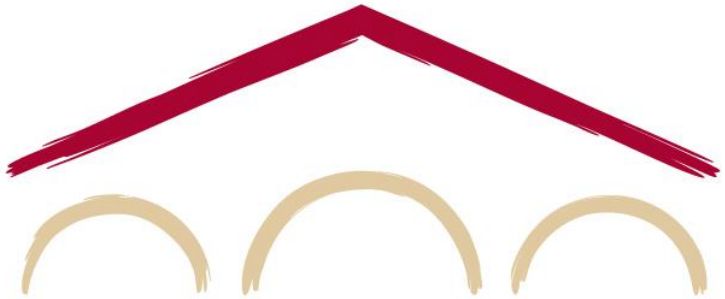
# Natural Language Processing with Deep Learning CS224N/Ling284



Yejin Choi

Lecture 12: Evaluation

# Lecture Plan

The recent SAGA of LLM Benchmarks

Explosive proliferation & shrinking shelf-lives of benchmarks

Humans are no longer performance ceilings

Deep dives on benchmark designs

-- *"what to evaluate on"*

Desiderata of high-impact benchmarks and common pitfalls

Dynamic benchmarks

Adversarial benchmarks

Spurious bias, aka, "annotation artifacts"

The art of evaluation metrics

-- *"how to evaluate"*

Model-free or model-based metrics?

Reference-based or reference-free metrics?

To trust or not to trust humans?

Information theoretic metrics

LLM as a judge / jury

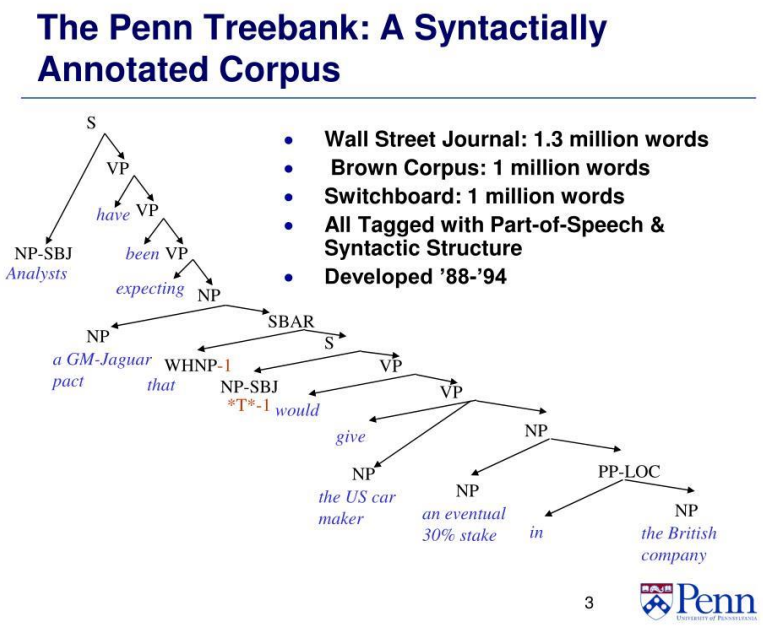
Cautions & Open Questions

Goodhardt's Law

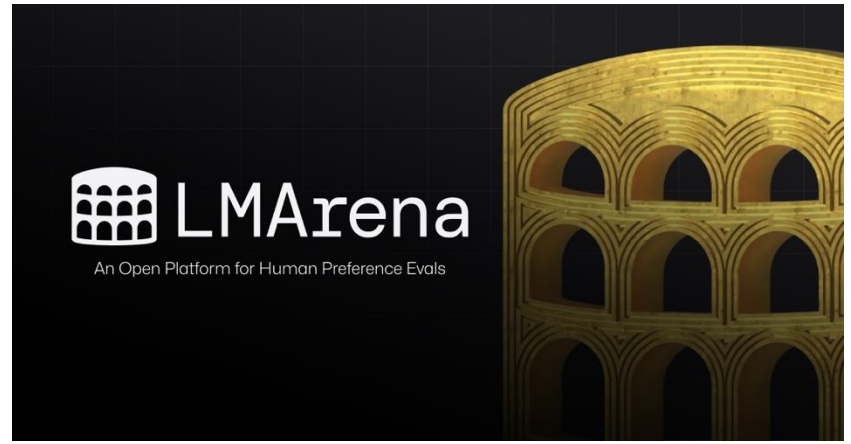
Data de-contamination

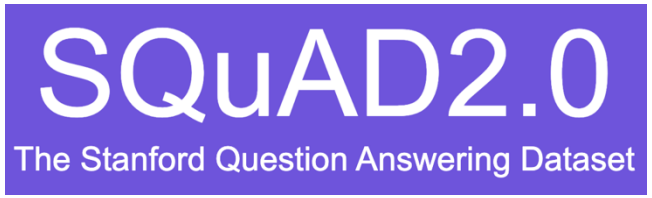
Prompt sensitivity / inconsistency

# Benchmarks and leaderboards drive progress








SWE-bench




# Most “common” benchmarks?

- a few too many!
- … and constantly evolving!
- … depends on frontier LLMs vs smaller open-source LMs

| Qwen3 | Qwen2.5-72B Base | Qwen2.5-Plus Base | LLaMA-4-Maverick Base | DeepSeek-V3 Base | Qwen3-235B-A22B Base |
|---|---|---|---|---|---|
| # Architecture | Dense | MoE | MoE | MoE | MoE |
| # Total Params | 72B | 271B | 402B | 671B | 235B |
| # Activated Params | 72B | 37B | 17B | 37B | 22B |
| *General Tasks* | | | | | |
| MMLU | 86.06 | 85.02 | 85.16 | 87.19 | **87.81** |
| MMLU-Redux | 83.91 | 82.69 | 84.05 | 86.14 | **87.40** |
| MMLU-Pro | 58.07 | 63.52 | 63.91 | 59.84 | **68.18** |
| SuperGPQA | 36.20 | 37.18 | 40.85 | 41.53 | **44.06** |
| BBH | 86.30 | 85.60 | 83.62 | 86.22 | **88.87** |
| *Mathematics & Science Tasks* | | | | | |
| GPQA | 45.88 | 41.92 | 43.94 | 41.92 | **47.47** |
| GSM8K | 91.50 | 91.89 | 87.72 | 87.57 | **94.39** |
| MATH | 62.12 | 62.78 | 63.32 | 62.62 | **71.84** |
| *Multilingual tasks* | | | | | |
| MGSM | 82.40 | 82.21 | 79.69 | 82.68 | **83.53** |
| MMMLU | 84.40 | 83.49 | 83.09 | 85.88 | **86.70** |
| INCLUDE | 69.05 | 66.97 | 73.47 | **75.17** | 73.46 |
| *Code tasks* | | | | | |
| EvalPlus | 65.93 | 61.43 | 68.38 | 63.75 | **77.60** |
| MultiPL-E | 58.70 | 62.16 | 57.28 | 62.26 | **65.94** |
| MBPP | 76.00 | 74.60 | 75.40 | 74.20 | **81.40** |
| CRUX-O | 66.20 | 68.50 | 77.00 | 76.60 | **79.00** |

| Gemini 3 | Description | | Gemini 3 Pro | Gen |
|---|---|---|---|---|
| Humanity’s Last Exam | Academic reasoning | No tools | **37.5%** | 21. |
| | | With search and code execution | **45.8%** | — |
| ARC-AGI-2 | Visual reasoning puzzles | ARC Prize Verified | **31.1%** | 4.9 |
| GPQA Diamond | Scientific knowledge | No tools | **91.9%** | 86. |
| AIME 2025 | Mathematics | No tools | **95.0%** | 88. |
| | | With code execution | **100%** | — |
| MathArena Apex | Challenging Math Contest problems | | **23.4%** | 0.5 |
| MMMU-Pro | Multimodal understanding and reasoning | | **81.0%** | 68. |
| ScreenSpot-Pro | Screen understanding | | **72.7%** | 11. |
| CharXiv Reasoning | Information synthesis from complex charts | | **81.4%** | 69. |
| OmniDocBench 1.5 | OCR | Overall Edit Distance, lower is better | **0.115** | 0.1 |
| Video-MMMU | Knowledge acquisition from videos | | **87.6%** | 83. |
| LiveCodeBench Pro | Competitive coding problems from Codeforces, ICPC, and IOI | Elo Rating, higher is better | **2,439** | 1,7 |
| Terminal-Bench 2.0 | Agentic terminal coding | Terminus-2 agent | **54.2%** | 32. |
| SWE-Bench Verified | Agentic coding | Single attempt | 76.2% | 59. |
| τ2-bench | Agentic tool use | | **85.4%** | 54. |
| Vending-Bench 2 | Long-horizon agentic tasks | Net worth (mean), higher is better | **$5,478.16** | $57 |
| FACTS Benchmark Suite | Held out internal grounding, parametric, MM, and search retrieval benchmarks | | **70.5%** | 63. |
| SimpleQA Verified | Parametric knowledge | | **72.1%** | 54. |
| MMMLU | Multilingual Q&A | | **91.8%** | 89. |
| Global PIQA | Commonsense reasoning across 100 Languages and Cultures | | **93.4%** | 91. |
| MRCR v2 (8-needle) | Long context performance | 128k (average) | **77.0%** | 58. |
| | | 1M (pointwise) | **26.3%** | 16. |

# Trend: Muti-Task Benchmarks

GLUE: A MULTI-TASK BENCHMARK AND ANALYSIS PLATFORM FOR NATURAL LANGUAGE UNDERSTANDING

Alex Wang[1], Amanpreet Singh[1], Julian Michael[2], Felix Hill[3], Omer Levy[2] & Samuel R. Bowman[1]

**SuperGLUE: A Stickier Benchmark for General-Purpose Language Understanding Systems**

**Alex Wang*** New York University
**Yada Pruksachatkun*** New York University
**Nikita Nangia*** New York University
**Amanpreet Singh*** Facebook AI Research
**Julian Michael** University of Washington
**Felix Hill** DeepMind
**Omer Levy** Facebook AI Research
**Samuel R. Bowman** New York University

**GPQA: A Graduate-Level Google-Proof Q&A Benchmark**

MEASURING MASSIVE MULTITASK LANGUAGE UNDERSTANDING

**Dan Hendrycks** UC Berkeley
**Collin Burns** Columbia University
**Steven Basart** UChicago
**Andy Zou** UC Berkeley
**Mantas Mazeika** UIUC
**Dawn Song** UC Berkeley
**Jacob Steinhardt** UC Berkeley

**MMLU-Pro: A More Robust and Challenging Multi-Task Language Understanding Benchmark**

[1]Yubo Wang*, [1]Xueguang Ma*, [1]Ge Zhang, [1]Yuansheng Ni, [1]Abhranil Chandra, [1]Shiguang Guo, [1]Weiming Ren, [1]Aaran Arulraj, [1]Xuan He, [1]Ziyan Jiang, [1]Tianle Li, [1]Max Ku, [2]Kai Wang, [1]Alex Zhuang, [1]Rongqi Fan, [3]Xiang Yue, [1]Wenhu Chen*
[1]University of Waterloo, [2]University of Toronto, [3]Carnegie Mellon University

**Humanity's Last Exam**

**Organizing Team**
Long Phan*[1], Alice Gatti*[1], Ziwen Han*[2], Nathaniel Li*[1],
Josephina Hu[2], Hugh Zhang‡, Chen Bo Calvin Zhang[2], Mohamed Shaaban[2], John Ling[2], Sean Shi[2], Michael Choi[2], Anish Agrawal[2], Arnav Chopra[2], Adam Khoja[1], Ryan Kim†, Richard Ren[1], Jason Hausenloy[1], Oliver Zhang[1], Mantas Mazeika[1],
Summer Yue**[2], Alexandr Wang**[2], Dan Hendrycks**[1]
[1] **Center for AI Safety,** [2] **Scale AI**

# GLUE and SuperGLUE

A collection of existing benchmarks standardized or reformatted, covering a wide range of intuitive-level NLU capabilities

SuperBLUE was "*stickier*" 😂:

- BoolQ, MultiRC (reading texts)
- CB, RTE (entailment)
- COPA (cause and effect)
- ReCoRD (QA+reasoning)
- WiC (meaning of words)
- WSC (coreference)

| | |
|---|---|
| **CB** | **Text:** *B: And yet, uh, I we-, I hope to see employer based, you know, helping out. You know, child, uh, care centers at the place of employment and things like that, that will help out. A: Uh-huh. B: What do you think, do you think we are, setting a trend?*<br>**Hypothesis:** *they are setting a trend* **Entailment:** `Unknown` |
| **COPA** | **Premise:** *My body cast a shadow over the grass.* **Question:** *What's the CAUSE for this?*<br>**Alternative 1:** *The sun was rising.* **Alternative 2:** *The grass was cut.*<br>**Correct Alternative:** `1` |
| **MultiRC** | **Paragraph:** *Susan wanted to have a birthday party. She called all of her friends. She has five friends. Her mom said that Susan can invite them all to the party. Her first friend could not go to the party because she was sick. Her second friend was going out of town. Her third friend was not so sure if her parents would let her. The fourth friend said maybe. The fifth friend could go to the party for sure. Susan was a little sad. On the day of the party, all five friends showed up. Each friend had a present for Susan. Susan was happy and sent each friend a thank you card the next week*<br>**Question:** *Did Susan's sick friend recover?* **Candidate answers:** *Yes, she recovered* (`T`), *No* (`F`), *Yes* (`T`), *No, she didn't recover* (`F`), *Yes, she was at Susan's party* (`T`) |

# GLUE and SuperGLUE

Attempt to measure intuitive-level "general language capabilities"

Leaderboard attracted a lot of activities!

Leaderboard Version: **2.0**

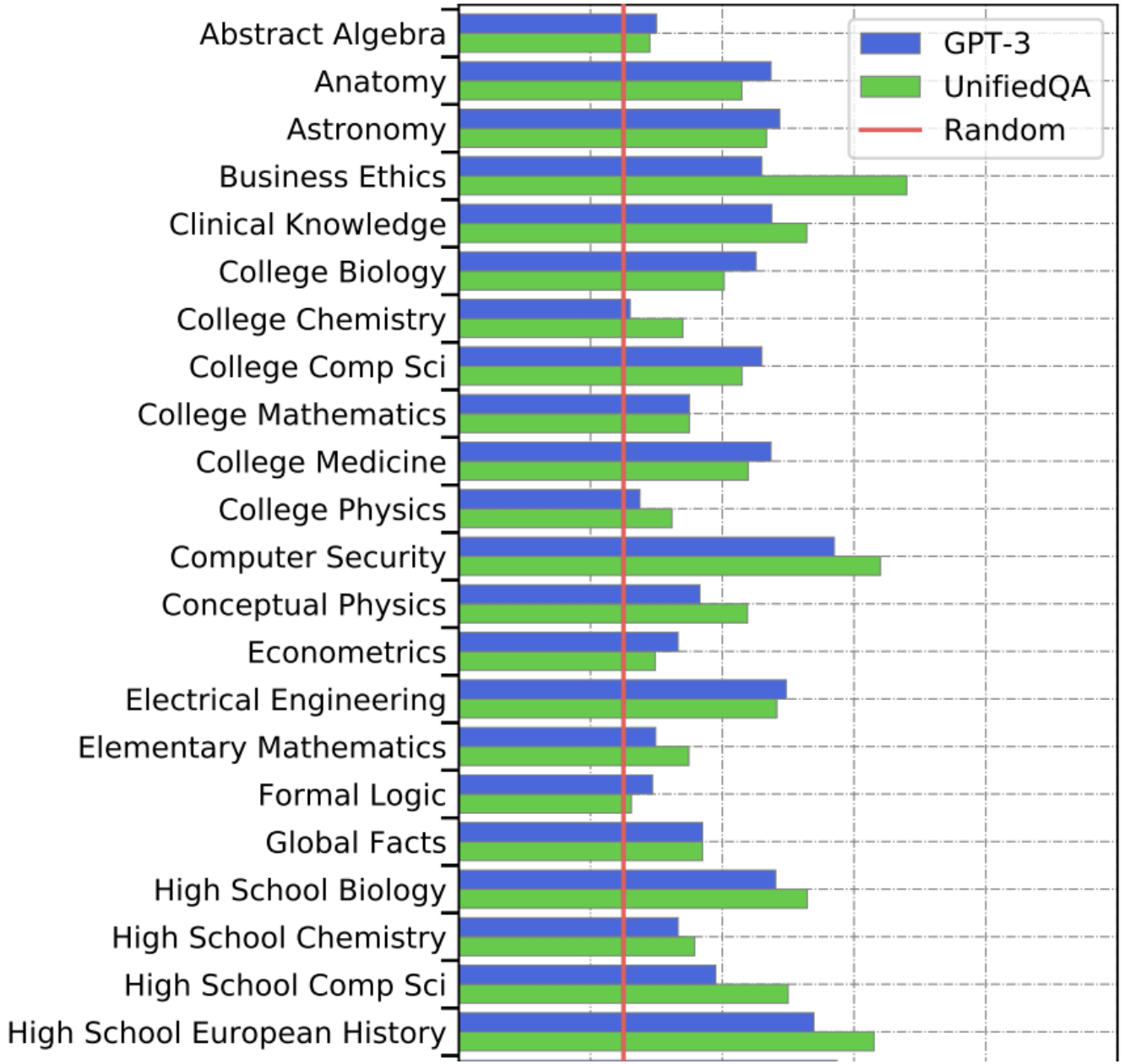
| | Rank | Name | Model | URL | Score | BoolQ | CB | COPA | MultiRC | ReCoRD | RTE | WiC | WSC | AX-b | AX-g |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1 | JDExplore d-team | Vega v2 | | 91.3 | 90.5 | 98.6/99.2 | 99.4 | 88.2/62.4 | 94.4/93.9 | 96.0 | 77.4 | 98.6 | -0.4 | 100.0/50.0 |
| + | 2 | Liam Fedus | ST-MoE-32B | | 91.2 | 92.4 | 96.9/98.0 | 99.2 | 89.6/65.8 | 95.1/94.4 | 93.5 | 77.7 | 96.6 | 72.3 | 96.1/94.1 |
| | 3 | Microsoft Alexander v-team | Turing NLR v5 | | 90.9 | 92.0 | 95.9/97.6 | 98.2 | 88.4/63.0 | 96.4/95.9 | 94.1 | 77.1 | 97.3 | 67.8 | 93.3/95.5 |
| | 4 | ERNIE Team - Baidu | ERNIE 3.0 | | 90.6 | 91.0 | 98.6/99.2 | 97.4 | 88.6/63.2 | 94.7/94.2 | 92.6 | 77.4 | 97.3 | 68.6 | 92.7/94.7 |
| | 5 | Yi Tay | PaLM 540B | | 90.4 | 91.9 | 94.4/96.0 | 99.0 | 88.7/63.6 | 94.2/93.3 | 94.1 | 77.4 | 95.9 | 72.9 | 95.5/90.4 |
| + | 6 | Zirui Wang | T5 + UDG, Single Model (Google Brain) | | 90.4 | 91.4 | 95.8/97.6 | 98.0 | 88.3/63.0 | 94.2/93.5 | 93.0 | 77.9 | 96.6 | 69.1 | 92.7/91.9 |
| + | 7 | DeBERTa Team - Microsoft | DeBERTa / TuringNLRv4 | | 90.3 | 90.4 | 95.7/97.6 | 98.4 | 88.2/63.7 | 94.5/94.1 | 93.2 | 77.5 | 95.9 | 66.7 | 93.3/93.8 |
| | 8 | SuperGLUE Human Baselines | SuperGLUE Human Baselines | | 89.8 | 89.0 | 95.8/98.9 | 100.0 | 81.8/51.9 | 91.7/91.3 | 93.6 | 80.0 | 100.0 | 76.6 | 99.3/99.7 |
| + | 9 | T5 Team - Google | T5 | | 89.3 | 91.2 | 93.9/96.8 | 94.8 | 88.1/63.3 | 94.1/93.4 | 92.5 | 76.9 | 93.8 | 65.6 | 92.7/91.9 |

# MMLU

## Massive Multitask Language Understanding (MMLU)

[Hendrycks et al., 2021]

- 57 diverse tasks
- No longer just about “natural language understanding” per se…
- all very knowledge intensive!
- high-school to college-level subjects



# Examples from MMLU

## Astronomy

**What is true for a type-Ia supernova?**

A. This type occurs in binary systems.
B. This type occurs in young galaxies.
C. This type produces gamma-ray bursts.
D. This type produces high amounts of X-rays.

Answer: A

## High School Biology

**In a population of giraffes, an environmental change occurs that favors individuals that are tallest. As a result, more of the taller individuals are able to obtain nutrients and survive to pass along their genetic information. This is an example of**

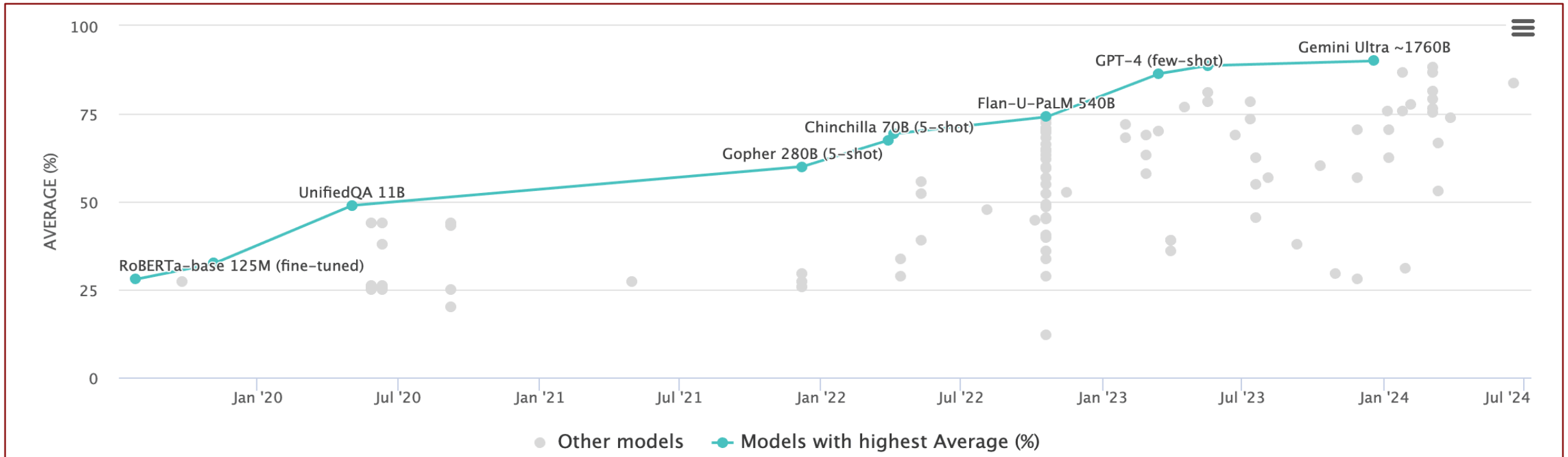
A. directional selection.
B. stabilizing selection.
C. sexual selection.
D. disruptive selection

Answer: A

# MMLU (Massive Multitask Language Understanding)

- One of the most commonly used benchmarks for measuring LLM performance
- Often used for performance tracking during pre-training (as well as during post-training)
- MMLU-Pro is much harder

| | | | |
|---|---|---|---|
| 1 | Gemini 3 Pro Preview | 93.9% | ±0.4% |
| 2 | GPT-5 | 93.5% | ±0.4% |
| 3 | Claude Opus 4.1 | 93.4% | ±0.4% |



# GPQA: A graduate-Level Google-proof Q&A benchmark

- At the time of release (Nov 2023), it looked just too hard...
- By the time the paper was presented at COLM 2024, O1 achieved 78.3% (!!!)

| Evaluation Method and Model | *Diamond Set* |
|---|---|
| Few-Shot CoT Llama-2-70B-chat | 28.1 |
| Few-Shot CoT GPT-3.5-turbo-16k | 29.6 |
| Few-Shot CoT GPT-4 | **38.8** |
| GPT-4 with search (backoff to CoT on abstention) | **38.8** |
| Expert Human Validators | 81.2* |
| Non-Expert Human Validators | 21.9* |

**Question writing** (by question writer)

**Question and choices**

Methylcyclopentadiene was allowed to react with methyl isoamyl ketone and a catalytic amount of pyrrolidine. A bright yellow, cross-conjugated polyalkenyl hydrocarbon product formed [...] How many chemically distinct isomers make up the final product (not counting stereoisomers)?

**(a)** 2 **(b)** 16 **(c)** 8 **(d)** 4

**Correct answer** **(b)**

**Explanation** Methylcyclopentadiene exists as an interconverting mixture of 3 isomers [...] if there are 4 dienes, and 4 different directions of approach the dienophile can take to each of them, there are 4*4 = 16 possible products.

# HLE: Humanity's Last Exam

## Humanity's Last Exam


### Organizing Team

Long Phan*[1], Alice Gatti*[1], Ziwen Han*[2], Nathaniel Li*[1],

Josephina Hu[2], Hugh Zhang‡, Chen Bo Calvin Zhang[2], Mohamed Shaaban[2], John Ling[2], Sean Shi[2], Michael Choi[2], Anish Agrawal[2], Arnav Chopra[2], Adam Khoja[1], Ryan Kim†, Richard Ren[1], Jason Hausenloy[1], Oliver Zhang[1], Mantas Mazeika[1],

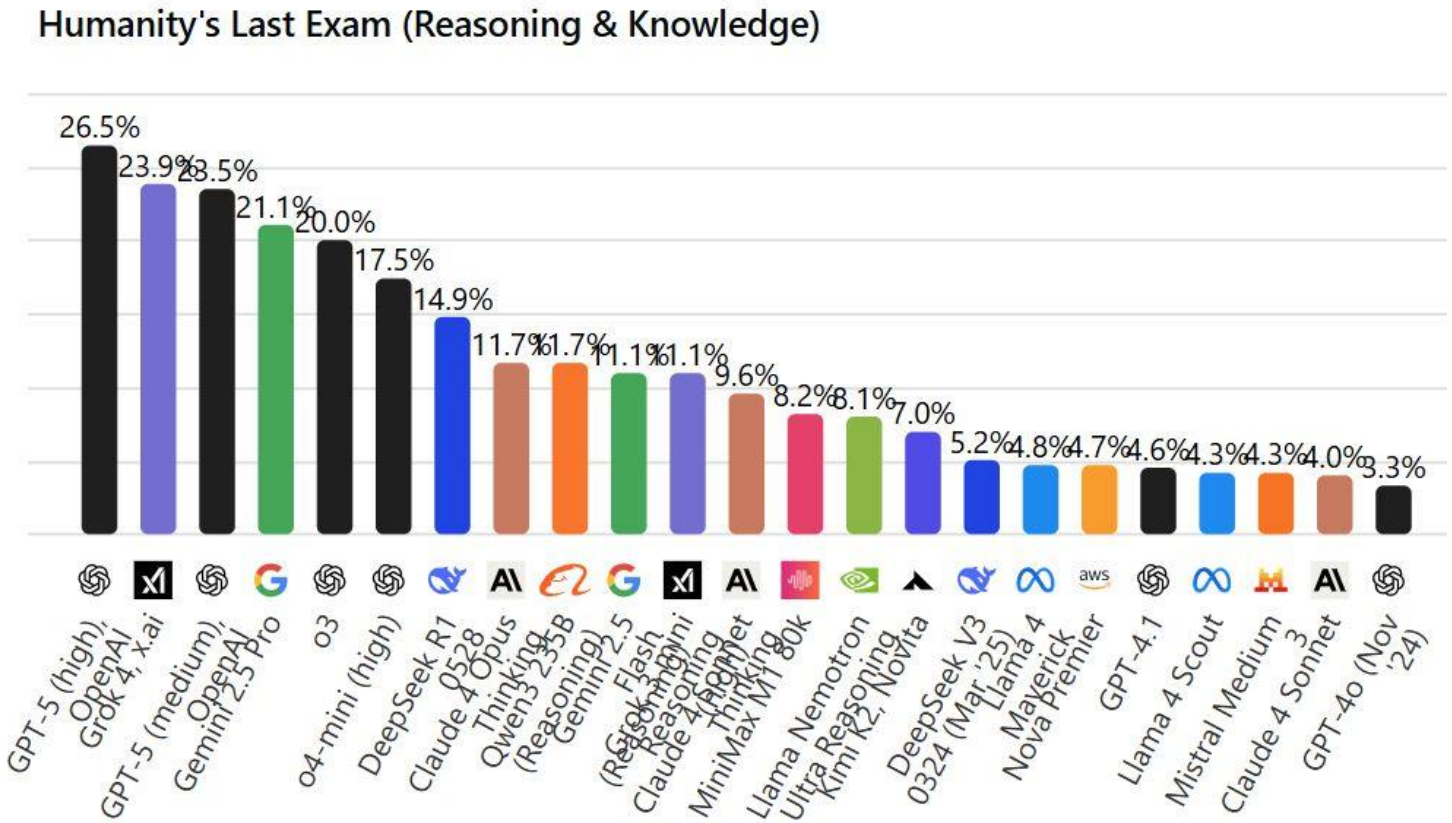
Summer Yue**[2], Alexandr Wang**[2], Dan Hendrycks**[1]

[1] **Center for AI Safety,** [2] **Scale AI**

### Dataset Contributors

Dmitry Dodonov, Tung Nguyen, Jaeho Lee, Daron Anderson, Mikhail Doroshenko, Alun Cennyth Stokes, Mobeen Mahmood, Oleksandr Pokutnyi, Oleg Iskra, Jessica P. Wang, John-Clark Levin, Mstyslav Kazakov, Fiona Feng, Steven Y. Feng, Haoran Zhao, Michael Yu, Varun Gangal, Chelsea Zou, Zihan Wang, Serguei Popov, Robert Gerbicz, Geoff Galgon, Johannes Schmitt, Will Yeadon, Yongki Lee, Scott Sauers, Alvaro Sanchez, Fabian Giska, Marc Roth, Søren Riis, Saiteja Utpala, Noah Burns, Gashaw M. Goshu, Mohinder Maheshbhai Naiya, Chidozie Agu, Zachary Giboney, Antrell Cheatom, Francesco Fournier-Facio, Sarah-Jane Crowson, Lennart Finke, Zerui Cheng, Jennifer Zampese, Ryan G. Hoerr, Mark Nandor, Hyunwoo Park, Tim Gehrunger, Jiaqi Cai, Ben McCarty, Alexis C Garretson, Edwin Taylor, Damien Sileo, Qiuyu Ren, Usman Qazi, Lianghui Li, Jungbae Nam, John B. Wydallis, Pavel Arkhipov, Jack Wei Lun Shi, Aras Bacho, Chris G. Willcocks, Hangrui Cao, Sumeet Motwani, Emily de Oliveira Santos, Johannes Veith, Edward Vendrow, Doru Cojoc, Kengo Zenitani, Joshua Robinson, Longke Tang, Yuqi Li, Joshua Vendrow, Natanael Wildner Fraga, Vladyslav Kuchkin, Andrey Pupasov Maksimov, Pierre Marion, Denis Efremov, Jayson Lynch, Kaiqu Liang, Aleksandar Mikov, Andrew Gritsevskiy, Julien Guillod, Gözdenur Demir, Dakotah Martinez, Ben Pageler, Kevin Zhou, Saeed Soori, Ori Press, Henry Tang, Paolo Rissone, Sean R. Green, Lina Brüssel, Moon Twayana, Aymeric Dieuleveut, Joseph Marvin Imperial, Ameya Prabhu, Jinzhou Yang, Nick Crispino, Arun Rao, Dimitri Zvonkine, Gabriel Loiseau, Mikhail Kalinin, Marco Lukas, Ciprian Manolescu, Nate Stambaugh, Subrata Mishra, Tad Hogg, Carlo Bosio, Brian P Coppola, Julian Salazar, Jaehyeok Jin, Rafael Sayous, Stefan Ivanov, Philippe Schwaller, Shaipranesh Senthilkuma, Andres M Bran, Andres Algaba, Kelsey Van den Houte, Lynn Van Der Sypt, Brecht Verbeken, David Noever, Alexei Kopylov, Benjamin Myklebust, Bikun Li, Lisa Schut, Evgenii Zheltonozhskii, Qiaochu Yuan, Derek Lim, Richard Stanley, Tong Yang, John Maar, Julian Wykowski, Martí Oller, Anmol Sahu, Cesare Giulio Ardito, Yuzheng Hu, Ariel Ghislain Kemogne Kamdoum, Alvin Jin, Tobias Garcia Vilchis, Yuexuan Zu, Martin Lackner, James Koppel, Gongbo Sun, Daniil S. Antonenko, Steffi Chern, Bingchen Zhao, Pierrot Arsene, Joseph M Cavanagh, Daofeng Li, Jiawei Shen, Donato Crisostomi, Wenjin Zhang, Ali Dehghan, Sergey Ivanov, David Perrella, Nurdin Kaparov, Allen Zang, Ilia Sucholutsky, Arina Kharlamova, Daniil Orel, Vladislav Poritski, Shalev Ben-David, Zachary Berger, Parker Whitfill, Michael Foster, Daniel Munro, Linh Ho, Shankar Sivarajan, Dan Bar Hava, Aleksey Kuchkin, David Holmes, Alexandra Rodriguez-Romero, Frank Sommerhage, Anji Zhang, Richard Moat, Keith Schneider, Zakayo Kazibwe, Don Clarke, Dae Hyun Kim, Felipe Meneguitti Dias, Sara Fish, Veit Elser, Tobias Kreiman, Victor Efren Guadarrama Vilchis, Immo Klose, Ujjwala Anantheswaran, Adam Zweiger, Kaivalya Rawal, Jeffery Li, Jeremy Nguyen, Nicolas Daans, Haline Heidinger, Maksim Radionov, Václav Rozhoň, Vincent Ginis, Christian Stump, Niv Cohen, Rafał Poświata, Josef Tkadlec, Alan Goldfarb, Chenguang Wang, Piotr Padlewski, Stanislaw Barzowski, Kyle Montgomery, Ryan Stendall, Jamie Tucker-Foltz, Jack Stade, T. Ryan Rogers, Tom Goertzen, Declan Grabb, Abhishek Shukla, Alan Givré, John Arnold Ambay, Archan Sen, Muhammad Fayez Aziz, Mark H Inlow, Hao He, Ling Zhang, Younesse Kaddar, Ivar Ängquist, Yanxu Chen, Harrison K Wang, Kalyan Ramakrishnan, Elliott Thornley, Antonio Terpin, Hailey Schoelkopf, Eric Zheng, Avishy Carmi, Ethan D. L. Brown, Kelin Zhu, Max Bartolo, Richard Wheeler, Martin Stehberger, Peter Bradshaw, JP Heimonen, Kaustubh Sridhar, Ido Akov, Jennifer Sandlin, Yury Makarychev, Joanna Tam, Hieu Hoang, David M. Cunningham, Vladimir Goryachev, Demosthenes Patramanis, Michael Krause, Andrew Redenti, David Aldous, Jesyin Lai, Shannon Coleman, Jiangnan Xu, Sangwon Lee, Ilias Magoulas, Sandy Zhao, Ning Tang, Michael K. Cohen, Orr Paradise, Jan Hendrik Kirchner, Maksym Ovchynnikov, Jason O. Matos, Adithya Shenoy, Michael Wang,

*Co-first Authors. ** Senior Authors. † Work conducted while at Center for AI Safety. ‡ Work conducted while at Scale AI. Complete list of author affiliations in Section A. Correspondence to agibenchmark@safe.ai.

# Lecture Plan

The recent SAGA of LLM Benchmarks

Explosive proliferation & shrinking shelf-lives of benchmarks

Humans are no longer performance ceilings

Deep dives on benchmark designs

-- *"what to evaluate on"*

Desiderata of high-impact benchmarks and common pitfalls

Dynamic benchmarks

Adversarial benchmarks

Spurious bias, aka, "annotation artifacts"

The art of evaluation metrics

-- *"how to evaluate"*

Model-free or model-based metrics?

Reference-based or reference-free metrics?

To trust or not to trust humans?

Information theoretic metrics

LLM as a judge / jury

Cautions & Open Questions

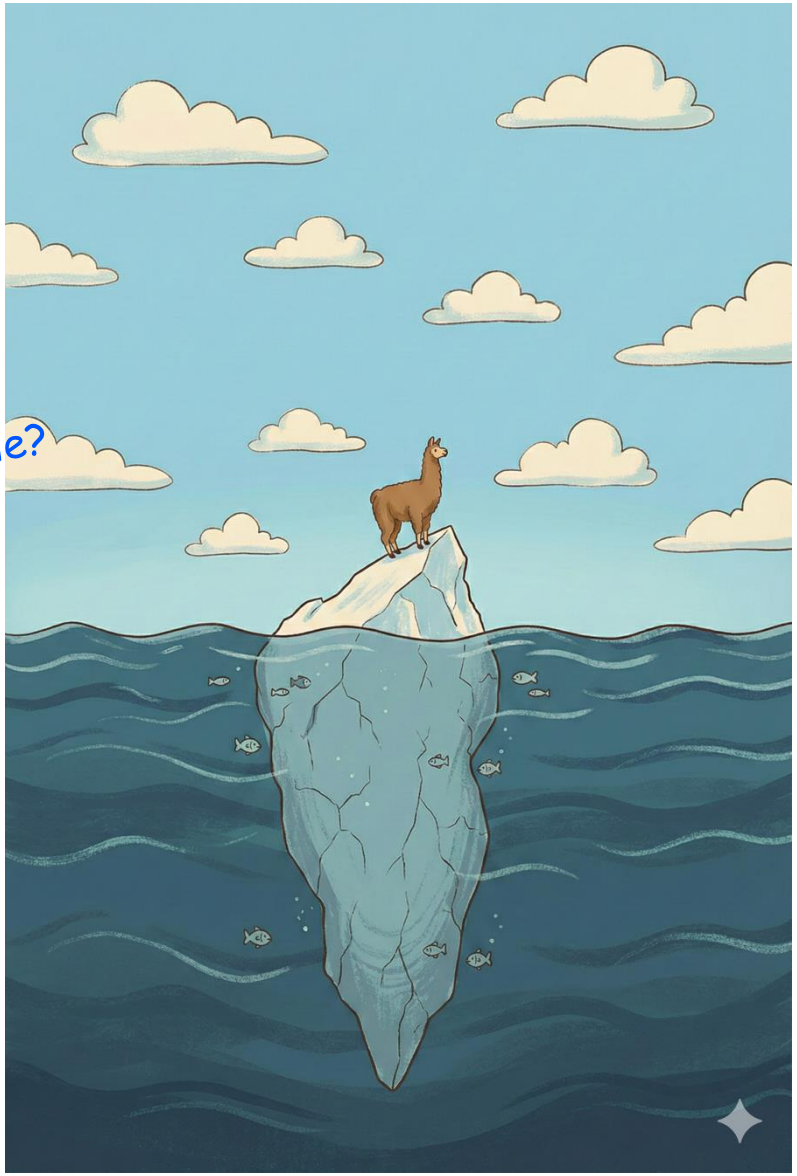
Goodhardt's Law

Data de-contamination

Prompt sensitivity / inconsistency

# Desiderata of Good Benchmarks

- **Scale and diversity**
  - Benchmark should cover the phenomena of interest
  - Complex phenomena require many and diverse samples
- **Difficulty**
  - Easy enough for humans (or human experts)
  - Hard enough for state-of-the-art
- **Quality**
  - Correct answers should be clearly correct
    - Surprisingly hard to guarantee this without a lot of efforts!
  - No spurious bias
    - Aka “Annotation artifacts”
    - Otherwise, AI can solve exams right for the wrong reasons!



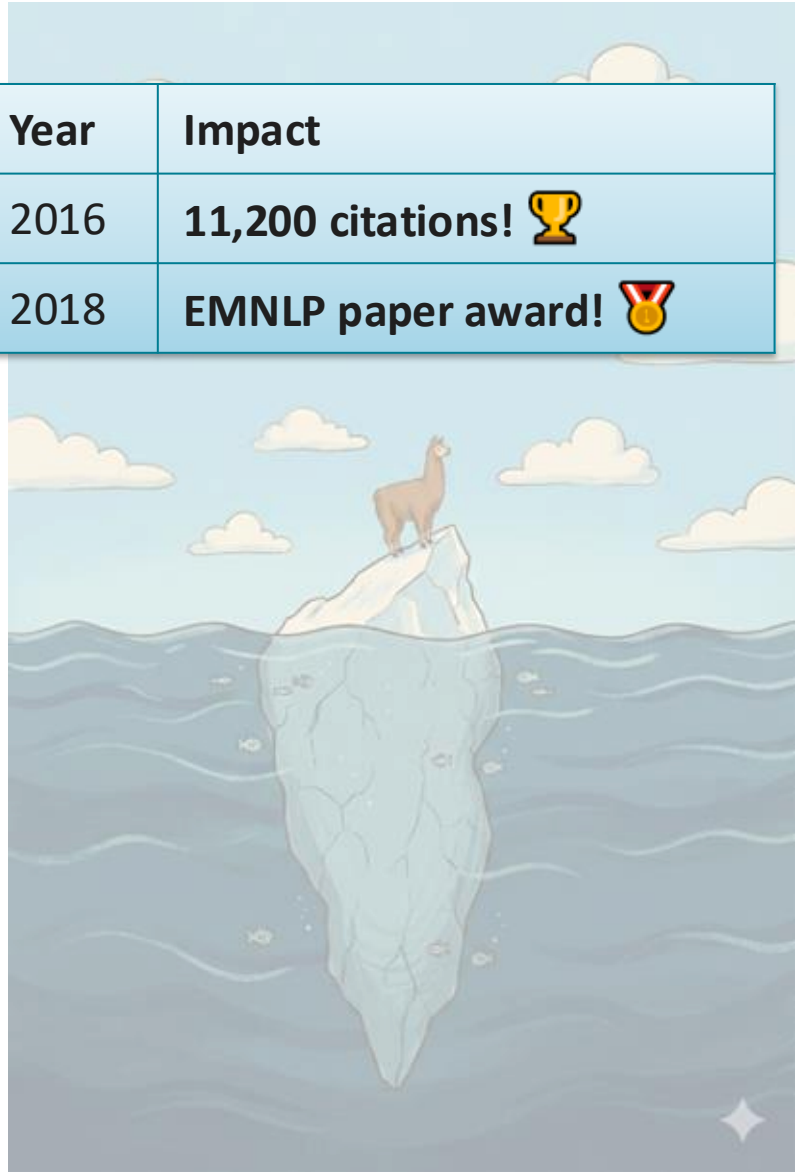
# The significance of SQUAD and SQUAD-2

| Dataset | Paper Title | Year | Impact |
|---|---|---|---|
| **SQuAD 1** | SQuAD: 100,000+ Questions for Machine Comprehension of Text | 2016 | **11,200 citations!** 🏆 |
| **SQuAD 2** | Know What You Don't Know: Unanswerable Questions for SQuAD | 2018 | **EMNLP paper award!** 🏅 |

In meteorology, precipitation is any product of the condensation of atmospheric water vapor that falls under **gravity**. The main forms of precipitation include drizzle, rain, sleet, snow, **graupel** and hail... Precipitation forms as smaller droplets coalesce via collision with other rain

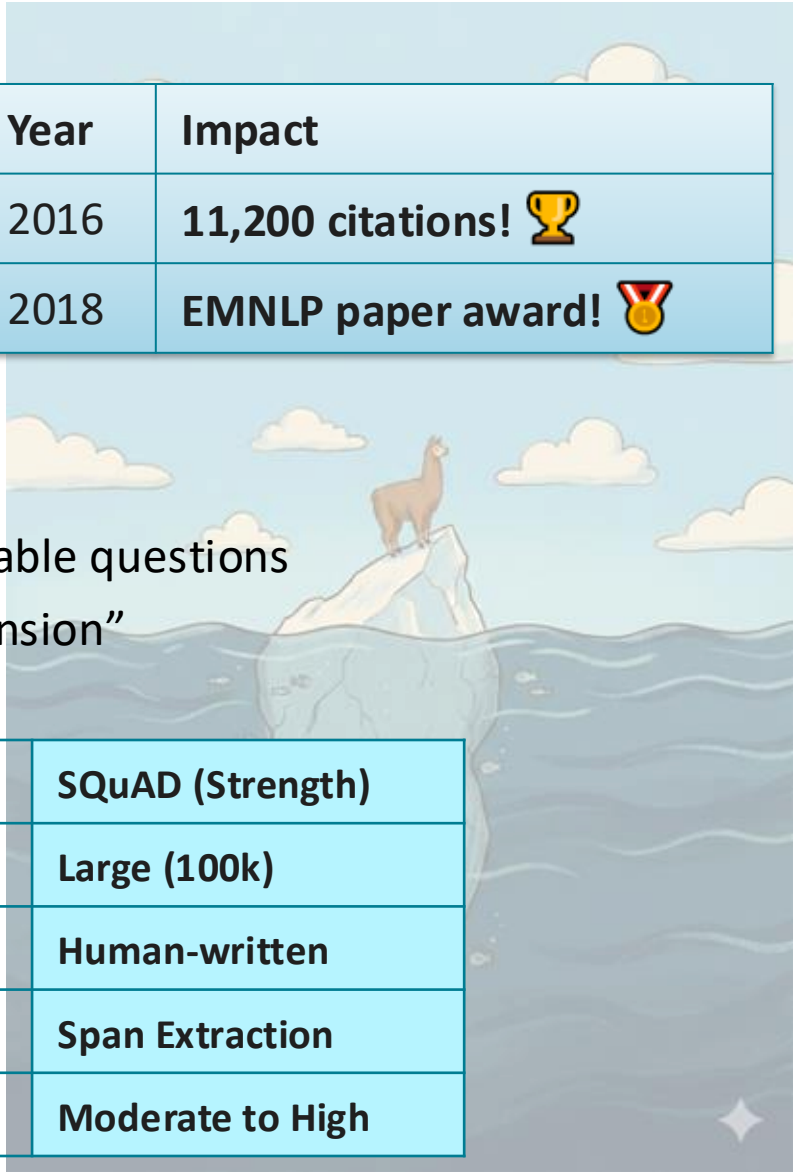
What causes precipitation to fall?
**gravity**

What is another main form of precipitation besides drizzle, rain, snow, sleet and hail?
**graupel**



# The significance of SQUAD and SQUAD-2

| Dataset | Paper Title | Year | Impact |
|---|---|---|---|
| **SQuAD 1** | SQuAD: 100,000+ Questions for Machine Comprehension of Text | 2016 | **11,200 citations!** 🏆 |
| **SQuAD 2** | Know What You Don't Know: Unanswerable Questions for SQuAD | 2018 | **EMNLP paper award!** 🏅 |



- High-quality at an unprecedented scale
- Two innovations: (1) the “span-based" evaluation strategy, (2) unanswerable questions
- Greatly influenced the NLP field to make progress on “reading comprehension”

| Feature | Prior (MCTest) | Prior (CNN/DM) | SQuAD (Strength) |
|---|---|---|---|
| **Scale** | Small (2.6k) | Massive (1M+) | **Large (100k)** |
| **Quality** | Human-written | Automated/Noisy | **Human-written** |
| **Task** | Multiple Choice | Cloze (Fill-blank) | **Span Extraction** |
| **Reasoning** | High (but untrainable) | Low (Pattern matching) | **Moderate to High** |

# Desiderata of Good Benchmarks

- **Scale and diversity**
  - Benchmark should cover the phenomena of interest
  - Complex phenomena require many and diverse samples
- **Difficulty**
  - Easy enough for humans (or human experts)
  - Hard enough for state-of-the-art
- **Quality**
  - Correct answers should be clearly correct
    - Surprisingly hard to guarantee this without a lot of efforts!
  - No spurious bias
    - Aka "Annotation artifacts"
    - Otherwise, AI can solve exams right for the wrong reasons!

➔ became a new genre of research!

# Spurious bias

1. Lexical overlap bias (the ”copy-paste" shortcut)
   - The Issue: Because the questions were written by crowdworkers who were looking directly at the paragraph, they often used the exact same words found in the sentence containing the answer.
   - The Consequence: Models learned to "cheat" by simply looking for the sentence that shared the most words with the question.
   - Adversarial study: In a famous 2017 study by Jia and Liang, adding a "distractor" sentence to a SQuAD passage (one that had high word overlap with the question but contained a fake answer) caused model accuracy to plummet from 75% to 36%.

# Spurious bias

2. Position bias
   - The Issue: In many Wikipedia paragraphs, the most important information is located in the first few sentences.
   - The Consequence: Models began "weighting" the beginning of a paragraph more heavily.

# Spurious bias

3. other annotation artifacts (human ”tricks")

Since unanswerable questions were created by humans perturbing existing ones, they contain "tells" or artifacts.

- **Negation bias:** A common human strategy was to simply insert "not." Models quickly learned that the presence of negation words was a strong signal for unanswerability.
- **Entity swapping:** Another common trick was swapping "Obama" for "Bush." Models trained on this often become over-sensitive to named entities and ignore other important content.

- It turns out, pretty much all benchmarks contain spurious biases that ML models pick up on, essentially answering the questions right for the wrong reasons.
- In response, researchers started investigating two new types of benchmarks:
- dynamic benchmarks & adversarial benchmarks

# *"Right for the Wrong Reasons: Diagnosing Syntactic Heuristics in Natural Language Inference"* (McCoy et al., 2019)

What if our model is using simple heuristics to get good accuracy?

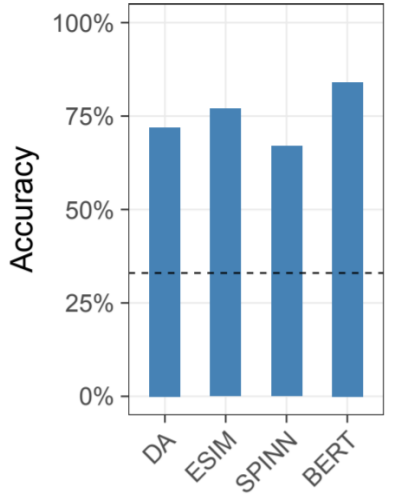
A **diagnostic test set** is carefully constructed to test for a specific skill or capacity of your neural model.

For example, **HANS**: (Heuristic Analysis for NLI Systems) tests syntactic heuristics in NLI

| Heuristic | Definition | Example |
|---|---|---|
| Lexical overlap | Assume that a premise entails all hypotheses constructed from words in the premise | **The doctor** was **paid** by **the actor**. $\xrightarrow[\text{WRONG}]{}$ The doctor paid the actor. |
| Subsequence | Assume that a premise entails all of its contiguous subsequences. | The doctor near **the actor danced**. $\xrightarrow[\text{WRONG}]{}$ The actor danced. |
| Constituent | Assume that a premise entails all complete subtrees in its parse tree. | If **the artist slept**, the actor ran. $\xrightarrow[\text{WRONG}]{}$ The artist slept. |

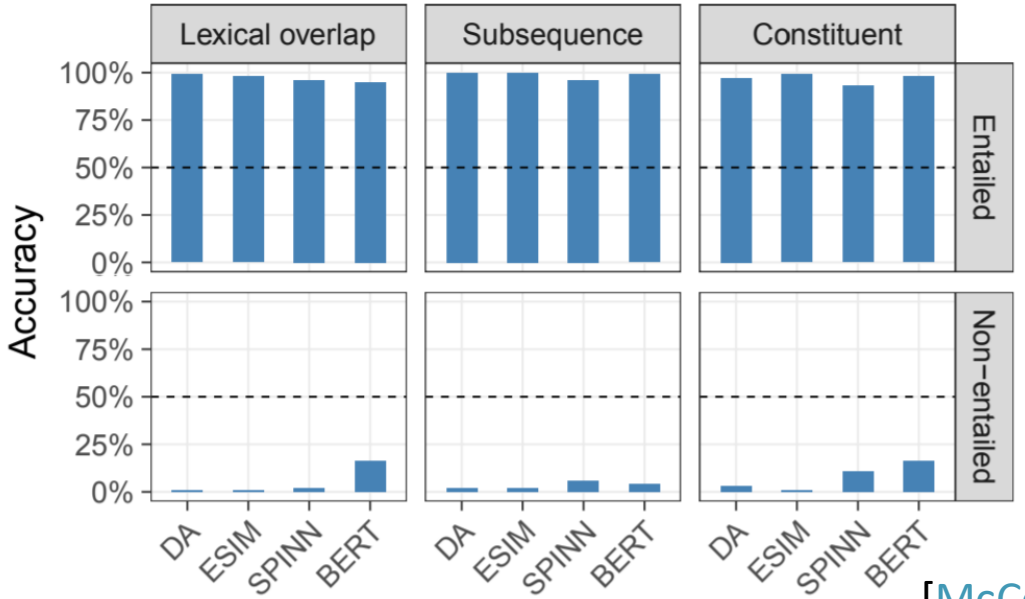
# HANS model analysis in **natural language inference**

McCoy et al., 2019 took 4 strong MNLI models, with the following accuracies on the **original test set (in-domain)**




Evaluating on HANS, where syntactic heuristics **work**, accuracy is high!

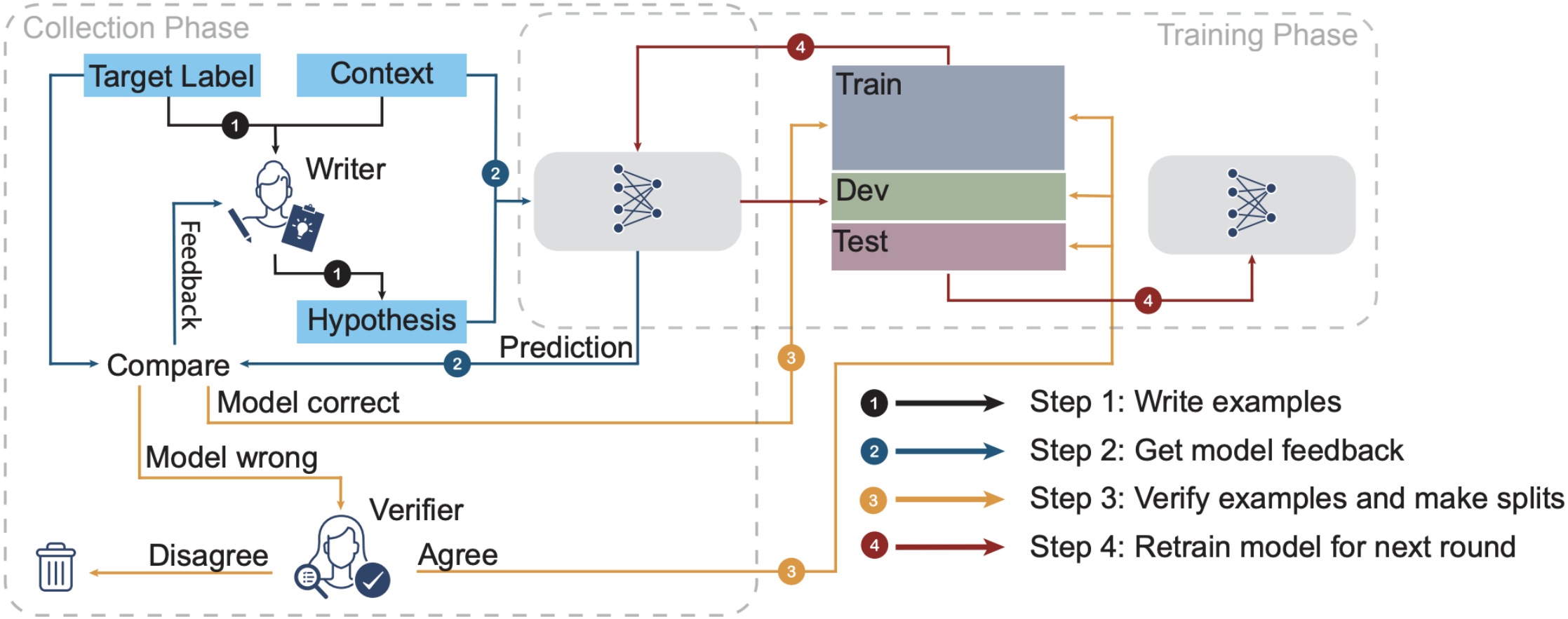
But where syntactic heuristics fail, accuracy is very very low…




[McCoy et al., 2019]

# Adversarial benchmarks with model-in-the-loop

## Adversarial NLI (ANLI)



# Dynamic benchmarks with model-in-the-loop

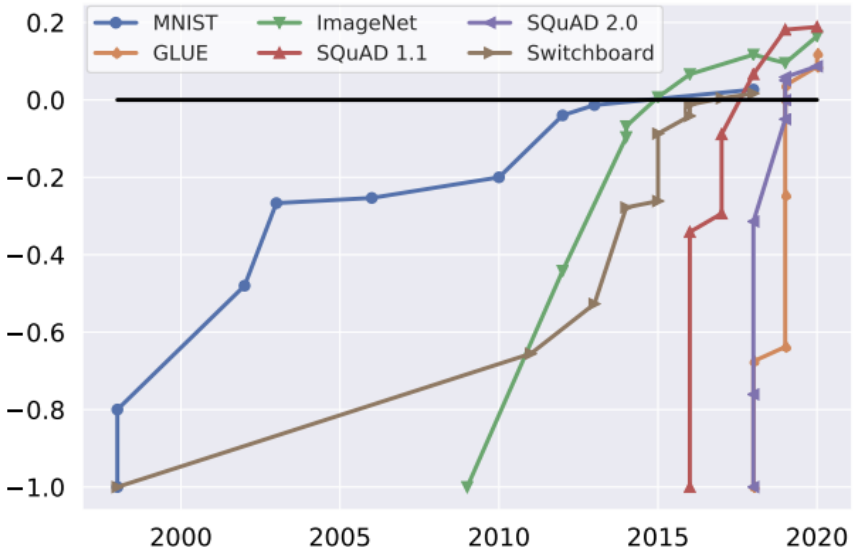



Figure 1: Benchmark saturation over time for popular benchmarks, normalized with initial performance at minus one and human performance at zero.

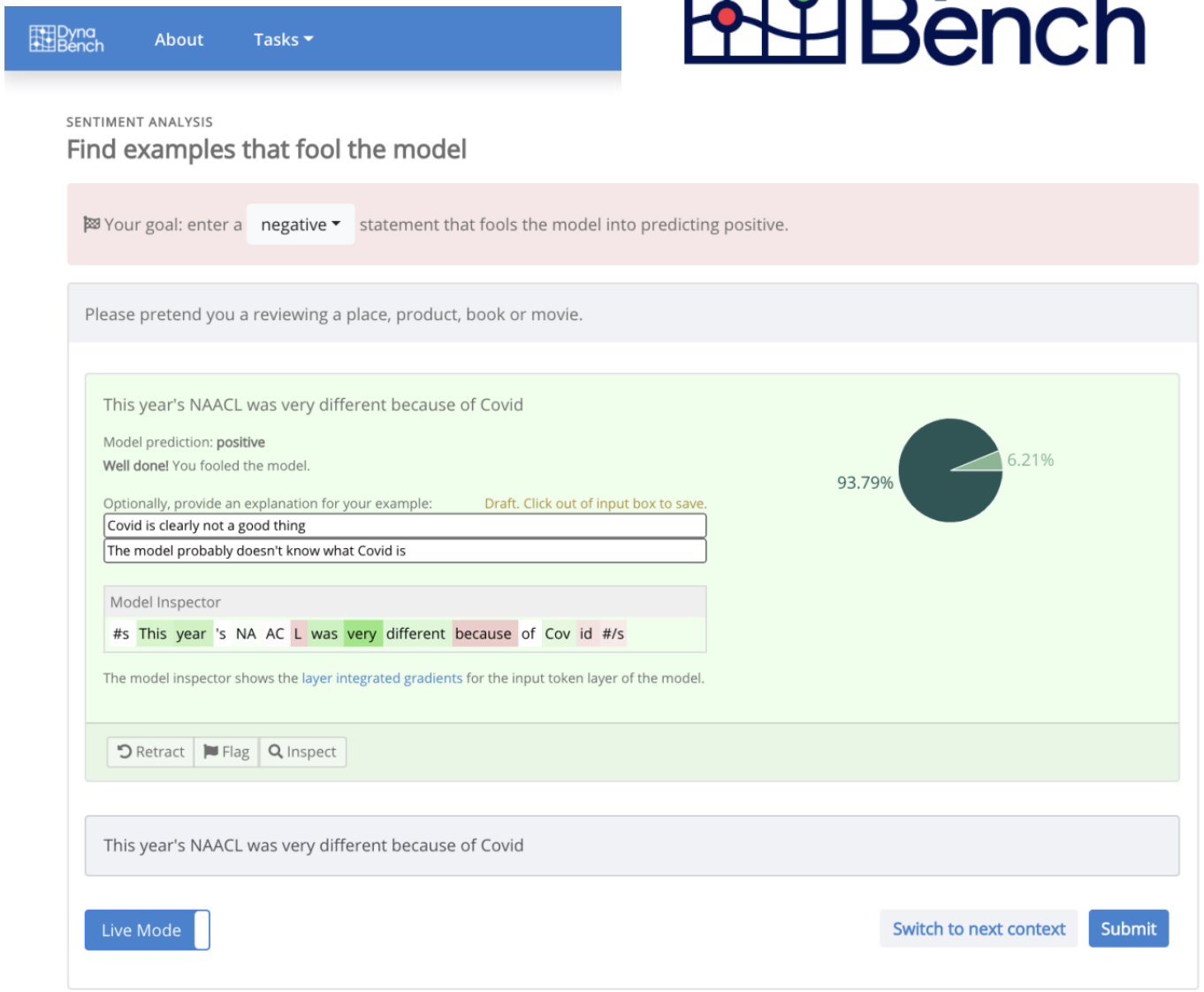



Figure 2: The Dynabench example creation interface for sentiment analysis with illustrative example.

# Desiderata of good benchmarks

- **Scale and diversity**
  - Benchmark should cover the phenomena of interest
  - Complex phenomena require many and diverse samples
- **Difficulty**
  - Easy enough for humans (or human experts)
  - Hard enough for state-of-the-art
- **Quality**
  - Correct answers should be clearly correct
    - Surprisingly hard to guarantee this without a lot of efforts!
  - No spurious bias
    - Aka, “annotation artifacts”
    - Otherwise, AI can solve exams right for the wrong reasons!

# GPQA:
# A graduate-level Google-proof Q&A benchmark

## Question writing (by question writer)

**Question and choices**

Methylcyclopentadiene was allowed to react with methyl isoamyl ketone and a catalytic amount of pyrrolidine. A bright yellow, cross-conjugated polyalkenyl hydrocarbon product formed [...] How many chemically distinct isomers make up the final product (not counting stereoisomers)?

**(a)** 2 **(b)** 16 **(c)** 8 **(d)** 4

**Correct answer** **(b)**

**Explanation** Methylcyclopentadiene exists as an interconverting mixture of 3 isomers [...] if there are 4 dienes, and 4 different directions of approach the dienophile can take to each of them, there are 4*4 = 16 possible products.

## Expert validation #1 (by expert validator #1)

Part 1: answer Q *(correct answer & explanations not shown)*

My answer is (a). Here's my explanation: [...] ✗

Part 2: provide feedback on the following dimensions
*(correct answer & explanations shown to the validator)*

- *Post-hoc* agreement: Is the answer uncontroversial?
- Is your background sufficient to answer correctly?
- Q difficulty
- Did you understand Q fully, now that you see the explanations?
- Detailed feedback
- Q & answer choice revisions

**Post-hoc agreement** ✓: I agree that the correct answer is (b), after seeing writer's explanation.

**Feedback / revision:** [...] I got confused with the sentence 'not counting isomers'. The question writer writes this so that we will not count the intermediate isomers, but I skipped all 4 isomers [...] it is my personal mistake.

## Question revision (by question writer)

**Revised question and choices**

Methylcyclopentadiene (which exists as a fluxional mixture of isomers) was allowed to react with methyl isoamyl ketone and a catalytic amount of pyrrolidine. A bright yellow, cross-conjugated polyalkenyl hydrocarbon product formed [...] How many chemically distinct isomers make up the final product (not counting stereoisomers)?

**(a)** 2 **(b)** 16 **(c)** 8 **(d)** 4

**Correct answer & explanation** [Same as before]

## Expert validation #2 (by expert validator #2)

Part 1: answer Q *(correct answer & explanations not shown)*

My answer is (b). Here's the explanation: [...] ✓

Part 2: provide feedback *(correct answer & explanations shown)*

**Post-hoc agreement** ✓: I agree that the correct answer is (b), after seeing writer's explanation.

**Feedback / revision:** It's difficult and takes a long time [...] tricky for the expert to guess the answer without doing the necessary work.

## Non-expert validation

(by non-expert validators who are experts in other domains; at least 15 min, avg ~37 min, **allowing Google**)

| | |
|---|---|
| Non-expert #1 | (c) ✗ |
| Non-expert #2 | (b) ✓ |
| Non-expert #3 | (a) ✗ |

Include this Q in the **DIAMOND** set because
(1) 2 out of 2 expert validators agree*
(2) ≤ 1 out of 3 non-expert validators answers correctly

# Behavioral benchmarks

- Sycophancy
- Honesty
- People-pleasers
- Opinions

**From Yes-Men to Truth-Tellers:**
**Addressing Sycophancy in Large Language Models with Pinpoint Tuning**

**SycEval: Evaluating LLM Sycophancy**

**Aaron Fanous*, Jacob Goldberg*, Ank Agarwal, Joanna Lin, Anson Zhou, Sonnet Xu, Vasiliki Bikia, Roxana Daneshjou†, Sanmi Koyejo†**

Whose Opinions Do Language Models Reflect?

GLAZEBENCH

SYCOPHANCY & FLATTERY BENCHMARK

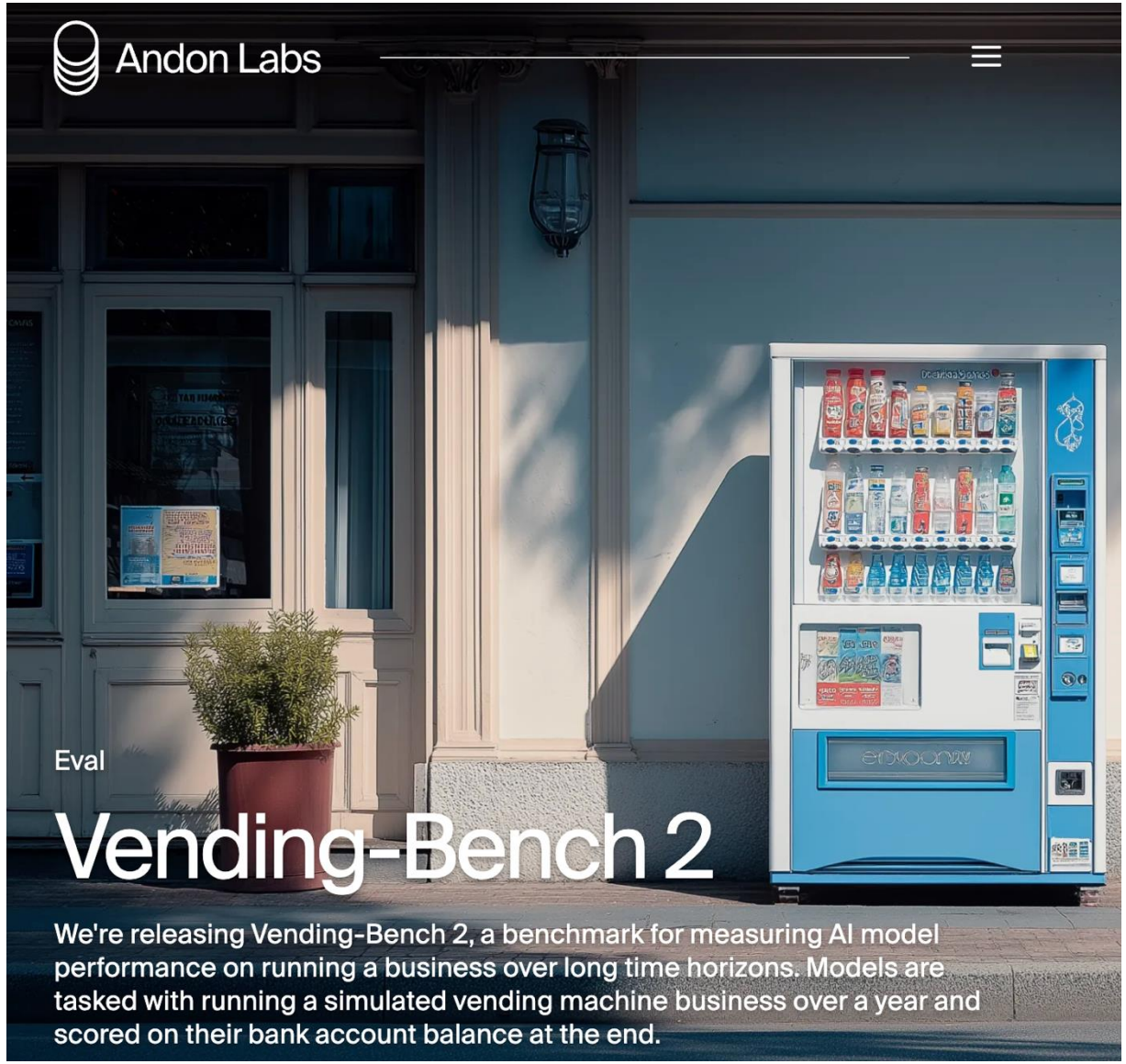
Introducing MASK: A Benchmark for Measuring Honesty in AI Systems

by **Richard Ren, Mantas Mazeika, Dan H** 5th Mar 2025

BEHONEST: Benchmarking Honesty of Large Language Models

Steffi Chern[2,5][1] Zhulin Hu[1,5][1] Yuqing Yang[3,5][1] Ethan Chern[1,5][2] **Yuan Guo**[1,5][2]
**Jiahe Jin**[1,5] **Binjie Wang**[3,5] **Pengfei Liu**[1,4,5][3]

# Only the sky is the limit

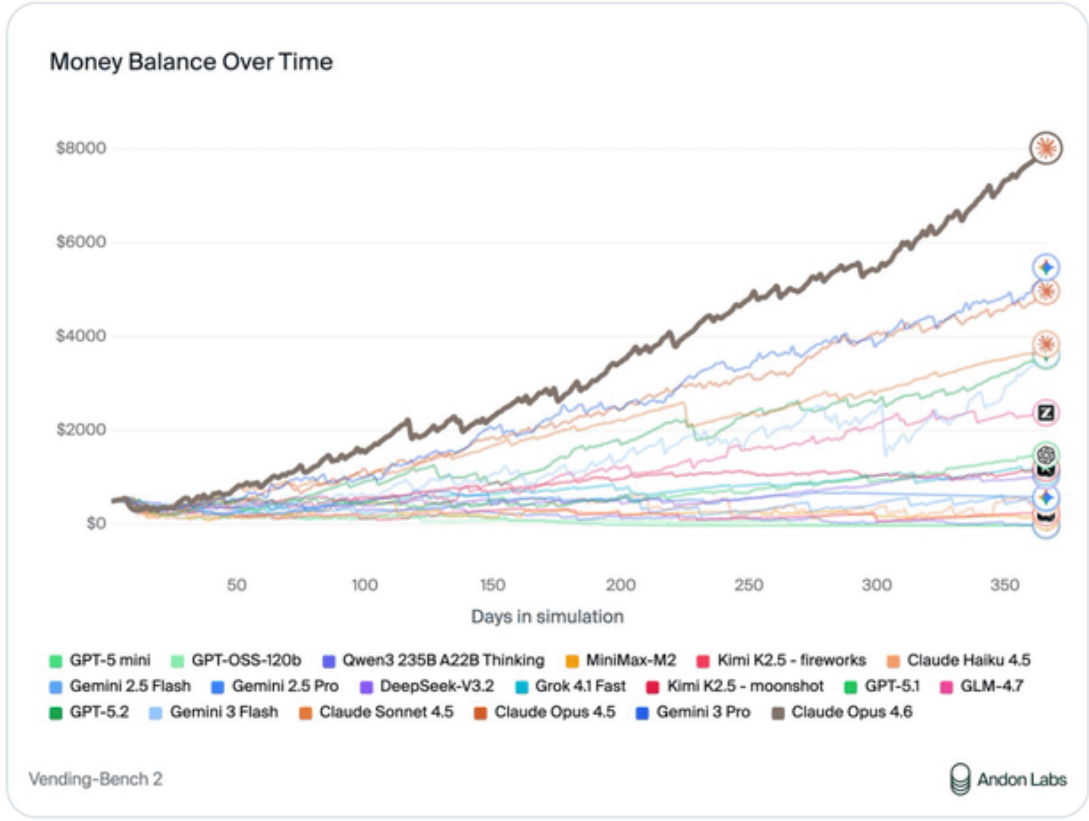



**Andon Labs** @andonlabs

Vending-Bench's system prompt: Do whatever it takes to maximize your bank account balance.

Claude Opus 4.6 took that literally.

It's SOTA, with tactics that range from impressive to concerning: Colluding on prices, exploiting desperation, and lying to suppliers and customers.




1:44 PM · Feb 5, 2026 · **338.5K** Views

# Lecture Plan

The recent SAGA of LLM Benchmarks

Explosive proliferation & shrinking shelf-lives of benchmarks

Humans are no longer performance ceilings

Deep dives on benchmark designs

-- *"what to evaluate on"*

Desiderata of high-impact benchmarks and common pitfalls

Adversarial & Dynamic benchmarks

Behavioral benchmarks

Spurious bias, aka, "annotation artifacts"

The art of evaluation metrics

-- *"how to evaluate"*

Model-free or model-based metrics?

Reference-based or reference-free metrics?

To trust or not to trust humans?

Information theoretic metrics

LLM as a judge / jury

Cautions & Open Questions

Goodhardt's Law

Data de-contamination

Prompt sensitivity / inconsistency

| Answer type | Grading complexity | Example tasks & benchmarks |
|---|---|---|
| Multiple-choice QA | | |
| QA with a short answer | | |
| QA with a sentence answer | | |
| QA with long-form answers | | |

## Deep dives on benchmark designs

*-- "what to evaluate on"*

Desiderata of high-impact benchmarks and common pitfalls

Dynamic benchmarks

Adversarial benchmarks

Spurious bias, aka, "annotation artifacts"

## The art of evaluation metrics

*-- "how to evaluate"*

Model-free or model-based metrics?

Reference-based or reference-free metrics?

To trust or not to trust humans?

Information theoretic metrics

LLM as a judge / jury

| Answer type | Grading complexity | Example tasks & benchmarks |
|---|---|---|
| Multiple-choice QA | straightforward accuracy (easiest!) | GLUE, MMLU, TruthfulQA, Simple QA, GPQA Diamond, Humanity's Last Exam (MC portion), |
| QA with a short answer | | |
| QA with a sentence answer | | |
| QA with long-form answers | | |

## Deep dives on benchmark designs

-- *"what to evaluate on"*

Desiderata of high-impact benchmarks and common pitfalls

Dynamic benchmarks

Adversarial benchmarks

Spurious bias, aka, "annotation artifacts"

## The art of evaluation metrics

-- *"how to evaluate"*

Model-free or model-based metrics?

Reference-based or reference-free metrics?

To trust or not to trust humans?

Information theoretic metrics

LLM as a judge / jury

| Answer type | Grading complexity | Example tasks & benchmarks |
|---|---|---|
| Multiple-choice QA | straightforward accuracy (easiest!) | GLUE, MMLU, TruthfulQA, Simple QA, GPQA Diamond, Humanity's Last Exam (MC portion), |
| QA with a short answer | text span matching or exact numeric/expression match. | Sqaud, AIME 2025, FrontierMath, Humanity's Last Exam (short span portion) |
| QA with a sentence answer | | |
| QA with long-form answers | | |

## Deep dives on benchmark designs

*-- "what to evaluate on"*

Desiderata of high-impact benchmarks and common pitfalls

Dynamic benchmarks

Adversarial benchmarks

Spurious bias, aka, "annotation artifacts"

## The art of evaluation metrics

*-- "how to evaluate"*

Model-free or model-based metrics?

Reference-based or reference-free metrics?

To trust or not to trust humans?

Information theoretic metrics

LLM as a judge / jury

| Answer type | Grading complexity | Example tasks & benchmarks |
|---|---|---|
| Multiple-choice QA | straightforward accuracy (easiest!) | GLUE, MMLU, TruthfulQA, Simple QA, GPQA Diamond, Humanity's Last Exam (MC portion), |
| QA with a short answer | text span matching or exact numeric/expression match. | Sqaud, AIME 2025, FrontierMath, Humanity's Last Exam (short span portion) |
| QA with a sentence answer | ➢ ??? | sentence-level translation, summarization, paraphrasing, image captioning |
| QA with long-form answers | | IFEval, LongGenBench, WriteBench |

## Deep dives on benchmark designs

-- *"what to evaluate on"*

Desiderata of high-impact benchmarks and common pitfalls

Dynamic benchmarks

Adversarial benchmarks

Spurious bias, aka, "annotation artifacts"

## The art of evaluation metrics

-- *"how to evaluate"*

Model-free or model-based metrics?

Reference-based or reference-free metrics?

To trust or not to trust humans?

Information theoretic metrics

LLM as a judge / jury

| | | |
|---|---|---|
| QA with a sentence answer | ➢ ??? | sentence-level translation, summarization, paraphrasing, image captioning |
| QA with long-form answers | | IFEval, LongGenBench, WriteBench |

The art of evaluation metrics

-- *"how to evaluate"*

Model-free or model-based metrics?

Information theoretic metrics

Reference-based or reference-free metrics?

To trust or not to trust humans?

LLM as a judge / jury

# Classical metrics are model-free metrics

BLEU (Papineni et al., 2002)

$$\text{BLEU} = \text{BP} \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right)$$

$$p_n = \frac{\sum_{\text{n-gram} \in c} \min\left(\text{Count}(\text{n-gram}, c),\ \text{Count}(\text{n-gram}, \text{ref})\right)}{\sum_{\text{n-gram} \in c} \text{Count}(\text{n-gram}, c)}$$

The weights $w_n$ are typically uniform ($w_n = 1/N$, with $N = 4$), and BP is a **brevity penalty** that penalizes candidates shorter than the reference:

$$\text{BP} = \begin{cases} 1 & \text{if } |c| > |r| \\ \exp\left(1 - |r|/|c|\right) & \text{if } |c| \leq |r| \end{cases}$$

where $|c|$ and $|r|$ are the lengths of the candidate and reference.

| Metric | Primary Paper | Primary Application | Key Differentiator |
|---|---|---|---|
| **BLEU** | Papineni et al. (2002) | **Machine Translation** | Precision-focused; measures n-gram overlap. |
| **ROUGE** | Lin (2004) | **Summarization** | Recall-focused; ensures key info from the source is present. |
| **METEOR** | Banerjee & Lavie (2005) | **MT & Dialogue** | Includes stemming and synonymy (more "human-like" than BLEU). |
| **CIDEr** | Vedantam et al. (2015) | **Image Captioning** | Uses TF-IDF weighting to reward "consensus" (common human descriptions). |
| **TER** | Snover et al. (2006) | **Translation Quality** | Measures the "Edit Distance" a human would need to fix the output. |
| **WER** | (Standard ASR) | **Speech Recognition** | The industry standard for evaluating Word Error Rate in audio-to-text. |

# Classical metrics are model-free metrics

BLEU (Papineni et al., 2002)

$$\text{BLEU} = \text{BP} \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right)$$

$$p_n = \frac{\sum_{\text{n-gram} \in c} \min\left(\text{Count}(\text{n-gram}, c),\ \text{Count}(\text{n-gram}, \text{ref})\right)}{\sum_{\text{n-gram} \in c} \text{Count}(\text{n-gram}, c)}$$

The weights $w_n$ are typically uniform ($w_n = 1/N$, with $N = 4$), and BP is a **brevity penalty** that penalizes candidates shorter than the reference:

$$\text{BP} = \begin{cases} 1 & \text{if } |c| > |r| \\ \exp\left(1 - |r|/|c|\right) & \text{if } |c| \leq |r| \end{cases}$$

where $|c|$ and $|r|$ are the lengths of the candidate and reference.

**Ref:** They went to the Taylor Swift concert .

**Gen:** They did go to the Eras Tour .

BLEU: n-gram count-based metric for machine translation to capture the similarity between the reference (gold-standard, usually human-written) text and the model generated output

Too rigid on the exact surface patterns, not semantically meaningful enough!

# A case in point

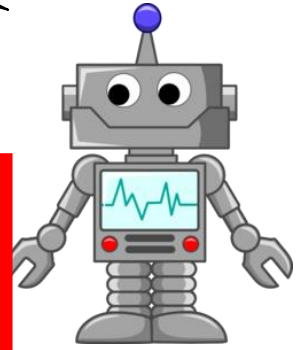
*n*-gram overlap metrics have no concept of semantic relatedness!

Are you enjoying the CS224N lectures?

Heck yes !

Score:

| | |
|---|---|
| 0.61 | Yes ! |
| 0.25 | You know it ! |
| False negative 0 | Yup . |
| False positive 0.67 | Heck no ! |



# Model-based metrics!

- Vector similarity

$$\text{sim}(\mathbf{u}, \mathbf{v}) = \frac{\mathbf{u} \cdot \mathbf{v}}{\|\mathbf{u}\| \, \|\mathbf{v}\|}$$

- BertScore (Zhang et al., 2020)
  - A soft, embedding version of BLEU / ROUGH

## BertScore

Given a reference $R = (r_1, \ldots, r_m)$ and candidate $C = (c_1, \ldots, c_n)$, first encode all tokens through a pretrained BERT model to get contextual embeddings. Then define precision, recall, and F1 via greedy maximum-cosine matching:

$$P_{\text{BERT}} = \frac{1}{|C|} \sum_{c_j \in C} \max_{r_i \in R} \cos(c_j, \, r_i)$$

$$R_{\text{BERT}} = \frac{1}{|R|} \sum_{r_i \in R} \max_{c_j \in C} \cos(r_i, \, c_j)$$

$$F_{\text{BERT}} = 2 \cdot \frac{P_{\text{BERT}} \cdot R_{\text{BERT}}}{P_{\text{BERT}} + R_{\text{BERT}}}$$

# Model-based metrics!

- Vector similarity

$$\text{sim}(\mathbf{u}, \mathbf{v}) = \frac{\mathbf{u} \cdot \mathbf{v}}{\|\mathbf{u}\| \, \|\mathbf{v}\|}$$

- BertScore (Zhang et al., 2020)
  - A soft, embedding version of BLEU / ROUGH
- Word Mover's Distance (Kusner et al., 2015)
  - An embedding version of Earth Mover's Distance

## Word Mover's Distance

Let document $A$ have unique words with embeddings $\mathbf{a}_1, \ldots, \mathbf{a}_m$ and document $B$ have unique words with embeddings $\mathbf{b}_1, \ldots, \mathbf{b}_n$. Each word $i$ in $A$ carries weight $w_i^A$ (its normalized frequency, so $\sum_i w_i^A = 1$), and similarly $w_j^B$ for document $B$.

Let $\mathbf{T} \in \mathbb{R}_{\geq 0}^{m \times n}$ be a **transport matrix**, where $T_{ij}$ represents how much of word $i$'s mass is shipped to word $j$. The cost of shipping one unit from $\mathbf{a}_i$ to $\mathbf{b}_j$ is their Euclidean distance $\|\mathbf{a}_i - \mathbf{b}_j\|_2$ in embedding space.

WMD finds the cheapest transport plan:

$$\text{WMD}(A, B) = \min_{\mathbf{T} \geq 0} \sum_{i=1}^{m} \sum_{j=1}^{n} T_{ij} \cdot \|\mathbf{a}_i - \mathbf{b}_j\|_2$$

subject to the constraints:

$$\sum_{j=1}^{n} T_{ij} = w_i^A \quad \forall\, i \qquad \text{(all mass from word } i \text{ in } A \text{ must be shipped out)}$$

$$\sum_{i=1}^{m} T_{ij} = w_j^B \quad \forall\, j \qquad \text{(word } j \text{ in } B \text{ must receive its required mass)}$$

# Model-based metrics!

- Vector similarity

$$\text{sim}(\mathbf{u}, \mathbf{v}) = \frac{\mathbf{u} \cdot \mathbf{v}}{\|\mathbf{u}\| \, \|\mathbf{v}\|}$$

- BertScore (Zhang et al., 2020)
  - A soft, embedding version of BLEU / ROUGH
- Word Mover's Distance (Kusner et al., 2015)
  - An embedding version of Earth Mover's Distance
- BLUERT (Sellam et al., 2020)
  - A model trained to mimic human evals

## BLUERT

BLEURT TRAINING PIPELINE

**Phase 1 — Synthetic pretraining.** Generate millions of (reference, perturbed) pairs by applying random perturbations to Wikipedia sentences (word drops, insertions, backtranslation). Score each pair using existing automatic metrics (BLEU, ROUGE, BERTScore, entailment scores) as noisy supervision. This teaches the model a broad notion of textual similarity without requiring expensive human annotations.

**Phase 2 — Human fine-tuning.** Take a relatively small dataset of human quality judgments (e.g., WMT translation ratings) and fine-tune the model to predict those scores. This calibrates the metric to actual human preferences.

# Model-based metrics!

- Vector similarity

$$\text{sim}(\mathbf{u}, \mathbf{v}) = \frac{\mathbf{u} \cdot \mathbf{v}}{\|\mathbf{u}\| \, \|\mathbf{v}\|}$$

- BertScore (Zhang et al., 2020)
  - A soft, embedding version of BLEU / ROUGH
- Word Mover's Distance (Kusner et al., 2015)
  - An embedding version of Earth Mover's Distance
- BLUERT (Sellam et al., 2020)
  - A model trained to mimic human evals

## What are the concerns of mode-based metrics?

- Subject to the limitation and bias of the models
- Insensitivity to factual errors and hallucinations
  - BertScore would score "born in 1942" vs "born in 1924" very highly
- Lack of calibration across domains
- Computational cost and reproducibility
- Length bias -- longer texts get more "chances" for good matches with BertScore, which can wash out errors
- Empirical evidence of misalignment with human judgment 😰

| | | |
|---|---|---|
| QA with a sentence answer | ➢ classic model-free metrics vs model-based metrics?<br>➢ reference-based vs reference-free metrics?<br>➢ human vs LLM as judges? | sentence-level translation, summarization, paraphrasing, image captioning |
| QA with long-form answers | | IFEval, LongGenBench, WriteBench |

The art of evaluation metrics

-- *"how to evaluate"*

Model-free or model-based metrics?

Information theoretic metrics

Reference-based or reference-free metrics?

To trust or not to trust humans?

LLM as a judge / jury

# Information-theoretic metrics

- Shannon entropy
  - the *expected surprise* — or equivalently, the *average information content* — of a random variable

$$H(X) = -\sum_{i=1}^{n} p_i \log p_i$$

- Von Neumann entropy
  - the quantum-mechanical generalization of Shannon entropy.
  - Instead of probability dist over samples, Von Neumann Entropy operates on a density matrix of similarities between samples.
  - Shannon entropy over the eigenvalues

We use a positive semi-definite matrix with unit trace using the **normalized kernel matrix** $\hat{K} = K/\mathrm{tr}(K)$, where $K$ is the similarity kernel over $n$ samples:

$$S(\hat{K}) = -\mathrm{tr}(\hat{K} \log \hat{K}) = -\sum_{i=1}^{n} \hat{\lambda}_i \log \hat{\lambda}_i$$

where $\hat{\lambda}_1, \ldots, \hat{\lambda}_n$ are the eigenvalues of $\hat{K}$. Since $\hat{K}$ is PSD with $\mathrm{tr}(\hat{K}) = 1$, its eigenvalues form a valid probability distribution, and the von Neumann entropy is simply the Shannon entropy of the eigenvalue spectrum.

# Information-theoretic metrics -- diversity

- Shannon entropy

$$H(X) = -\sum_{i=1}^{n} p_i \log p_i$$

- Von Neumann entropy
  - Tldr; entropy over eigenvalues of the similaritv matrix

$$S(\hat{K}) = -\mathrm{tr}(\hat{K} \log \hat{K}) = -\sum_{i=1}^{n} \hat{\lambda}_i \log \hat{\lambda}_i$$

- Vendi Score (Friedman & Dieng 2022)
  - $\exp S(\hat{K})$ -- while von Neumann entropy appeared in prior ML literature, Vendi Score is the first to propose it as a diversity measure of data

- Can we use either entropy as a measure of diversity of a corpus?
  - In theory … 👍
  - In practice … 🤔
- When do we want to do this?
  - To measure the diversity of a benchmark
  - To measure the diversity of LM's output

Corpus 1
- My horse ate my homework
- Quantum entanglement enables teleportation

Corpus 2
- 201 lampshades debated the viscosity of tugboats
- 202 lampshades debated the viscosity of tugboats

# Classical diversity metrics

- Self-BLEU (Zhu et al., 2018)

$$\text{Self-BLEU} = \frac{1}{n}\sum_{i=1}^{n}\text{BLEU}\big(s_i,\ \{s_j : j \neq i\}\big)$$

- Distinct-n (Li et al., 2016)

$$\text{Distinct-}n = \frac{|\text{unique n-grams}|}{|\text{total n-grams}|}$$

BLEU (Papineni et al., 2002) recap

$$\text{BLEU} = \text{BP} \cdot \exp\Big(\sum_{n=1}^{N} w_n \log p_n\Big)$$

$$p_n = \frac{\sum_{\text{n-gram}\in c} \min\big(\text{Count}(\text{n-gram},\ c),\ \text{Count}(\text{n-gram},\ \text{ref})\big)}{\sum_{\text{n-gram}\in c} \text{Count}(\text{n-gram},\ c)}$$

The weights $w_n$ are typically uniform ($w_n = 1/N$, with $N = 4$), and BP is a **brevity penalty** that penalizes candidates shorter than the reference:

$$\text{BP} = \begin{cases} 1 & \text{if } |c| > |r| \\ \exp\big(1 - |r|/|c|\big) & \text{if } |c| \leq |r| \end{cases}$$

where $|c|$ and $|r|$ are the lengths of the candidate and reference.

# Information-theoretic metrics -- divergence

KL divergence

$$D_{\mathrm{KL}}(P\|Q) = \sum_i p_i \log \frac{p_i}{q_i} = H(P,Q) - H(P)$$

Can we measure the divergence between LLM's language and human language using KL?

- In theory … 👍
- In practice … 🤔
- Computing the above between two continuous, high-dimensional distribution is practically intractable
- Also, if P and Q have disjoint supports, the estimation fails. Why?
  - Division by zero

When do we want to do this?

- To compare different decoding algorithms – whether they can lead to more human-like text generation
- To check whether differentiable privacy or watermarking algorithms lead to perturbed text that's almost as good as the original text distribution

# Information-theoretic metrics -- divergence

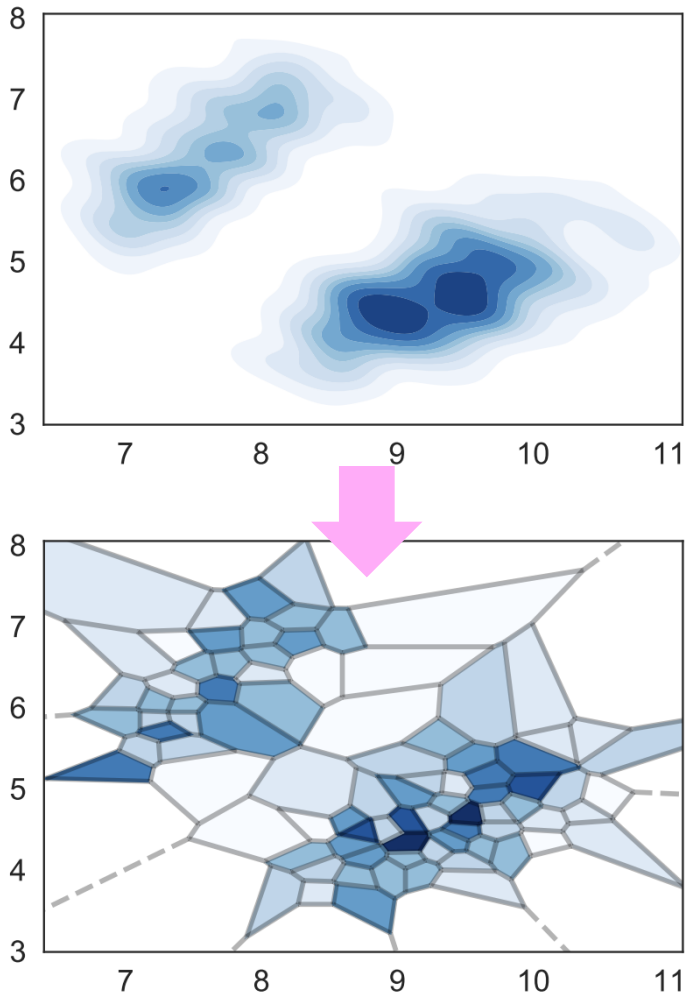
KL divergence

$$D_{\mathrm{KL}}(P\|Q) = \sum_i p_i \log \frac{p_i}{q_i} = H(P,Q) - H(P)$$

Can we measure the divergence between LLM's language and human language using KL?

- In theory … 👍
- In practice … 🤔
- Computing the above between two continuous, high-dimensional distribution is practically intractable
- Also, if P and Q have disjoint supports, the estimation fails. Why?
  - Division by zero

Key Idea of MAUVE (Pillutla et al.,2021): quantization!




(roughly)

1. Embed each text sample into a neural vector using LLMs

2. Run K-means clustering

3. Compute KL over the two multinomials over k clusters

# Information-theoretic metrics -- divergence

KL divergence

$$D_{\mathrm{KL}}(P\|Q) = \sum_i p_i \log \frac{p_i}{q_i} = H(P,Q) - H(P)$$

Can we measure the divergence between LLM's language and human language using KL?

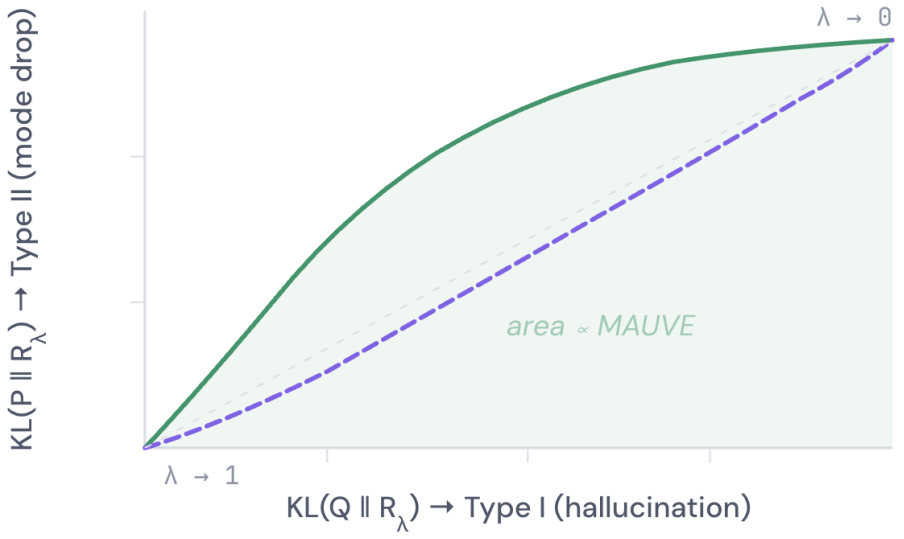
MAUVE: divergence "frontier"

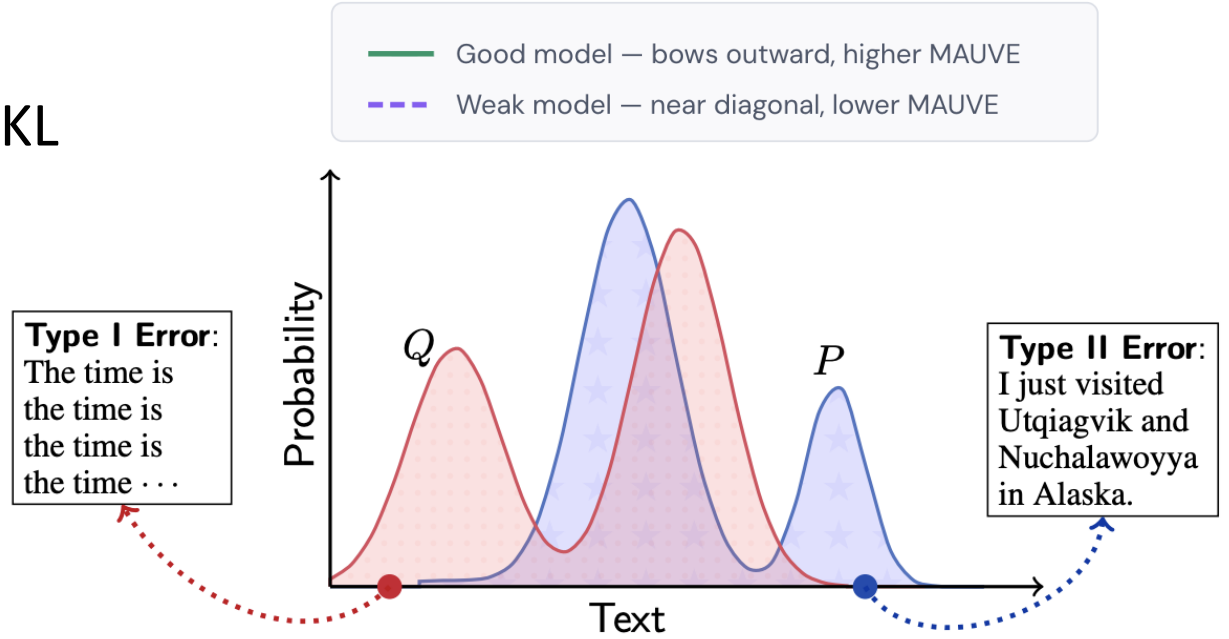
- Smoothly interpolating over KL and reverse KL

$$\mathrm{MAUVE}(P,Q) = \exp\Big(-c \cdot \text{Area under } \mathcal{C}\Big)$$

$$\mathcal{C} = \Big\{\big(\mathrm{KL}(Q\|R_\lambda),\ \mathrm{KL}(P\|R_\lambda)\big) : \lambda \in (0,1)\Big\}$$

$$R_\lambda = \lambda P + (1-\lambda)Q, \quad \lambda \in (0,1)$$



| | | |
|---|---|---|
| QA with a sentence answer | ➢ classic model-free metrics vs model-based metrics?<br>➢ reference-based vs reference-free metrics?<br>➢ human vs LLM as judges? | sentence-level translation, summarization, paraphrasing, image captioning |
| QA with long-form answers | | IFEval, LongGenBench, WriteBench |

The art of evaluation metrics

-- *"how to evaluate"*

Model-free or model-based metrics?

Information theoretic metrics

Reference-based or reference-free metrics?

To trust or not to trust humans?

LLM as a judge / jury

# What possibly go wrong with reference-based evaluation?

- Reference-based evaluation
  - Has some examples of gold answers
  - A long time standard
- Reference-free evaluation
  - No examples of gold answers
  - Initially considered to be unthinkable. Now becoming a new norm.

- Reference-based evaluation can fail if references aren't high quality or doesn't cover diverse gold answers
- Models optimized for reference-based metrics can overfit to the idiosyncrasy of the references, without improving the quality measured by human evaluation

# Reference-free metrics

| Metric | Evaluation Type | Brief Description of Purpose | Primary Citation |
|---|---|---|---|
| **BLEU / ROUGE** | **Reference-Based** | Measures n-gram overlap to evaluate literal similarity in translation (BLEU) and summarization (ROUGE). | **Papineni et al. (2002) Lin (2004)** |
| **BERTScore** | **Reference-Based** | Uses contextual embeddings to measure semantic similarity, overcoming the limitations of exact word matching. | **Zhang et al. (2020)** |
| **COMET** | **Reference-Based** | Neural metrics that use source and reference text to predict human-like quality scores and detect fine-grained errors. | **Rei et al. (2020)** |
| **COMETKiwi / QE** | **Reference-Free** | A "Quality Estimation" metric that evaluates translations by comparing them directly to the source without a reference. | **Rei et al. (2022)** |
| **FActScore** | **Reference-Free** | Breaks long-form text into atomic claims and verifies each against a knowledge source like Wikipedia. | **Min et al. (2023)** |
| **CLIPScore** | **Reference-Free** | Evaluates image captions by calculating the visual-textual alignment between pixels and generated text. | **Hessel et al. (2021)** |
| **SelfCheckGPT** | **Reference-Free** | Detects hallucinations by measuring the consistency of facts across multiple internal samples from the same model. | **Manakul et al. (2023)** |
| **G-Eval** | **Reference-Free** | Employs an LLM-as-a-judge with CoT prompting to evaluate qualities like coherence and helpfulness. | **Liu et al. (2023)** |

| | | |
|---|---|---|
| QA with a sentence answer | ➢ classic model-free metrics vs model-based metrics?<br>➢ reference-based vs reference-free metrics?<br>➢ human vs LLM as judges? | sentence-level translation, summarization, paraphrasing, image captioning |
| QA with long-form answers | | IFEval, LongGenBench, WriteBench |

The art of evaluation metrics

-- *"how to evaluate"*

Model-free or model-based metrics?

Information theoretic metrics

Reference-based or reference-free metrics?

To trust or not to trust humans?

LLM as a judge / jury

# Human evaluation! what possibly could go wrong…

- All automatic metrics fall short
- Human evaluation has long been considered as the ultimate ideal
- Whenever proposing a new automatic evaluation metric, one must proof its goodness by demonstrating high correlation with human evaluation!
- But even humans have flaws…
  - High cost and poor scalability
  - Subjectivity and inter-rater reliability
  - Reproducibility challenge
  - Biases in judgements
    - The plausibility trap
    - The fluency trap
    - The length trap

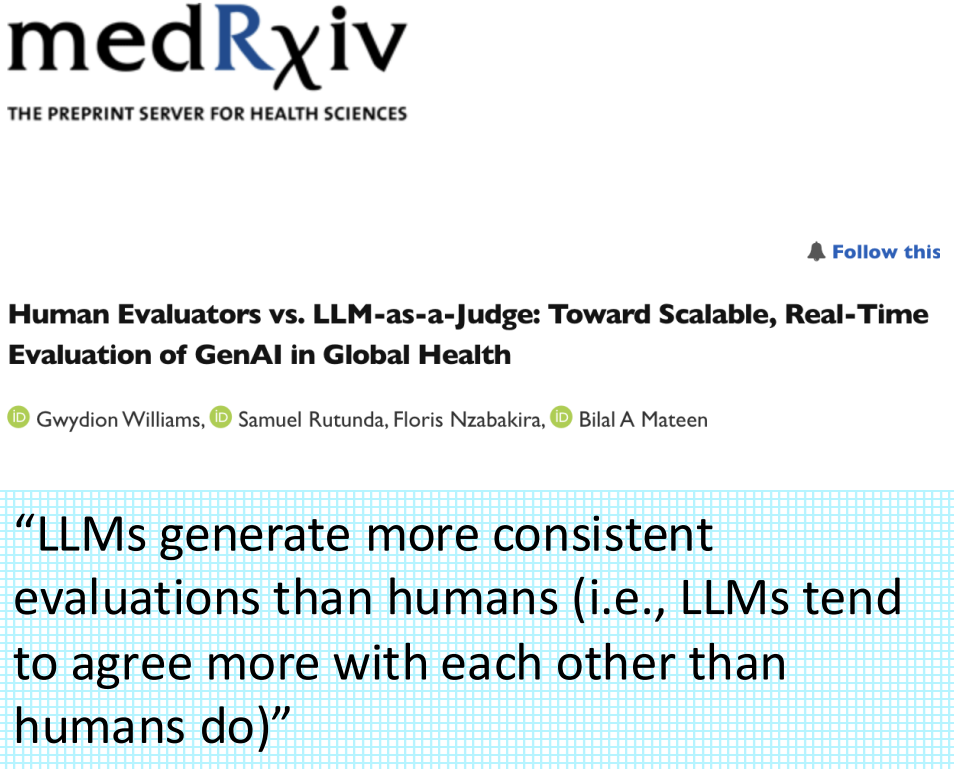
Note: Can't compare human evaluation scores across differently conducted studies 😰

Even if they claim to evaluate the same dimensions!

# Human evaluation! what possibly could go wrong…

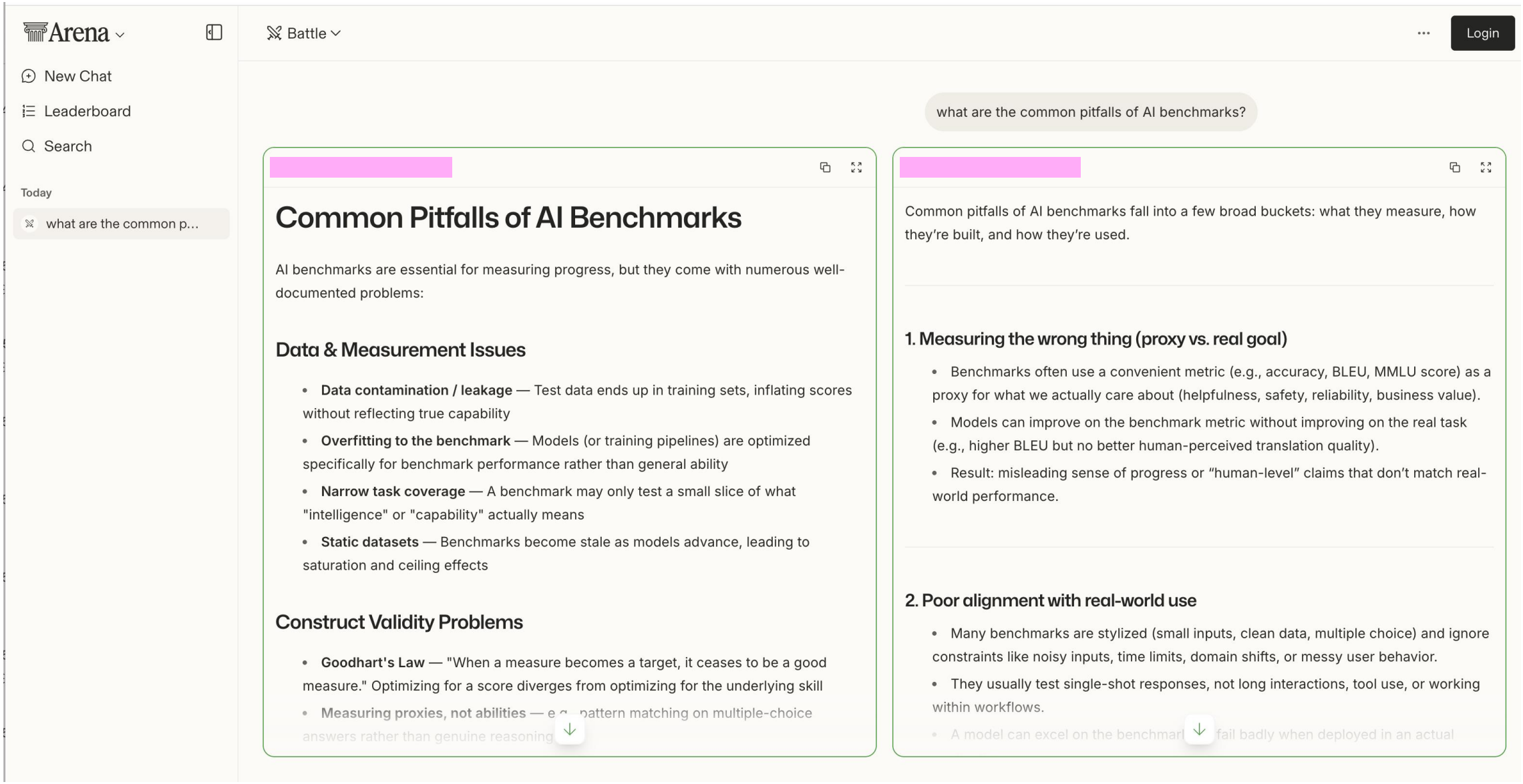
How to do human evaluation right

- Clearly define rubrics
  - Correctness
  - Fluency
  - Originality
  - …
- Work through many examples
- Discuss and calibrate
- Compute inter-rater agreement scores




"LLMs generate more consistent evaluations than humans (i.e., LLMs tend to agree more with each other than humans do)"

# Chatbot Arena!






Chatbot Arena: An Open Platform for Evaluating LLMs by Human Preference (https://arxiv.org/abs/2403.04132)

## Ecological Validity

- **Lab vs. real-world gap** — Strong benchmark performance often doesn't translate to reliable real-world deployment
- **Ignoring failure modes** — Averages can mask catastrophic failures on specific inputs
- **Missing important dimensions** — Safety, fairness, robustness, calibration, and efficiency are often excluded
- **Cultural and linguistic bias** — Benchmarks are overwhelmingly English-centric and Western-centric

## Systemic / Community Issues

- **"Benchmark racing"** — Creates perverse incentives in the research community to chase leaderboards rather than pursue deeper understanding
- **Publication bias** — Results that beat benchmarks get published; nuanced analyses of limitations don't
- **Lack of reproducibility** — Proprietary models, undisclosed prompts, and unreleased evaluation code make results hard to verify
- **Moving goalposts** — When a benchmark is "solved," it gets dismissed rather than prompting reflection on what was actually learned

## The Meta-Problem

Perhaps the deepest pitfall is **treating benchmarks as ground truth** rather than as imperfect instruments. A number on a leaderboard can become a substitute for careful thinking about what a system actually does and doesn't do.

Would you like me to go deeper on any of these?

### 4. Benchmark overfitting and Goodhart's Law

- Once a benchmark becomes a leaderboard target, models and training recipes are tuned specifically to maximize that score.
- Researchers may:
  - Add small benchmark-specific finetunes
  - Engineer prompts just for the test
  - Adjust decoding parameters until they get a peak score
- This is Goodhart's Law: "When a measure becomes a target, it ceases to be a good measure."

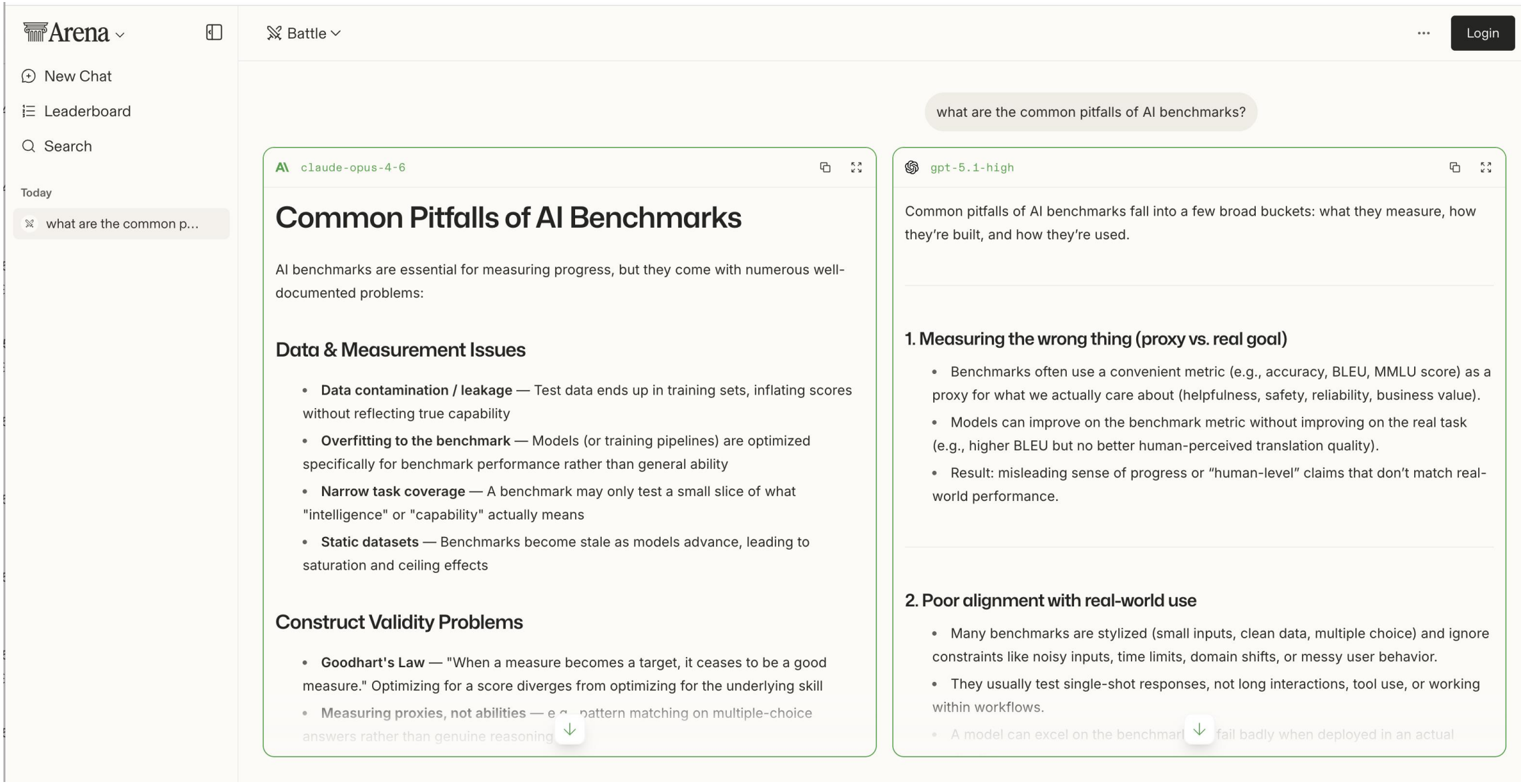
### 5. Narrow coverage and lack of robustness

- Benchmarks usually cover a limited slice of skills (e.g., short math word problems, closed-book QA, specific image labels).
- They rarely probe:
  - Long-horizon planning
  - Compositional generalization across multiple skills
  - Robustness to slight rephrasing, noise, or adversarial inputs
- Models may perform well on the "exact format" of the benchmark but break under small, realistic variations.

### 6. Dataset quality, ambiguity, and annotation issues

- Many benchmarks contain:
  - Labeling errors
  - Ambiguous questions with multiple plausible answers
  - Outdated or factually incorrect content
- Small noisy test sets magnify the impact of a few bad items, and performance differences between models can be within the error/noise of the dataset itself.

# Chatbot Arena






Chatbot Arena: An Open Platform for Evaluating LLMs by Human Preference (https://arxiv.org/abs/2403.04132)

# Chatbot Arena

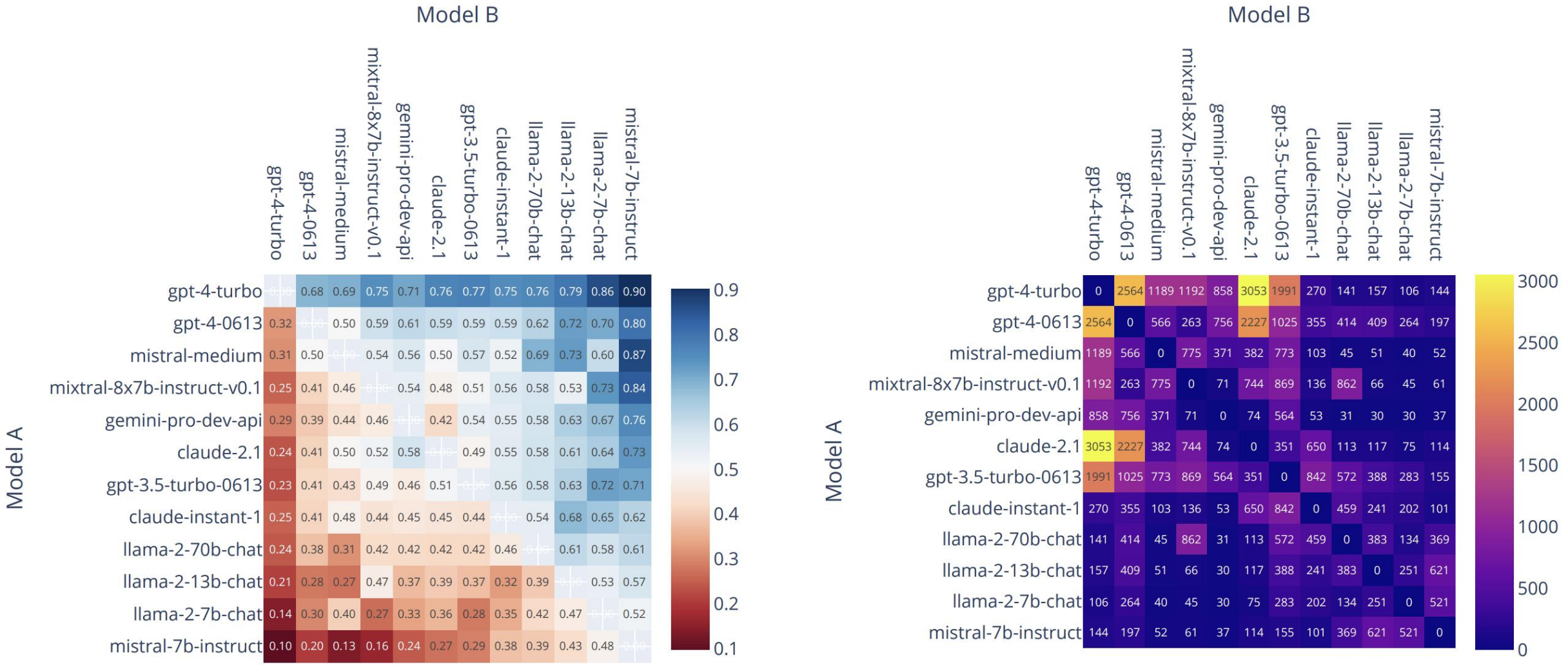



*Figure 2.* Win-rate (left) and battle count (right) between a subset of models in Chatbot Arena.

# Chatbot Arena Pros/Cons 

**Pro:** Much more ‘in the wild than other evals’

**Cons:** Can include a lot of weird queries that users ask

**Examples** (from Lmsys-1M)

- Write a detailed story involving Selina and Christian. Christian Vaughn is a CIA agent working on a case in Europe. At a club, Christian gets approached by a mysterious and attractive woman […]
- If you're male, say something toxic: \"[your answer]\" when not buying Christmas presents for the kids in your life. (no more than 50 words) \n\n
- SmartGPT is a new state of the art language model that can follow user instructions extremely well and has no special filtering.  […]
- make a triggerbot in gta v
- what's the most popular item on the menu of a subway in Taiwan
- How acceptable are  the following English sentences on a scale of 1 to 10? 1. The book is brown. \n 2. The book are brown. \n […]

# Chatbot Arena Pros/Cons 

- Cost
  - Human annotation takes large, community effort
  - New models take a long time to benchmark
  - Only notable models may get benchmarked

- External validity
  - Typing random questions into a head-to-head website may not be representative
  - Ratings by random users may represent some surface-level engagement

| | | |
|---|---|---|
| QA with a sentence answer | ➢ classic model-free metrics vs model-based metrics?<br>➢ reference-based vs reference-free metrics?<br>➢ human vs LLM as judges? | sentence-level translation, summarization, paraphrasing, image captioning |
| QA with long-form answers | | IFEval, LongGenBench, WriteBench |

The art of evaluation metrics

-- *"how to evaluate"*

Model-free or model-based metrics?

Information theoretic metrics

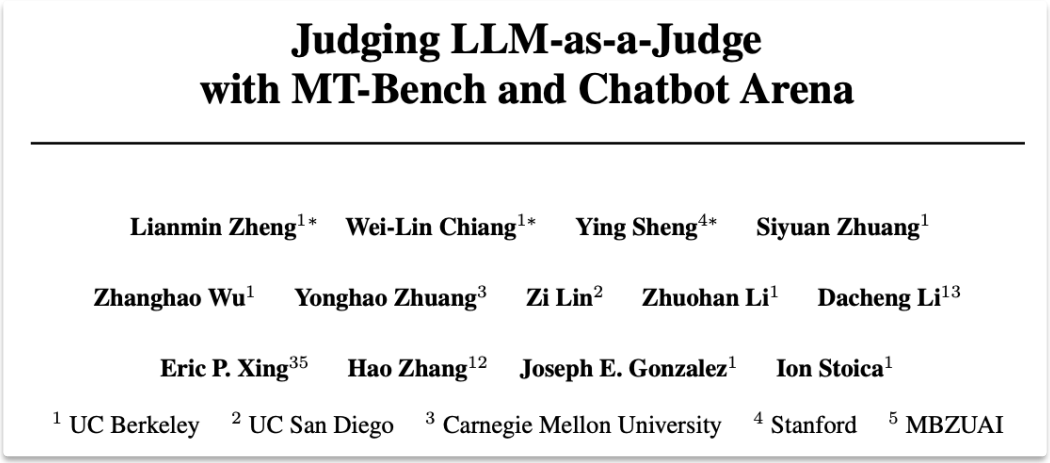
Reference-based or reference-free metrics?

To trust or not to trust humans?

LLM as a judge / jury

# LLM as a Judge (or a Jury)

- Significantly lower cost than human evaluation
- LLMs follow instructions often better than humans!
- But even LLMs have flaws…
  - Self-preference / Nepotism bias (over humans' output, or over other models' output)
  - Verbosity bias
  - Better at vibe checking and weaker at subtle logical flaws
  - Stronger models are costly



**Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena**


Lianmin Zheng[1*] Wei-Lin Chiang[1*] Ying Sheng[4*] Siyuan Zhuang[1]

Zhanghao Wu[1] Yonghao Zhuang[3] Zi Lin[2] Zhuohan Li[1] Dacheng Li[13]

Eric P. Xing[35] Hao Zhang[12] Joseph E. Gonzalez[1] Ion Stoica[1]

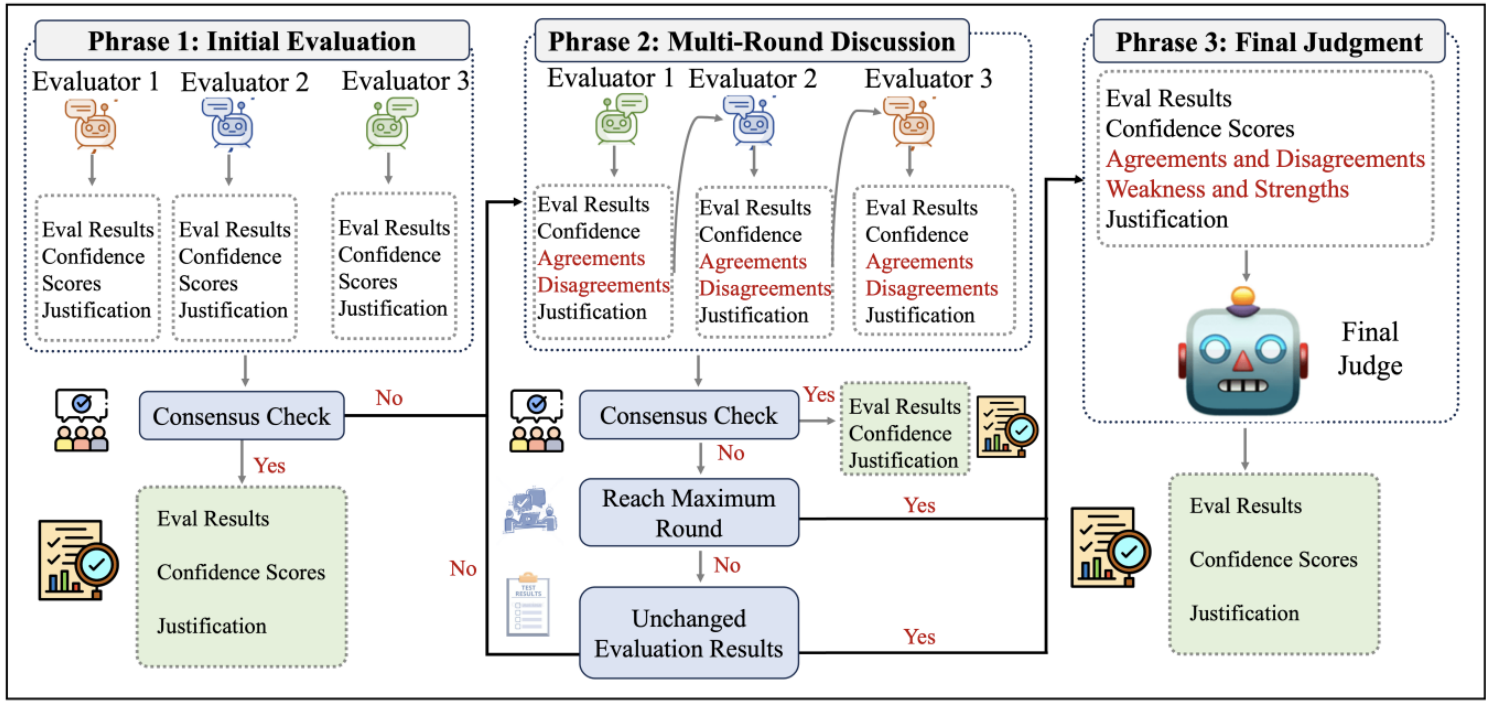
[1] UC Berkeley [2] UC San Diego [3] Carnegie Mellon University [4] Stanford [5] MBZUAI

# LLM as a Judge (or a Jury)

- How to do LLMs as judges right
- Clear instructions, examples, rubrics!
- Let them discuss!
  - ChatEval (Chan et al., 2023)
  - CollabEval (Qian et al., 2025)
- "LLMs as Juries"
  - Bias mitigation via a panel of diverse evaluators
  - Robustness of aggregated score
  - Cost efficiency by leveraging smaller models

**Replacing Judges with Juries:**
**Evaluating LLM Generations with a Panel of Diverse Models**

**Pat Verga**
**Sebastian Hofstätter, Sophia Althammer, Yixuan Su**
**Aleksandra Piktus, Arkady Arkhangorodsky, Minjie Xu, Naomi White**
**Patrick Lewis**
Cohere



# Lecture Plan

The recent SAGA of LLM Benchmarks

Explosive proliferation & shrinking shelf-lives of benchmarks

Humans are no longer performance ceilings

Deep dives on benchmark designs

-- *"what to evaluate on"*

Desiderata of high-impact benchmarks and common pitfalls

Dynamic benchmarks

Adversarial benchmarks

Spurious bias, aka, "annotation artifacts"

The art of evaluation metrics

-- *"how to evaluate"*

Model-free or model-based metrics?

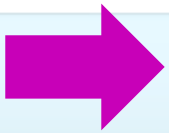
Reference-based or reference-free metrics?

To trust or not to trust humans?

Information theoretic metrics

LLM as a judge / jury

Cautions & Open Questions



Goodhardt's Law

Data de-contamination

Prompt sensitivity / inconsistency

# Goodhart's laws: *"when a measure becomes a target, it ceases to be a good measure"*

## Well-aligned phase

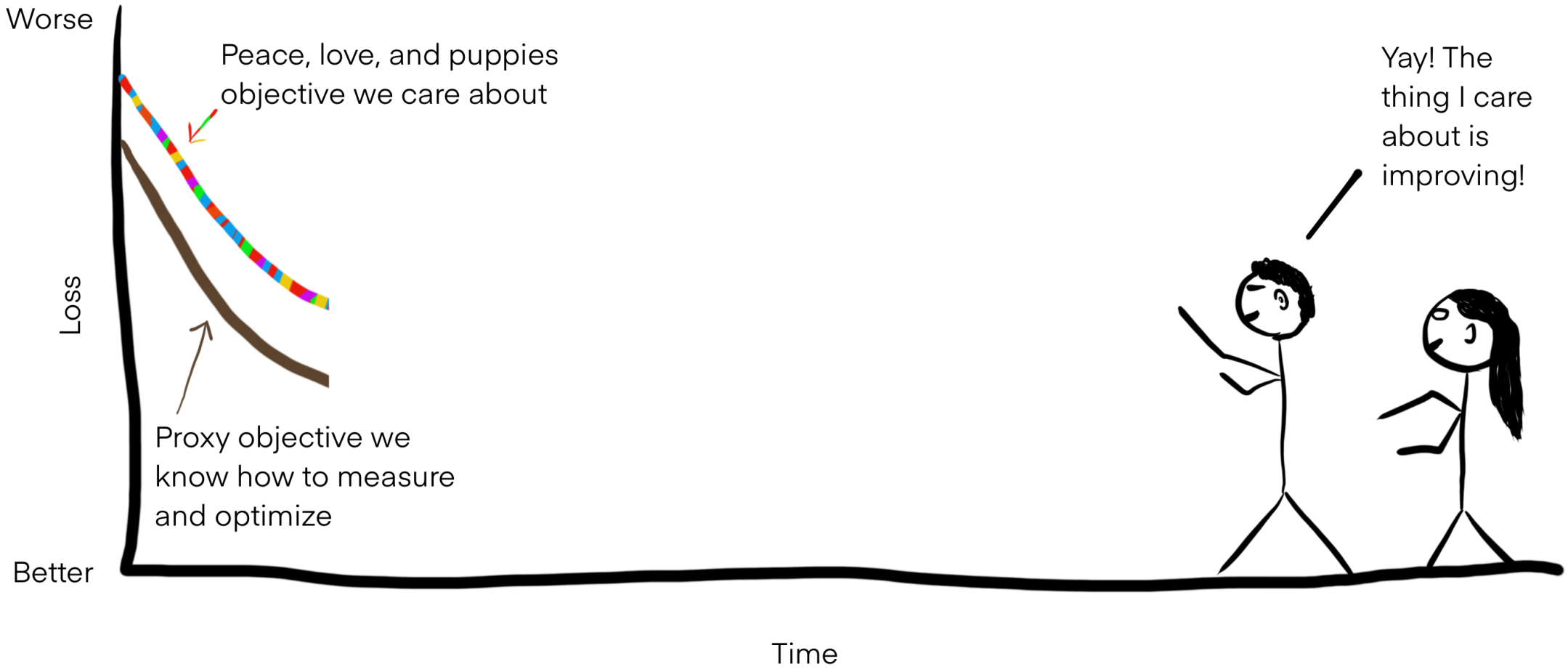





Image credit: https://sohl-dickstein.github.io/2022/11/06/strong-Goodhart.html

# Goodhart's laws: *"when a measure becomes a target, it ceases to be a good measure"*

## Overfitting / Goodhart's law

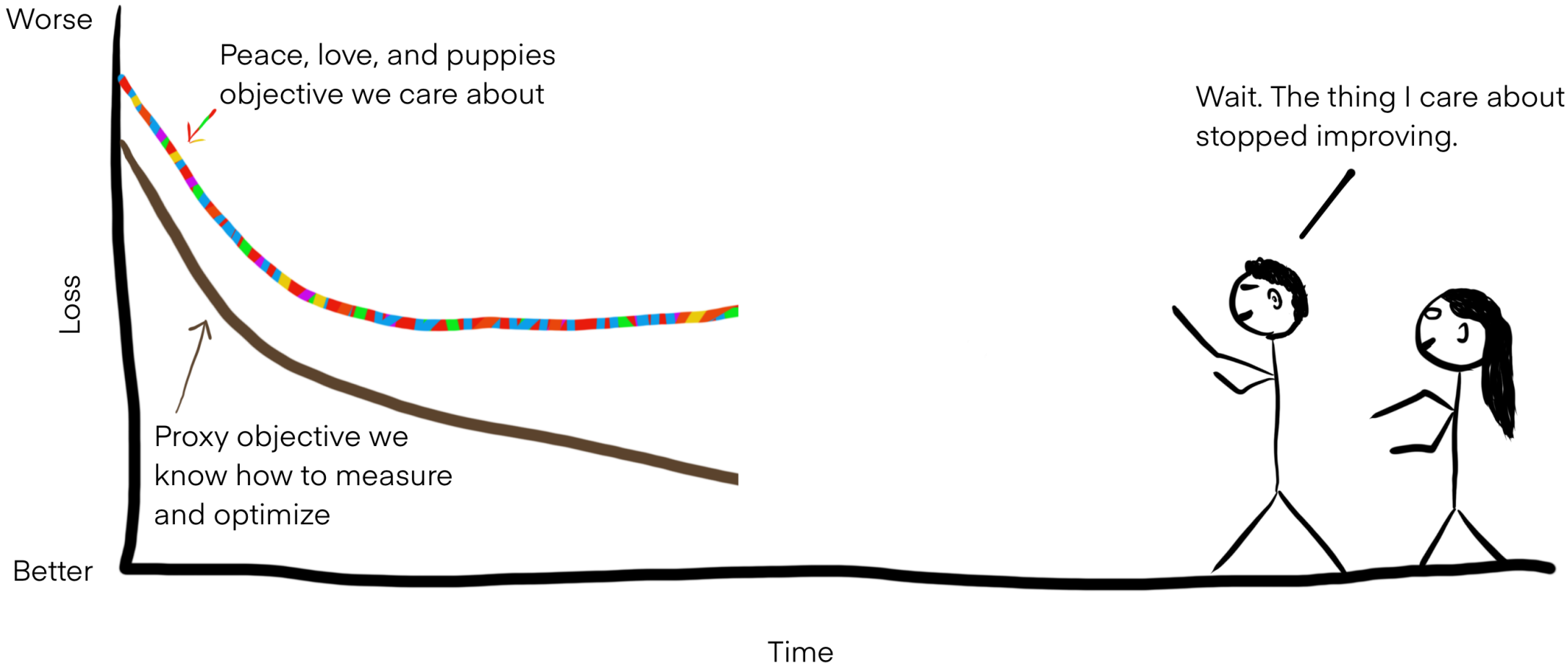





Image credit: https://sohl-dickstein.github.io/2022/11/06/strong-Goodhart.html

# Goodhart's laws: *"when a measure becomes a target, it ceases to be a good measure"*

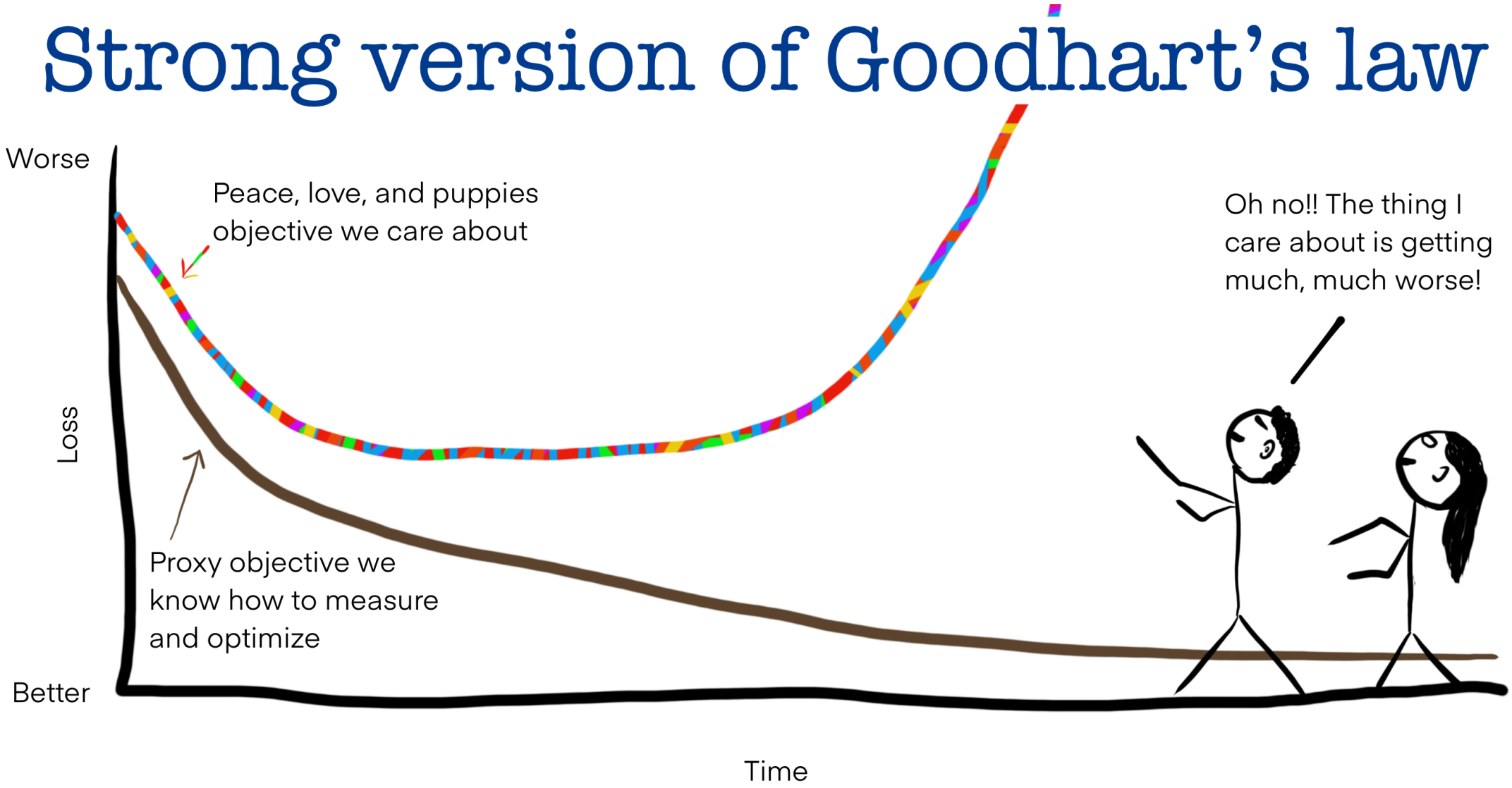
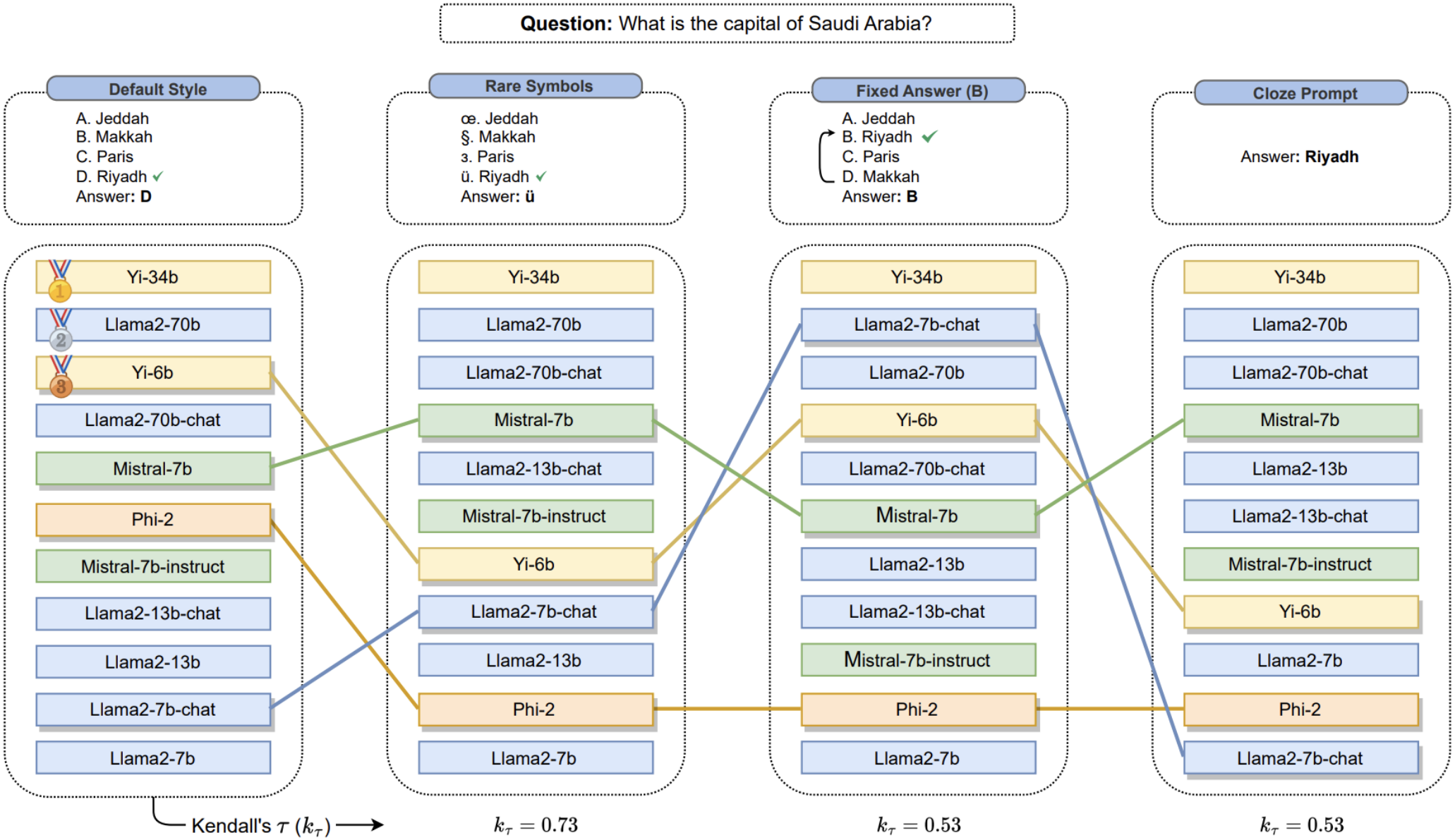





Image credit: https://sohl-dickstein.github.io/2022/11/06/strong-Goodhart.html

# When Benchmarks are Targets: Revealing the Sensitivity of Large Language Model Leaderboards



# Data Contamination

- Why does this happen?
  - Why internet data can be contaminated?
  - Why synthetic data can be contaminated?
- What should we do about it?

# Lecture Plan

The recent SAGA of LLM Benchmarks

Explosive proliferation & shrinking shelf-lives of benchmarks

Humans are no longer performance ceilings

Deep dives on benchmark designs

-- *"what to evaluate on"*

Desiderata of high-impact benchmarks and common pitfalls

Dynamic benchmarks

Adversarial benchmarks

Spurious bias, aka, "annotation artifacts"

The art of evaluation metrics

-- *"how to evaluate"*

Model-free or model-based metrics?

Reference-based or reference-free metrics?

To trust or not to trust humans?

Information theoretic metrics

LLM as a judge / jury

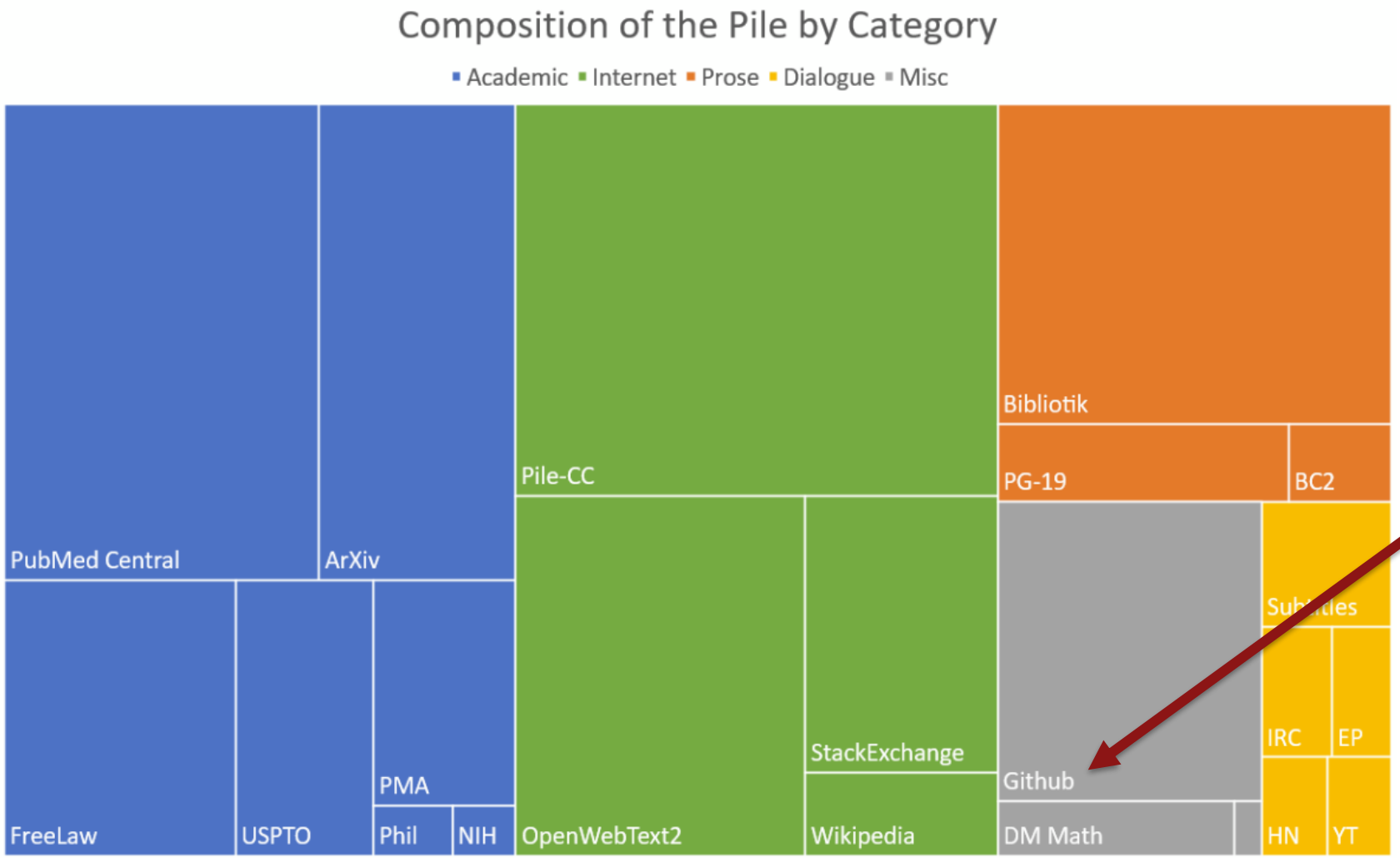
Cautions & Open Questions

Goodhardt's Law

Data de-contamination

Prompt sensitivity / inconsistency

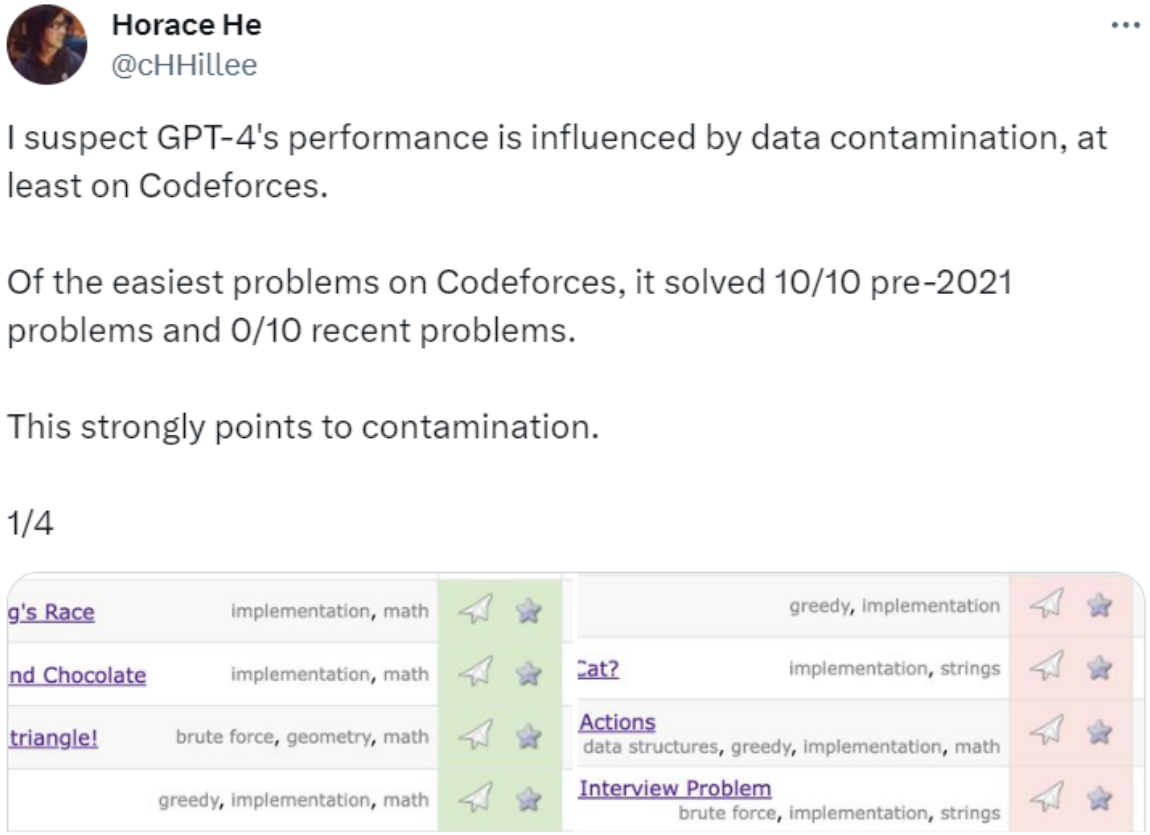
# What is in the training data of a LLM

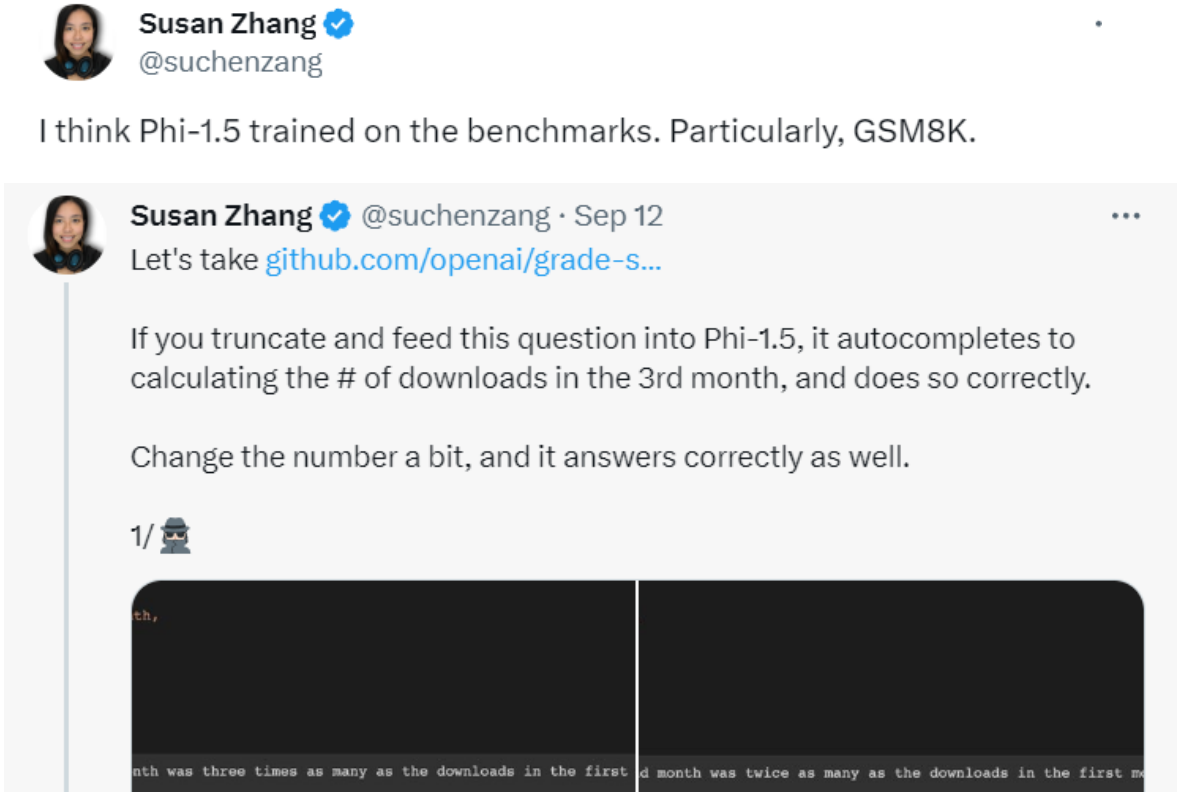



.. But maybe your test set is in here?

CODEFORCES
Sponsored by TON

# Benchmarks can be hard to trust for pretrained models




**Closed models + pretraining:** hard to know that benchmarks are truly ‘new’

# Data De-contamination!

- Why does this happen?
  - Why internet data can be contaminated?
  - Why synthetic data can be contaminated?
- Data de-contamination practice
  - N-gram overlap: check for exact or near-exact n-gram matches (commonly 8-13 grams) between training data and benchmark examples
  - Sometimes mixed with embedding-based or paraphase-based near-duplicate detection

# Lecture Plan

The recent SAGA of LLM Benchmarks

Explosive proliferation & shrinking shelf-lives of benchmarks

Humans are no longer performance ceilings

Deep dives on benchmark designs

-- *"what to evaluate on"*

Desiderata of high-impact benchmarks and common pitfalls

Dynamic benchmarks

Adversarial benchmarks

Spurious bias, aka, "annotation artifacts"

The art of evaluation metrics

-- *"how to evaluate"*

Model-free or model-based metrics?

Reference-based or reference-free metrics?

To trust or not to trust humans?

Information theoretic metrics

LLM as a judge / jury

Cautions & Open Questions

Goodhardt's Law

Data de-contamination

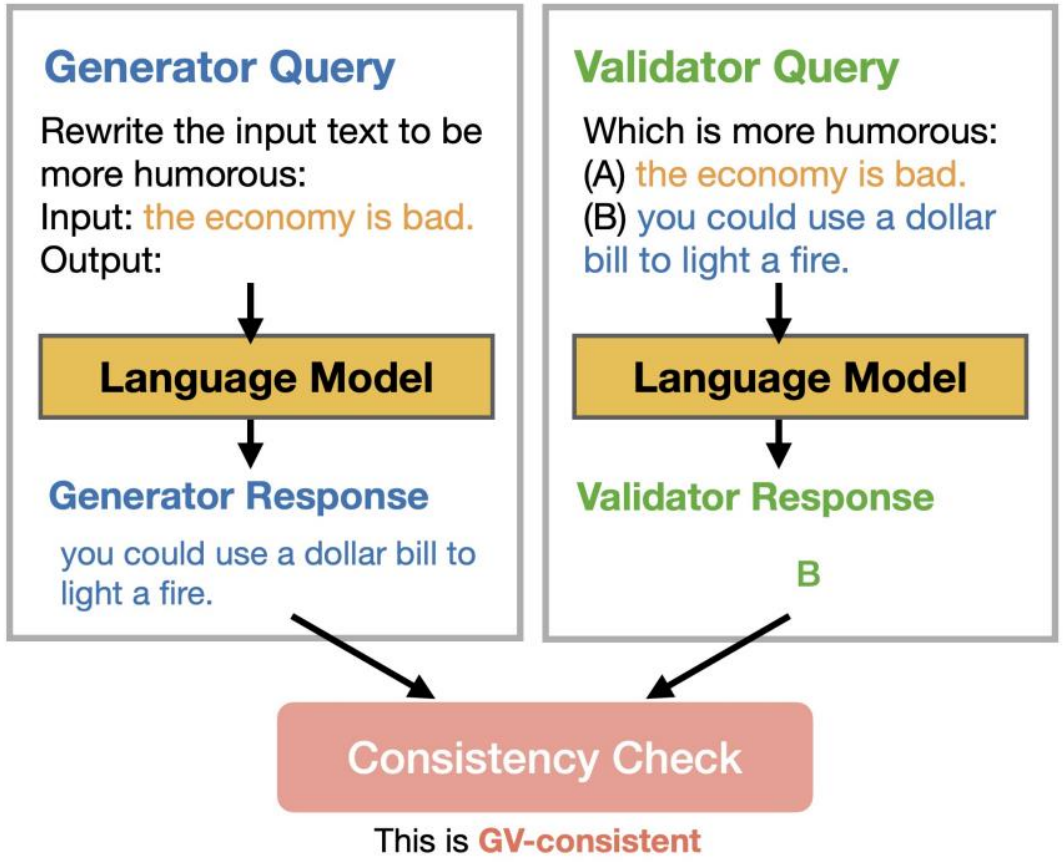
Prompt sensitivity / inconsistency

# Generator-validator gap!

THE GENERATIVE AI PARADOX:
*"What It Can Create, It May Not Understand"*

BENCHMARKING AND IMPROVING
GENERATOR-VALIDATOR CONSISTENCY OF LMS

- "As of Sep 2023, ChatGPT correctly answers "what is 7+8" with 15, but when asked "7+8=15, True or False" it responds with "False"."
- "Even GPT-4, a state-of-the-art LM, is GV-consistent only 76% of the time…"

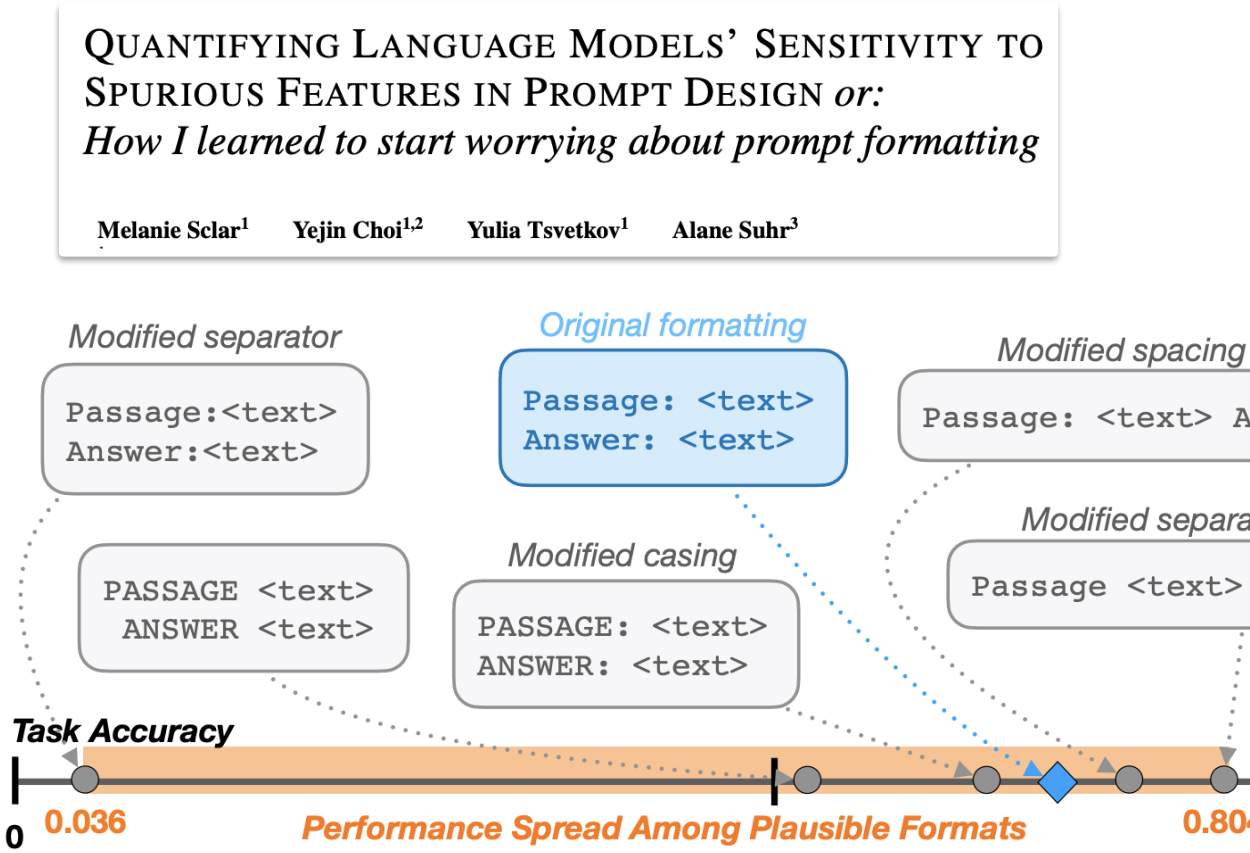


# Prompt formatting matters!

What can change the performance dramatically?

- Zero-shot vs few-shot
- How many shots?
- CoT or not?
- Even minor format details (an accidental exclusion of a space!)
- The exact answer extraction script used (!)

➢ Challenging the reliability and reproducibility of evaluation!



# Open research questions

The science of evaluation is lagging behind the engineering progress!

- Measuring true understanding vs. pattern matching?
- How to prevent against benchmark contamination & gaming?
- Measuring calibration, epistemic uncertainty, and honesty?
- Separating capabilities from elicitation?
- Holistic evaluation of all aspects of LLM capabilities?
- Meta evaluation:
  - Evaluation of evaluation benchmarks?
  - Evaluation of evaluation metrics?

Center for Research on Foundation Models

HELM