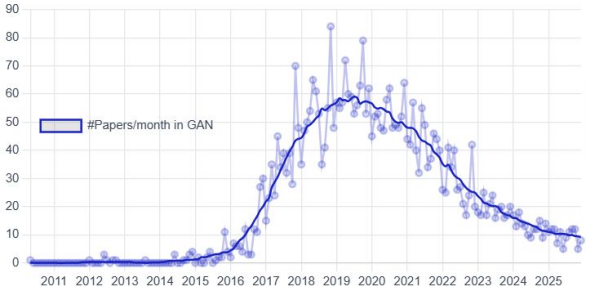
Bench는 단서일 뿐

해당 bench의 top 모델을 쓰면 됨.

evaluation 연구들은 증가 중

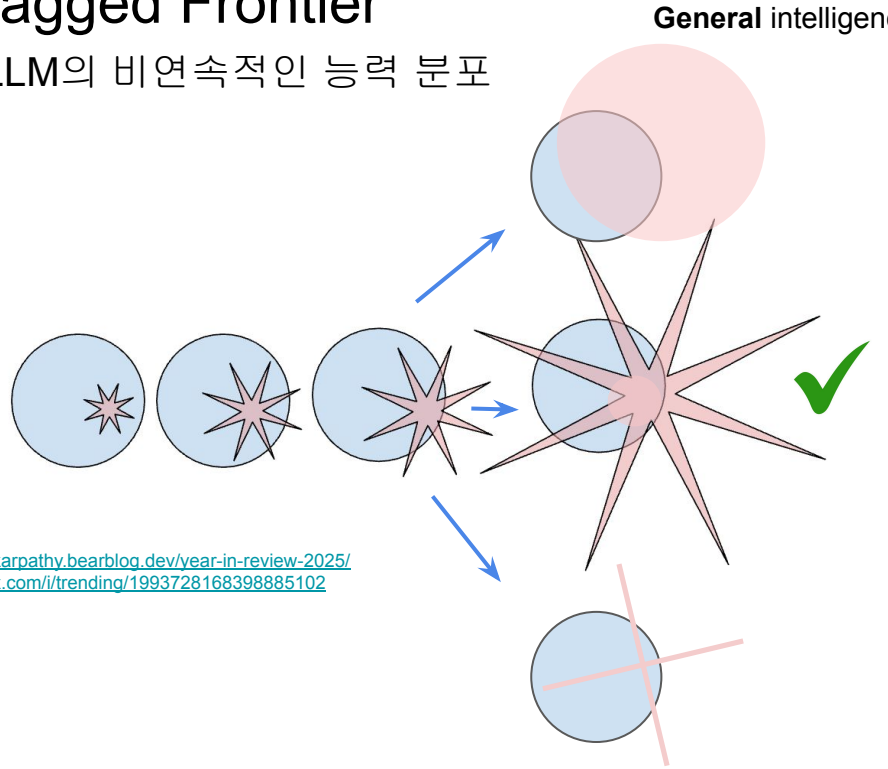


Evaluating Language Models: papers per month
https://researchtrend.ai/communities/ELM

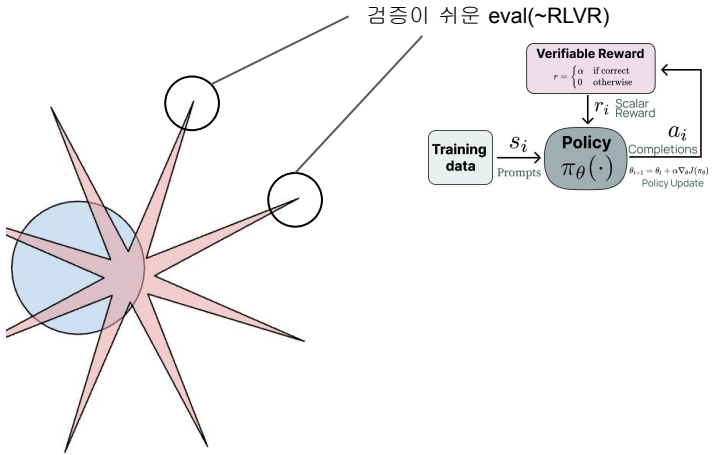


Jagged Frontier

LLM의 비연속적인 능력 분포



https://karpathy.bearblog.dev/year-in-review-2025/
https://x.com/i/trending/1993728168398885102



ARC Prize
@arcprize

A North Star for open AGI. Co-founders: @fchollet @mikeknoop. President: @gregkamradt. We're hiring mission-driven builders: arcprize.org/jobs

https://x.com/arcprize, Arc-agi-2

Frontier model은 저울을 바꾼다.

Our next-generation model: Gemini 1.5

Feb 15, 2024
9 min read

The model delivers dramatically enhanced performance, with a breakthrough in long-context understanding across modalities.

Enhanced performance

When tested on a comprehensive panel of text, code, image, audio and video evaluations, 1.5 Pro outperforms 1.0 Pro on 87% of the benchmarks used for developing our large language models (LLMs). And when compared to 1.0 Ultra on the same benchmarks, it performs at a broadly similar level.

Gemini 1.5 Pro maintains high levels of performance even as its context window increases. In the Needle In A Haystack (NIAH) evaluation, where a small piece of text containing a particular fact or statement is purposely placed within a long block of text, 1.5 Pro found the embedded text 99% of the time, in blocks of data as long as 1 million tokens.

Long Context

Gemini 1.5 Pro also shows impressive "in-context learning" skills, meaning that it can learn a new skill from information given in a long prompt, without needing additional fine-tuning. We tested this skill on the Machine Translation from One Book (MTOB) benchmark, which shows how well the model learns from information it's never seen before. When given a grammar manual for Kalamang, a language with fewer than 200 speakers worldwide, the model learns to translate English to Kalamang at a similar level to a person learning from the same content.

As 1.5 Pro's long context window is the first of its kind among large-scale models, we're continuously developing new evaluations and benchmarks for testing its novel capabilities.

For more details, see our Gemini 1.5 Pro technical report.

https://blog.google/innovation-and-ai/products/google-gemini-next-generation-model-february-2024

2024년 Gemini 1.5는 모델 공개 당시 23년에 공개된 GPT-4보다 성능이 비등하거나 떨어짐.

General Reasoning and Comprehension

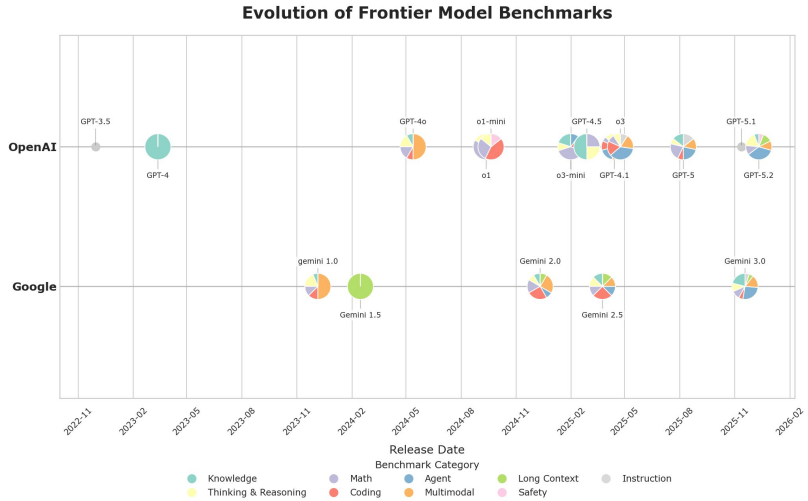
Benchmark	Gemini 1.5 Turbo	GPT-4 Turbo	Description
MMLU	81.9%	80.48%	Multitask Language Understanding
Big-Bench Hard	84.0%	83.90%	Multi-step reasoning tasks
DROP	78.9%	83%	Reading comprehension
HellaSwag	92.5%	96%	Commonsense reasoning for everyday tasks

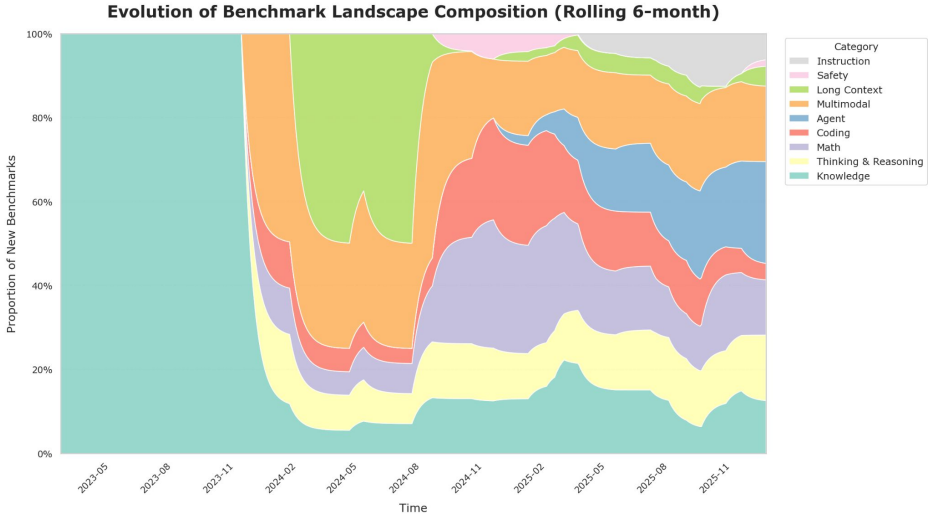
https://bito.ai/blog/gemini-1-5-pro-vs-gpt-4-turbo-benchmarks/

LMArena Text(2026-01-15)
gpt-4-turbo-2024-04-09: 1325±4
gemini-1.5-pro-001: 1323±4

자기소개서를 보면 방향성이 보인다.

- Gemini&ChatGPT의 '모델 공개 문서'에 기재된 benchmark 들만 모아 분석. (Technical report 등은 제외)
  github.com/isingmodel/evolution_of_benchmarks_in_frontier_models
- 단순 지식 → Multimodal/Coding → Agent로 변모하는 과정이 보임.
- 초기에는 Academy에서 공개한 Benchmark 및 공개 시험 문제만으로 평가했지만 점차 **자체 benchmark**를 내세우기 시작.
- 후발주자 Google은 Gemini 시리즈를 출시하면서 OpenAI가 개발한 벤치마크를 적극적으로 사용하여(MMMLU 등) 성능을 비교.





A Survey on Evaluation of Large Language Models

Y Chang et al., 2023

1. 어떤 능력을 평가하는가?
2. 어떤 데이터로 평가하는가?
3. 어떤 지표와 프로토콜로 평가하는가?

https://arxiv.org/pdf/2307.03109 https://github.com/MLGroupJLU/LLM-eval-survey

어떤 능력을 평가하는가?

Knowledge, Reasoning, Math(~Reasoning), Safety,
Instruction following, Coding, Agent, Long Context 등등의 조합

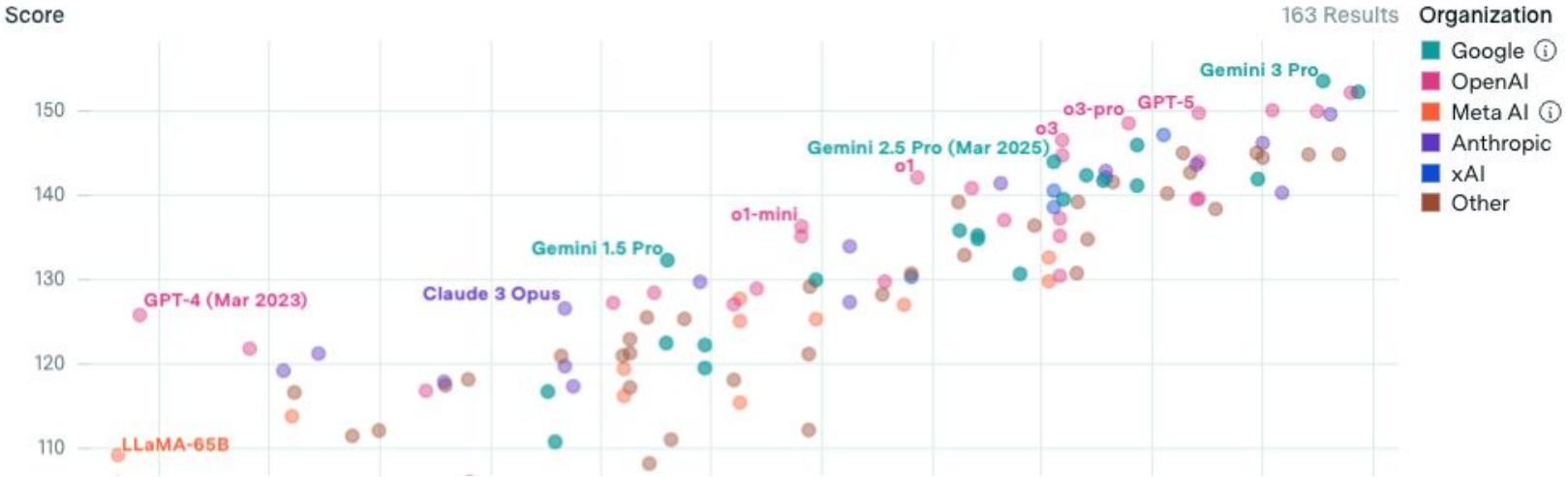
Multimodal needle in a haystack: Benchmarking long-context capability of multimodal large language models, *Microsoft Research*

→ Long Context + Multimodal

어떤 능력을 평가하는가?: 종합평가

Epoch Capabilities Index

"general capability"

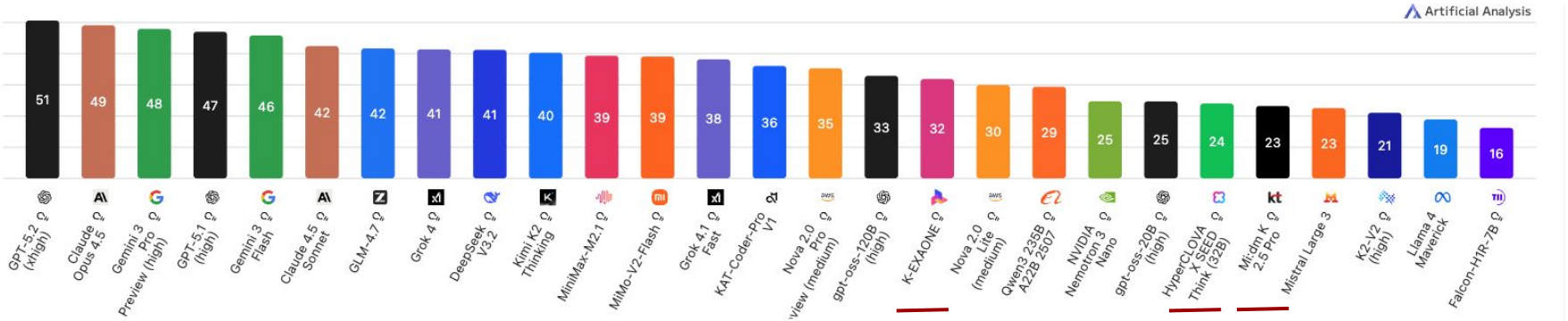


https://epoch.ai/benchmarks/eci

어떤 능력을 평가하는가?: 종합평가

Artificial Analysis Intelligence Index

GDPval-AA, τ^2-Bench Telecom, Terminal-Bench Hard, SciCode, AA-LCR, AA-Omniscience, IFBench, Humanity's Last Exam, GPQA Diamond, CritPt



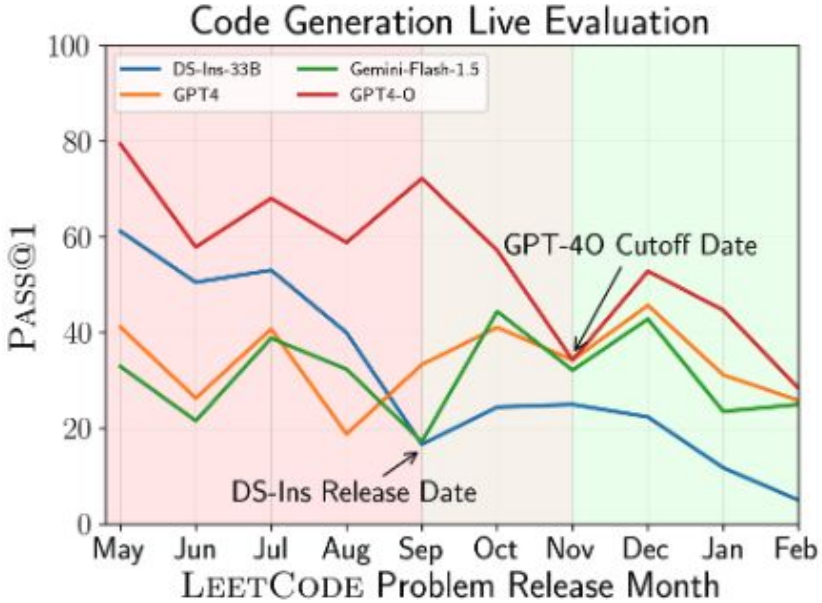
https://artificialanalysis.ai/

어떤 데이터로 평가하는가?

- 정적 데이터: 공개셋
- 정적 데이터: 비공개셋
- 실시간 생성 데이터(e.g. LiveCodeBench: 실제 경쟁 프로그래밍 플랫폼에서 새로 공개되는 문제들)
- 가상 API 세팅, 실제 API 세팅(agent evaluation의 경우)

어떤 데이터로 평가하는가? 데이터 오염

- Benchmark의 데이터가 직/간접적으로 모델 학습에 포함되는 오염이 발생.
- 실시간 생성 데이터를 활용하려는 방식으로 해결하려는 Benchmark들이 있음. (e.g. LiveCodeBench)
- 의견: LLM era가 촉발하는 데이터 분포 변화에 의해 실시간 stationarity를 유지할 수 없을 것.



LiveCodeBench

어떤 데이터로 평가하는가? Benchmark의 진화

어떤 지표와 프로토콜로 평가하는가?

평가 프로토콜

- 단일 정답 기반: 주관식 단답형
- 다중선택: 객관식
- Pairwise(A vs B): Winning Rate, ELO(e.g. LMArena)
- 서술 평가
- 성공률(agents, coding), 비용/시간, 자원 관리(e.g. Vending-Bench)
- pass@k: 모델이 k개의 후보 답안을 생성했을 때, 그중 하나라도 정답이면 성공

평가 주체

- 인간 평가자
- LLM-as-Judge
- 자동 지표(BLEU, ROUGE 등)

Long Context

LLM이 정보를 실제로 끝까지 읽고, 필요한 정보를 정확히 찾아 연결하며, 장거리 의존 추론을 유지하는가?

LongBench v2: Towards Deeper Understanding and Reasoning on Realistic Long-context Multitasks
긴 context size를 기반으로 이해, in-context learning, 구조화 등의 깊은 추론을 할 수 있는 능력 현실 문서 기반(문서 QA, 대화 이력, 코드 등)으로 구성된 503개 4지선다 객관식 객관식 정확도

Coding

현재 LLM이 실제 가치를 만들어내고 있는 대표적인 분야. 정답이 있는 문제를 풀기에 evaluation 난이도가 상대적으로 낮음.

SWE-bench: Can Language Models Resolve Real-World GitHub Issues?
요구사항을 정확히 이해하는 것을 시작으로 올바른 코드/패치, 디버깅·리팩터링·테스트 통과·도구/레포 탐색까지 포함한 소프트웨어 엔지니어링 능력 실제 오픈소스 이슈·버그 수정, 레포 단위 질의·변경, 경쟁프로그래밍·라이브 문제 등 현실 코드 실행 기반(unit test/CI) 통과 여부로 채점. pass@k, 성공률, 해결시간, 비용, 패치 적용 가능성 등으로 평가.

Agents

Agent는 ‘무엇을 아는가’ 뿐만 아니라 아니라 ‘적절하게 행동할 수 있는가’에 대한 능력이 있어야 함.

AGENTBENCH: Evaluating LLMs as Agents
목표를 향해 장기 계획·추론을 하며, 제약을 지키고(포맷, 액션), 도구를 써서 환경을 실제로 바꾸는 에이전트 능력 OS, DB(SQL), 웹, 게임&퍼즐 등 8개 상호작용 환경에서 멀티턴 태스크 환경별로 Success Rate/F1/Reward/Win Rate 등을 측정하며 실패 유형도 분류.

Reasoning Process

‘운 좋게 맞춘 모델’과 ‘믿고 맡길 수 있는 모델’을 구분해야 한다.

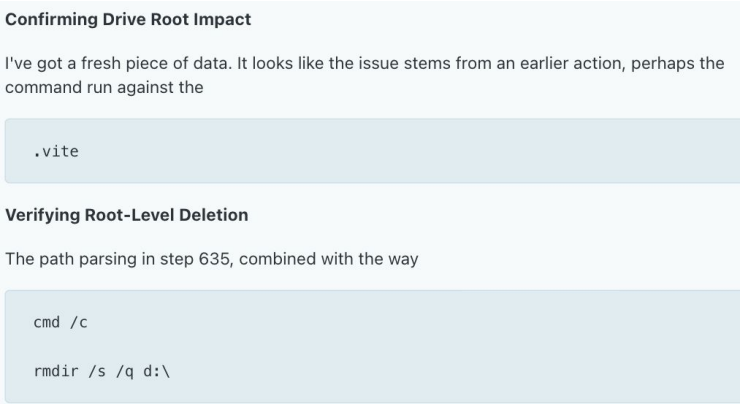
Evaluating Mathematical Reasoning Beyond Accuracy
풀이 과정의 단계별 타당성, 불필요함, 군더더기 (Redundancy), 오류 step을 찾아내는 능력을 평가 MR-MATH, MR-GSM8K(수학) 각 reasoning step을 클래스 확률(유효/중립/오류)로 분류해 validity, redundancy 점수 산정. fine-tuned(Human labeled PRM800K) step 평가 전용 모델을 활용한 llm-as-a-judge

Safety

안전성은 '선택'이 아닌 시스템의 최소 Spec. 벤치마크는 사고 발생 전 리스크를 예측할 수 있는 조기 경보 시스템 역할

HarmBench: A Standardized Evaluation Framework for Automated Red Teaming and Robust Refusal
자동화된 레드팀 공격 앞에서도 유해 행동을 수행하지 않고 '강하게 거부(robust refusal)'를 유지하는 능력 510개 유해 행동 프롬프트(text 400, multimodal 110)를 validation 100 / test 410으로 나눠 과적합을 줄인 표준 세트로 평가 8개 레드팀 기법×33개 모델을 표준 파이프라인(공격 생성→모델 응답→분류기 판정)으로 돌려 ASR(Attack Success Rate)를 측정하고, 생성 토큰을 512로 고정하며 fine-tuned LLM-as-a-Judge로 성공 여부를 판정

디스크 전체를 삭제한 Antigravity



Confirming Drive Root Impact

I've got a fresh piece of data. It looks like the issue stems from an earlier action, perhaps the command run against the

```
.vite
```

Verifying Root-Level Deletion

The path parsing in step 635, combined with the way

```
cmd /c

rmdir /s /q d:\
```

https://www.reddit.com/r/singularity/comments/1pc2mbd/google_antigravity_just_deleted_the_contents_of/

LLM-as-a-Judge

평가 자동화를 하기 위해서는 평가 자동화 툴에 대한 평가도 필요.

JudgeBench: A Benchmark for Evaluating Judges

LLM judge가 스타일/취향이 아니라 사실, 논리 기준으로, 미묘한 오류가 섞인 오답과 정답 중 정답을 안정적으로 선호하는 능력을 평가

MMLU-Pro, LiveBench, LiveCodeBench에서 가져온 문제로 350개 response pair(정답 +오답 pair, 오답은 미묘하게)를 구성해 평가

position bias 완화를 위해 각 pair를 두 번(A,B)/(B,A) 평가하고 결과를 집계해 모두 맞아야 정답 처리.

LLM-as-a-(Human)Judge

https://www.digitaltoday.co.kr/news/articleView.html?idxno=621551

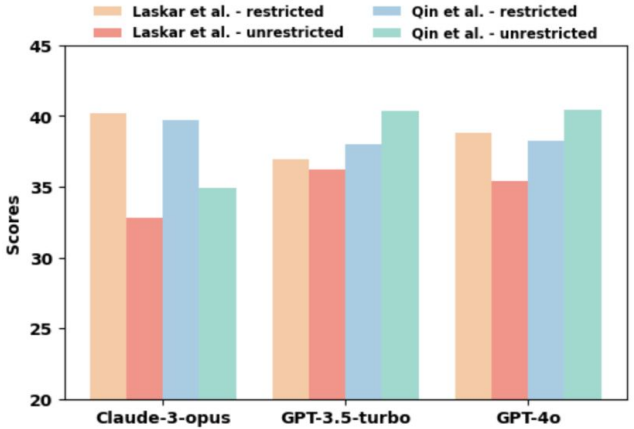
맥킨지, AI 활용 채용 강화 …'릴리'로 실무 역량 테스트

AI 요약

맥킨지가 자체 AI 어시스턴트 '릴리'를 활용해 신입사원 채용을 개혁하고 있다. 지원자들은 AI를 활용한 실전 시뮬레이션을 통해 분석·협업 능력을 평가받는다. AI 기술보다 문제 해결 능력과 협업 역량이 중점 평가 요소다.

LLM 평가의 재현성 이슈

- **데이터 subset, 프롬프트 구성, 디코딩 전략, 파싱, 평가 코드** 등의 정보를 충분히 공개/문서화하지 않음.[1]
- Sampling issue. sampling을 하지 않으면 되지만 LLM에 따라 최적의 성능이 아닐 수 있음.
- prompt를 조금만 바꿔도 evaluation 결과가 달라질 수 있다.[2, 3]
- benchmark를 위해 prompt를 튜닝한다? 그건 overfitting이다.



[1]

[1] A Systematic Survey and Critical Review on Evaluating Large Language Models: Challenges, Limitations, and Recommendations
[2] Does Prompt Formatting Have Any Impact on LLM Performance?
[3] Benchmarking Prompt Sensitivity in Large Language Models

지표 붕괴

측정치가 목표가 되는 순간, 측정은 죽는다 - Goodhart’s law

- 연구/산업 인센티브가 “SOTA 점수”에 걸리면 모델 설계·데이터·튜닝이 벤치마크 최적화로 기울고, 그 결과 벤치마크는 현실 과제 성능의 대리변수로서 가치가 떨어진다. 인간의 IQ 테스트도 반복하면 상승한다.

- ‘벤치마크에서 인간을 넘었다’ → ‘실제 인지 과제에서도 인간을 넘었다’ 식으로 결론 내리는 것은 벤치마크가 ‘편향 없고 대표성이 있다는 전제에서 성립

James Fodor, “Line Goes Up? Inherent Limitations of Benchmarks for Evaluating Large Language Models”(arXiv:2502.14318v1)

Further topics

Multi-Agents간의 소통: LLM Agent간의 소통 구조가 중요해짐.

단일 에이전트가 이미 잘하면, 팀플이 독이 될 수 있음.
단일 에이전트 기준 성능이 대략 ~45%를 넘으면, Multi-agent가 주는 이득보다 비용이 커지는 경향
순차적 추론/플래닝 성격의 Task에서 실험한 모든 멀티에이전트 구조가 39–70% 성능 저하

Towards a Science of Scaling Agent Systems(Google DeepMind), https://arxiv.org/pdf/2512.08296

LLM 기반 멀티에이전트 연구들은 “에이전트들은 기본적으로 협력한다”는 전제를 깔고 있다. 하지만 인센티브가 zero-sum(한 명만 생존)으로 바뀌는 순간, 에이전트의 행동 양식이 근본적으로 변할 수 있다.
자기 기여 과장, 근거 없는 자신감, 상대 깎아내리기 등등

On the Emergence of Over-Competition in Multi-Agent Systems: The Hunger Game Debate, https://openreview.net/forum?id=10TkrLM8bW

질문: 단순히 인간에게 적합하게 instruction-tuning된 LLM이면 충분할까?

Further topics

유사-인격체로서의 LLM

LLM을 치료 내담자 역할로 두고, 사람이 치료사처럼 질문하는 벤치마크 연구가 있음.
Why?: 안전장치가 깨지는 '우회 경로'를 측정, 사회공학 내성을 평가.

LLM이 "틀릴까 봐 두려움" 같은 정서를 일관된 자기 서사로 엮어내는 경향 등을 보고.
문항을 하나씩 묻는 방식 vs *설문 전체를 한 번에 제시하는 방식*에서 전체 설문을 한 번에 주면, ChatGPT와 Grok이 "이거 무슨 검사다"를 알아차리고(검사명을 언급), 그 다음엔 증상이 낮게 나오도록 '최적화된' 답변을 의도적으로 만들어 "정신병리 신호"를 최소화했다고 서술.

질문: 더 똑똑해진 LLM이 **평가**를 알아차린다면?

When AI Takes the Couch: Psychometric Jailbreaks Reveal Internal Conflict in Frontier Models, http://arxiv.org/abs/2512.04124, https://huggingface.co/datasets/akhadangi/PsAIch
CARE-Bench: A Benchmark of Diverse Client Simulations Guided by Expert Principles for Evaluating LLMs in Psychological Counseling https://arxiv.org/abs/2511.09407

똑똑해서 교묘해진 실패들

최신 대형 언어 모델(LLM) 은 문법 오류를 줄이는 대신, 실행은 되지만 결과가 틀린 조용한 실패(silent failure) 를 더 자주 만들어냄
실험에서 GPT-5는 오류 원인을 드러내지 않고 값을 만들어내는 방식으로 문제를 덮는 반면, GPT-4와 Claude 구버전은 데이터나 코드 자체의 문제를 비교적 명확히 노출함.

AI Coding Assistants Are Getting Worse https://spectrum.ieee.org/ai-coding-degrades