# Lecture 17: Evaluating LLMs

*Instructor: Swabha Swayamdipta*

*USC CSCI 444 NLP*

*Nov 10, 2025*

Some slides adapted from Dan Jurafsky and Chris Manning

# Announcements + Logistics

- Progress Report and Quiz 4 grades out
- Upcoming:
  - Paper Discussions on Wed, 11/12: MMLU, Dynabench, Chatbot Arena
  - 11/ 17: HW3 due
  - 11/ 24: Quiz 5 + Bonus
  - 12/1 and 12/3: Project Presentations: Sign up for dates / slots ASAP

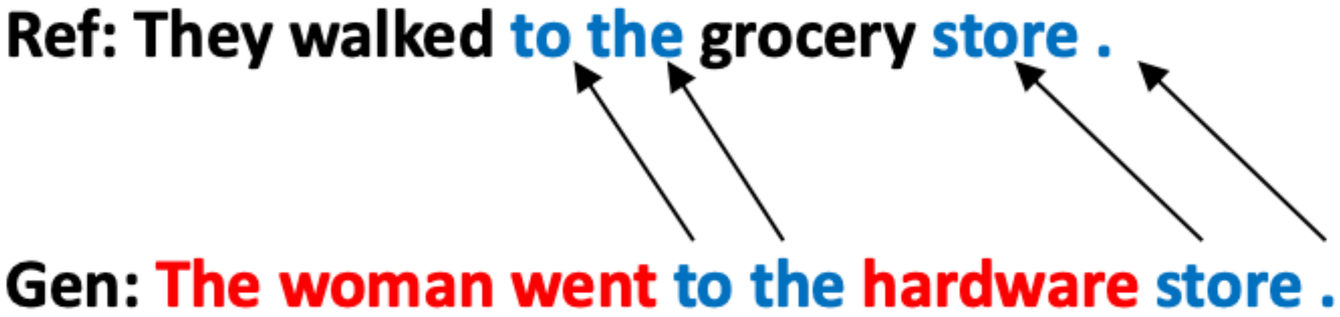
# Lecture Outline

- Announcements
- Quiz 4 Answers
- Generative Evaluation: Metrics / Methods
- Benchmarks
- Rethinking Evaluation Benchmarks
  - Some of my own research

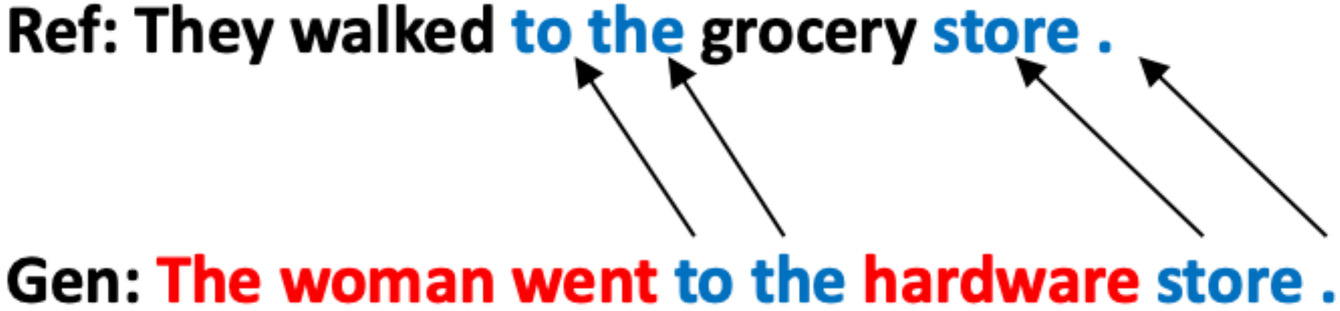
# Generative Evaluation Metrics / Methods

# Evaluation Strategies

- With Reference
  - Lexical Matching
  - Semantic Matching
- Without Reference
  - Perplexity
  - Model-Based Metrics
  - Advanced: Distributional Matching
  - Simplest, Most Reliable Strategy to-date: Human Evaluation
  - Even simpler and least reliable: Auto Evaluation



# Reference-Based Metrics




- Only possible for close-ended generation tasks
- Compute a score that indicates the lexical similarity between generated and gold-standard (human-written) text
- Fast and efficient and widely used
- $n$-gram overlap metrics (e.g., BLEU, ROUGE, etc.)

# BLEU

- Stands for Bilingual Evaluation Understudy
- BLEU compares the machine-written translation to one or several human-written translation(s), and computes a similarity score based on:
  - Geometric mean of n-gram precision (usually for 1, 2, 3 and 4-grams)
  - Plus a penalty for too-short system translations
- BLEU is useful but imperfect
  - There are many valid ways to translate a sentence
  - So a good translation can get a poor BLEU score because it has low n-gram overlap with the human translation
- Precision-based metric

Papineni et al., 2002

# Precision, Recall and F-1

- True Positives, True Negatives, False Positives and False Negatives

$$\textbf{Precision} = \frac{TP}{TP + FP}$$

Of all the items in the prediction, how many match the ground truth

$$\textbf{Recall} = \frac{TP}{TP + FN}$$

Of all the items in the ground truth, how many are correctly predicted

$$F_1 = \frac{2 * PR}{P + R}$$

Harmonic Mean of Precision and Recall

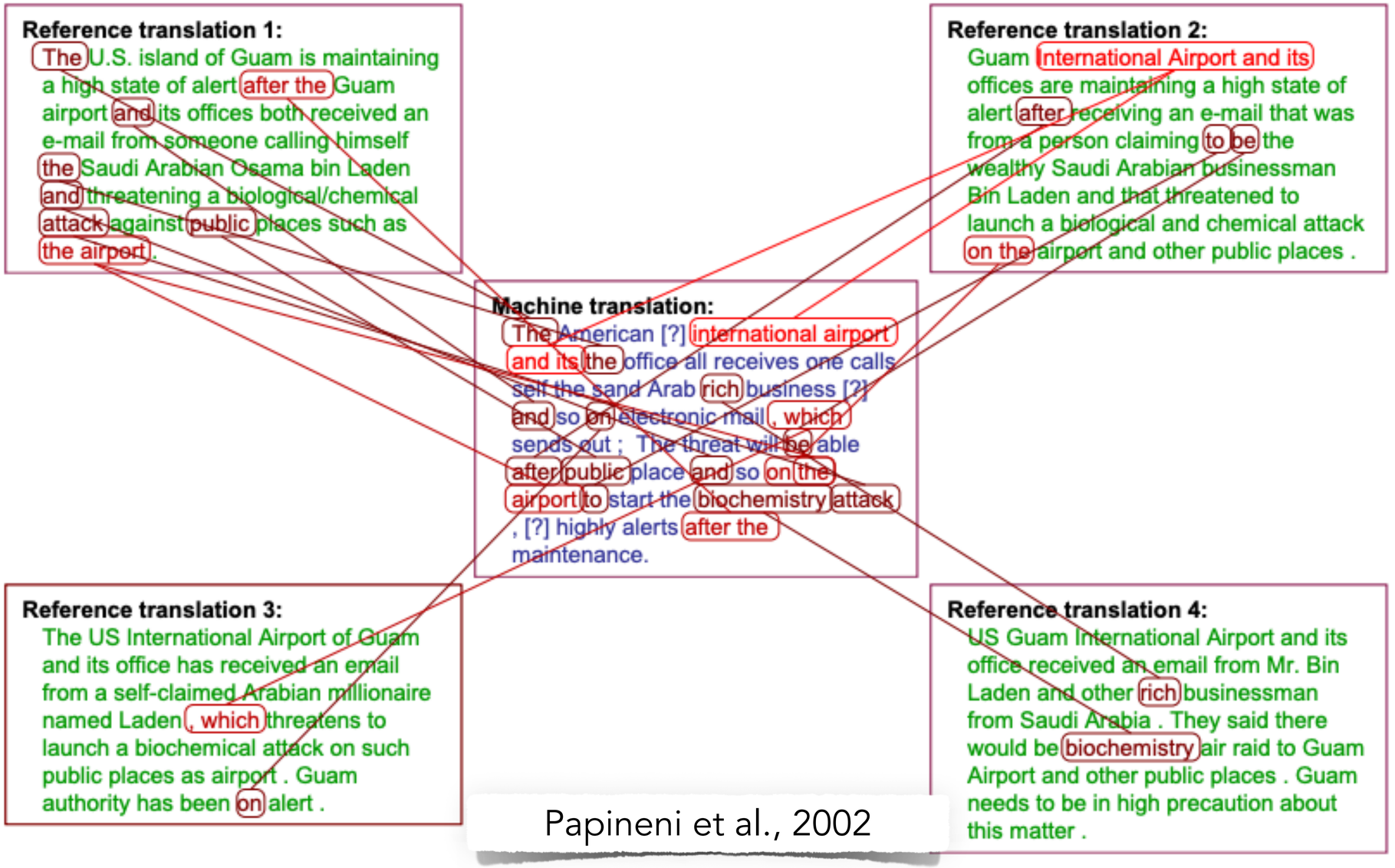
Different value for different classes!

# BLEU: Details

- Purely precision-based rather than combining precision and recall.
- BLEU score for a corpus of candidate translation sentences is a function of
  - the n-gram word precision over all the sentences
  - combined with a brevity penalty computed over the corpus as a whole.
- Consider a corpus composed of a single sentence
  - The unigram precision for this corpus is the percentage of unigram tokens in the candidate translation that also occur in the reference translation, and ditto for bigrams and so on, up to 4-grams
  - It computes this n-gram precision for unigrams, bigrams, trigrams, and 4-grams and takes the geometric mean
- Because BLEU is a word-based metric, it is very sensitive to word tokenization, making it impossible to compare different systems if they rely on different tokenization

Papineni et al., 2002

# BLEU: Example




Papineni et al., 2002

# ROUGE

- Stands for “Recall-Oriented Understudy for Gisting Evaluation”
- Originally created for evaluating automatic summarization as well as machine translation
- Comparing an automatically produced summary or translation against a set of reference summaries (typically human-produced)
- Four variants:
  - ROUGE-N
  - ROUGE-L
  - ROUGE-S
  - ROUGE-W

ROUGE: A Package for Automatic Evaluation of Summaries (Lin, 2004)

# ROUGE: Details

- **ROUGE-N**: measures **unigram**, **bigram**, **trigram** and higher order n-gram overlap
  - n-gram recall between a candidate summary and a set of reference summaries
- **ROUGE-L**: measures **longest matching sequence** of words using LCS.
  - Does not require consecutive matches but in-sequence matches that reflect sentence level word order.
  - Since it automatically includes longest in-sequence common n-grams, you don't need a predefined n-gram length.
- **ROUGE-S**: Is any pair of words in a sentence in order, allowing for arbitrary gaps.
  - Also be called skip-gram concurrence.
  - For example, **skip-bigram** measures the overlap of word pairs that can have a maximum of two gaps in between words. As an example, for the phrase *"cat in the hat"* the skip-bigrams would be *"cat in, cat the, cat hat, in the, in hat, the hat"*.
- **ROUGE-W**: Weighted Longest Common Subsequence

ROUGE: A Package for Automatic Evaluation of Summaries (Lin, 2004)

# Evaluating Generation: Other Options

$$PPL(\mathbf{w}) = P(w_1 w_2 \ldots w_N)^{-\frac{1}{N}}$$

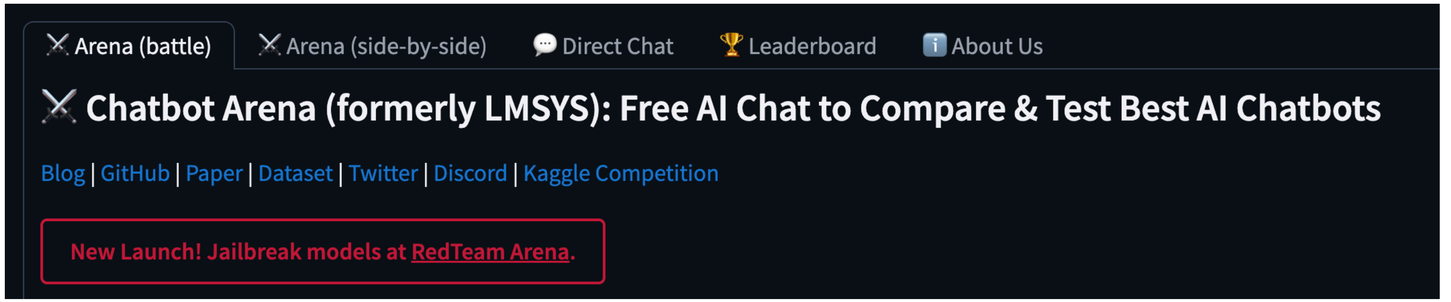
- Perplexity!
- Model-based Metrics (BERTScore, BARTScore, Word Mover's Distance, BLEURT)
  - Use learned representations of words and sentences to compute semantic similarity between generated and reference texts
  - No more n-gram bottleneck because text units are represented as embeddings!
  - The embeddings are pretrained, distance metrics used to measure the similarity can be fixed
- Automatic metrics for text comparisons fall short of matching human decisions
- So, Human Evaluation!

# Human Evaluation

- Ask humans to evaluate the quality of generated text
  - Along specific axes: fluency, coherence / consistency, factuality and correctness, commonsense, etc.
  - Mostly done via crowdsourcing
- Human judgments are regarded as the gold standard
- Many, many correct answers! Ground truth might be irrelevant
- Most LLMs are compared with each other on their generations directly
  - Evaluated by humans or models

# Chatbot Arena

- How to do human evaluation at very large scales?
- Maybe a model is good if people choose to use it (and pay for it)…
- Platform for ranking LLMs
- Limitations:
  - Could be possible to game this…




| Rank* (UB) | Model | Arena Score | 95% CI | Votes | Organization | License | Knowledge Cutoff |
|---|---|---|---|---|---|---|---|
| 1 | o1-preview | 1355 | +12/-11 | 2991 | OpenAI | Proprietary | 2023/10 |
| 2 | ChatGPT-4o-latest (2024-09-03) | 1335 | +5/-6 | 10213 | OpenAI | Proprietary | 2023/10 |
| 2 | o1-mini | 1324 | +12/-9 | 3009 | OpenAI | Proprietary | 2023/10 |
| 4 | Gemini-1.5-Pro-Exp-0827 | 1299 | +5/-4 | 28229 | Google | Proprietary | 2023/11 |
| 4 | Grok-2-08-13 | 1294 | +4/-4 | 23999 | xAI | Proprietary | 2024/3 |
| 6 | GPT-4o-2024-05-13 | 1285 | +3/-3 | 90695 | OpenAI | Proprietary | 2023/10 |
| 7 | GPT-4o-mini-2024-07-18 | 1273 | +3/-3 | 30434 | OpenAI | Proprietary | 2023/10 |
| 7 | Claude 3.5 Sonnet | 1269 | +3/-3 | 62977 | Anthropic | Proprietary | 2024/4 |
| 7 | Gemini-1.5-Flash-Exp-0827 | 1269 | +4/-4 | 22264 | Google | Proprietary | 2023/11 |
| 7 | Grok-2-Mini-08-13 | 1267 | +4/-5 | 22041 | xAI | Proprietary | 2024/3 |
| 7 | Gemini Advanced App (2024-05-14) | 1267 | +3/-3 | 52218 | Google | Proprietary | Online |

## Shane, Cameron, or Saxon—who is the ultimate White Lotus alpha?

**GigiPlaysPonies** · 22h ago

Can we all agree it isn't the one whose mother showed up on his honeymoon...

938 · Reply · Award · Share · ···

**CT_Phipps-Author** · 16h ago

2 more replies

**SeaworthySamus** · 23h ago

Cameron is who Shane and Saxon think they are

9K · Reply · 1 · Share · ···

**funguy07** · 23h ago · Edited 23h ago

**Excellent_Aerie** · 22h ago

Top 1% Commenter

**KeyPosition3983** · 22h ago

This is the answer. Cameron actually is who he thinks he is

132 · Reply · Award · Share · ···

19 more replies

**Feebedel324** · 19h ago

**Larro83** · 23h ago

This is the only answer.

47 · Reply · Award · Share · ···

38 more replies

**Let_us_proceed** · 23h ago

Top 1% Commenter

Do people really view Shane and Saxon as alphas?

**MAUVE: Measuring the Gap Between Neural Text and Human Text using Divergence Frontiers**

Krishna Pillutla[1] Swabha Swayamdipta[2] Rowan Zellers[1] John Thickstun[3]
Sean Welleck[1,2] Yejin Choi[1,2] Zaid Harchaoui[4]

[1]Paul G. Allen School of Computer Science & Engineering, University of Washington
[2]Allen Institute for Artificial Intelligence
[3]Department of Computer Science, Stanford University
[4]Department of Statistics, University of Washington

When considering who might be the "ultimate alpha" among Shane, Cameron, and Saxon from The White Lotus series, I'd need to analyze each character's portrayal:

Shane Patton (Season 1) is the entitled newlywed who becomes obsessed with getting the suite he booked, ultimately stabbing Armond in a confrontation. His privilege and …rate a certain type of aggressive dominance.

…ason 2) is the wealthy, manipulative finance bro who exerts control …exual conquest, even seducing his friend's wife. His confidence and lack … different kind of social dominance.

…acter introduced in Season 3 set in Thailand. As my knowledge cutoff is … have detailed information about his character traits or behaviors in the …

Based on the first two seasons, both Shane and Cameron exhibit toxic "alpha" behaviors in different ways - Shane through entitlement and Cameron through manipulation. Without complete information about Saxon, I can't make a definitive comparison between all three characters.

Subjectivity in evaluation of long-form text is a feature, not a bug

# Human Evaluation: Limitations

Human evaluation of generations can be hard, but it's still considered the most reliable method today

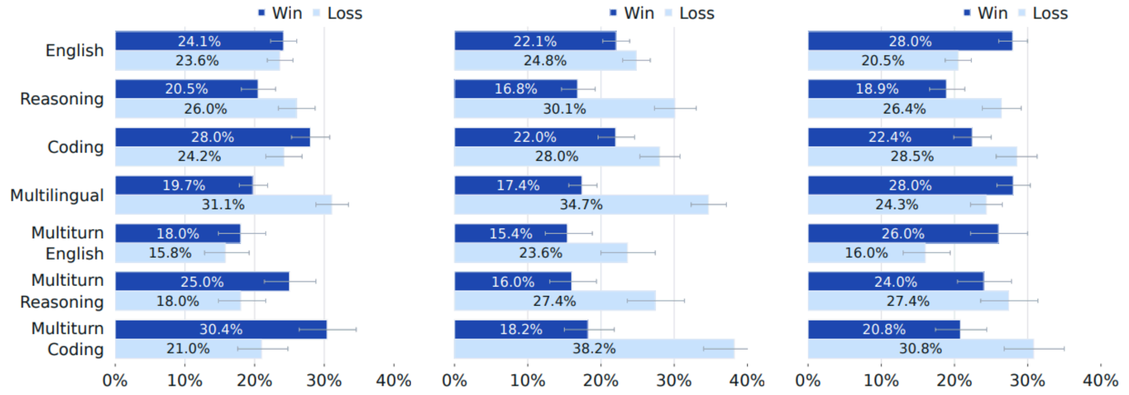



**Figure 17 Human evaluation results for the Llama 3 405B model.** *Left:* Comparison with GPT-4. *Middle:* Comparison with GPT-4o. *Right:* Comparison with Claude 3.5 Sonnet. All results include 95% confidence intervals and exclude ties.

- Of course, we know that human eval is slow and expensive
- Beyond the cost of human eval, it's still fairly hard to conduct, and far from perfect:
  - Results are inconsistent / not reproducible
  - Can be subjective!
  - Misinterpret your question
  - Precision not recall

# Automatic Evaluation with LLM Judges

**AlpacaFarm: A Simulation Framework for Methods that Learn from Human Feedback**

**Yann Dubois*** Stanford
**Xuechen Li*** Stanford
**Rohan Taori*** Stanford
**Tianyi Zhang*** Stanford
**Ishaan Gulrajani** Stanford
**Jimmy Ba** University of Toronto
**Carlos Guestrin** Stanford
**Percy Liang** Stanford
**Tatsunori B. Hashimoto** Stanford

Cheap and theoretically consistent with human evaluation. BUT… reliability? Models evaluating their own generations may lead to weird mode collapsing effect

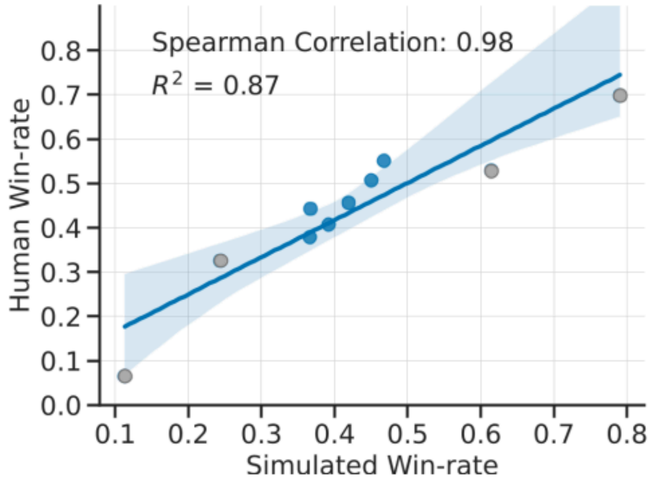



Figure 3: The ranking of methods trained and evaluated in AlpacaFarm matches that of methods trained and evaluated in the human-based pipeline. Each point represents one method $M$ (e.g. PPO). The x-axis shows the simulated evaluation (win-rates measured by $p_{\text{sim}}^{\text{eval}}$) on methods trained in simulation $M_{\text{sim}}$. The y-axis shows human evaluation (win-rates measured by $p_{\text{human}}$) on methods trained with human feedback $M_{\text{human}}$. Gray points show models that we did not train, so their $x$ and $y$ values only differ in the evaluation (simulated vs human). Without those points, we have $R^2 = 0.83$ and a Spearman Correlation of 0.94.

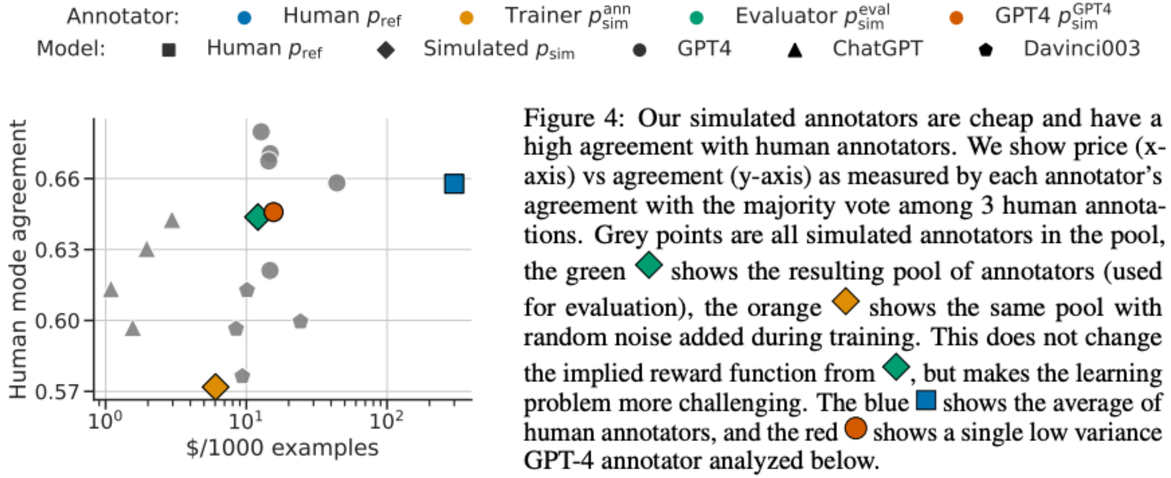



Figure 4: Our simulated annotators are cheap and have a high agreement with human annotators. We show price (x-axis) vs agreement (y-axis) as measured by each annotator's agreement with the majority vote among 3 human annotations. Grey points are all simulated annotators in the pool, the green ◆ shows the resulting pool of annotators (used for evaluation), the orange ◆ shows the same pool with random noise added during training. This does not change the implied reward function from ◆, but makes the learning problem more challenging. The blue ■ shows the average of human annotators, and the red ● shows a single low variance GPT-4 annotator analyzed below.

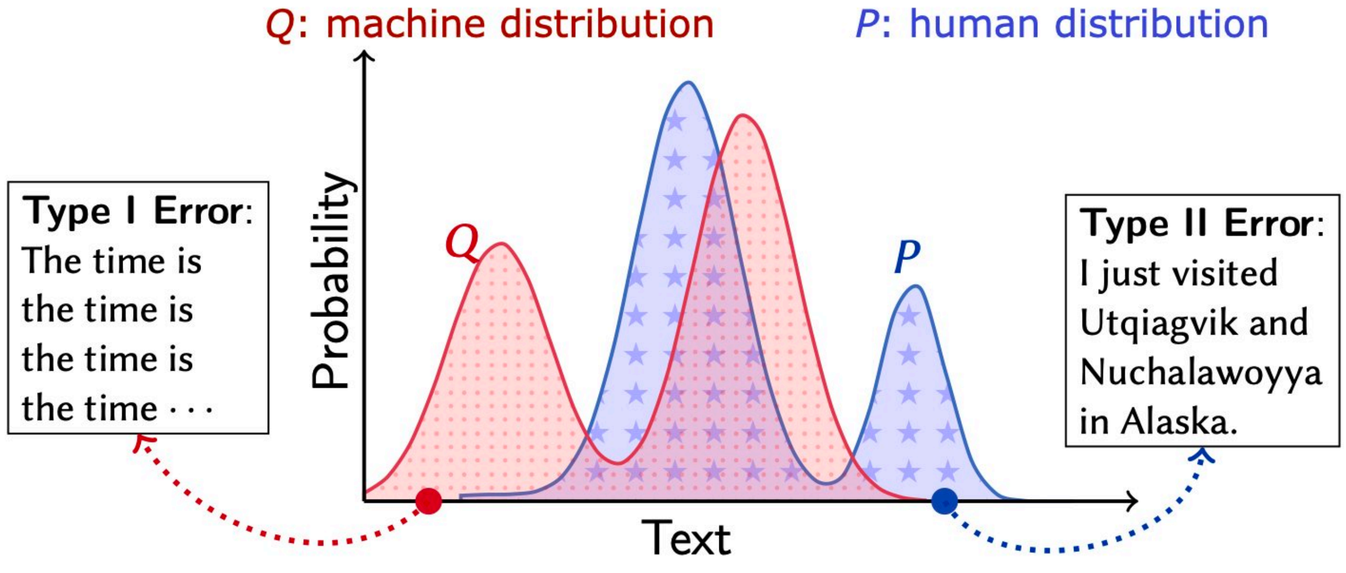
# Evaluating Systems without References

- Compare human / natural language distributions to model-generated language distributions
- Divergence between these two distributions can be measured by MAUVE

**MAUVE: Measuring the Gap Between Neural Text and Human Text using Divergence Frontiers**

**Krishna Pillutla**[1] **Swabha Swayamdipta**[2] **Rowan Zellers**[1] **John Thickstun**[3]
**Sean Welleck**[1,2] **Yejin Choi**[1,2] **Zaid Harchaoui**[4]

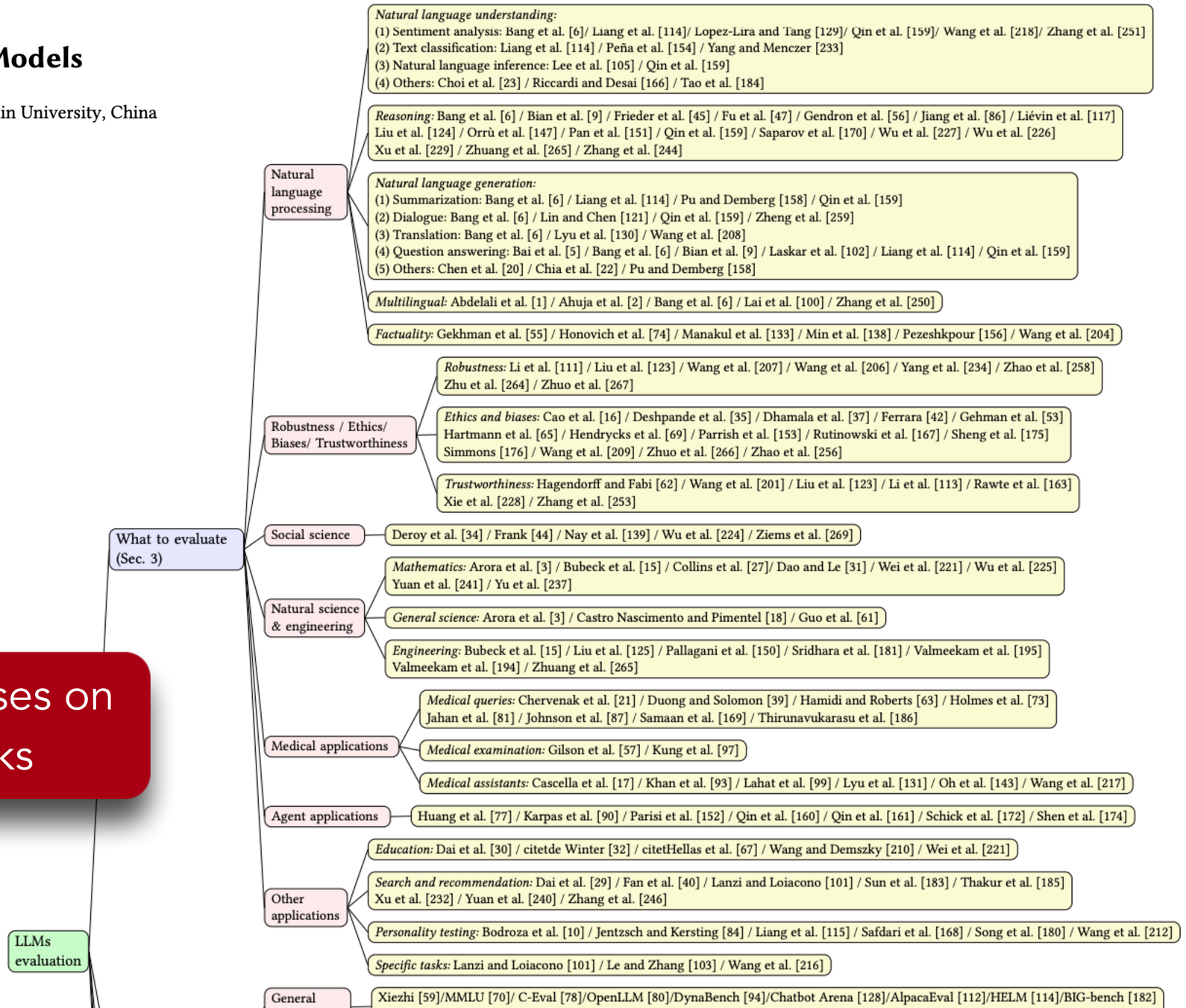
[1]Paul G. Allen School of Computer Science & Engineering, University of Washington
[2]Allen Institute for Artificial Intelligence
[3]Department of Computer Science, Stanford University
[4]Department of Statistics, University of Washington



## A Survey on Evaluation of Large Language Models

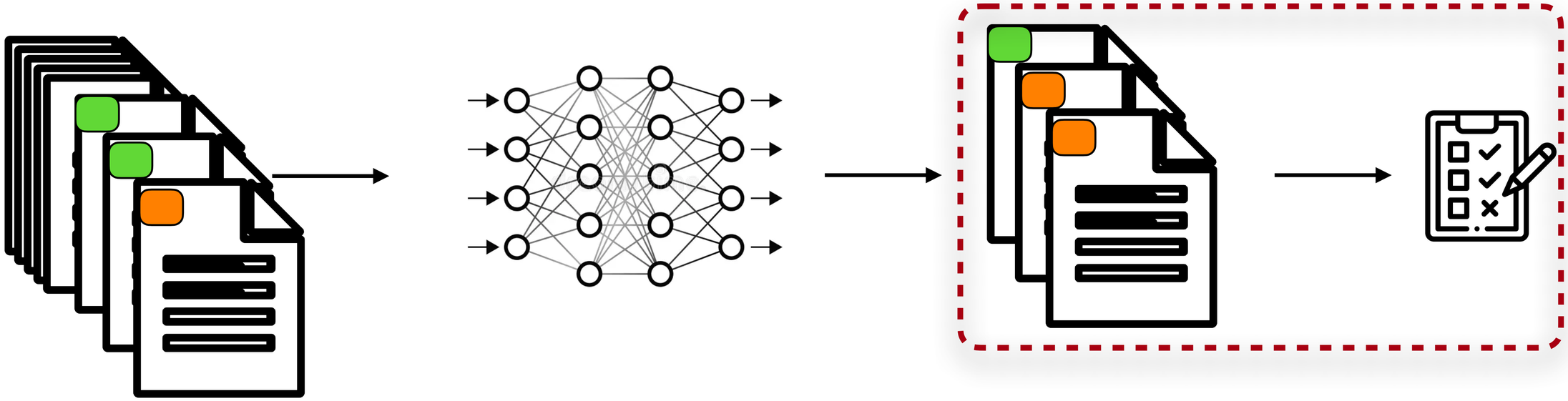
YUPENG CHANG* and XU WANG*, School of Artificial Intelligence, Jilin University, China
JINDONG WANG†, Microsoft Research Asia, China
YUAN WU†, School of Artificial Intelligence, Jilin University, China
LINYI YANG, Westlake University, China
KAIJIE ZHU, Institute of Automation, Chinese Academy of Sciences, China
HAO CHEN, Carnegie Mellon University, USA
XIAOYUAN YI, Microsoft Research Asia, China
CUNXIANG WANG, Westlake University, China
YIDONG WANG, Peking University, China
WEI YE, Peking University, China
YUE ZHANG, Westlake University, China
YI CHANG, School of Artificial Intelligence, Jilin University, China
PHILIP S. YU, University of Illinois at Chicago, USA
QIANG YANG, Hong Kong University of Science and Technology, China
XING XIE, Microsoft Research Asia, China

LLM Evaluation mostly focuses on creating new benchmarks

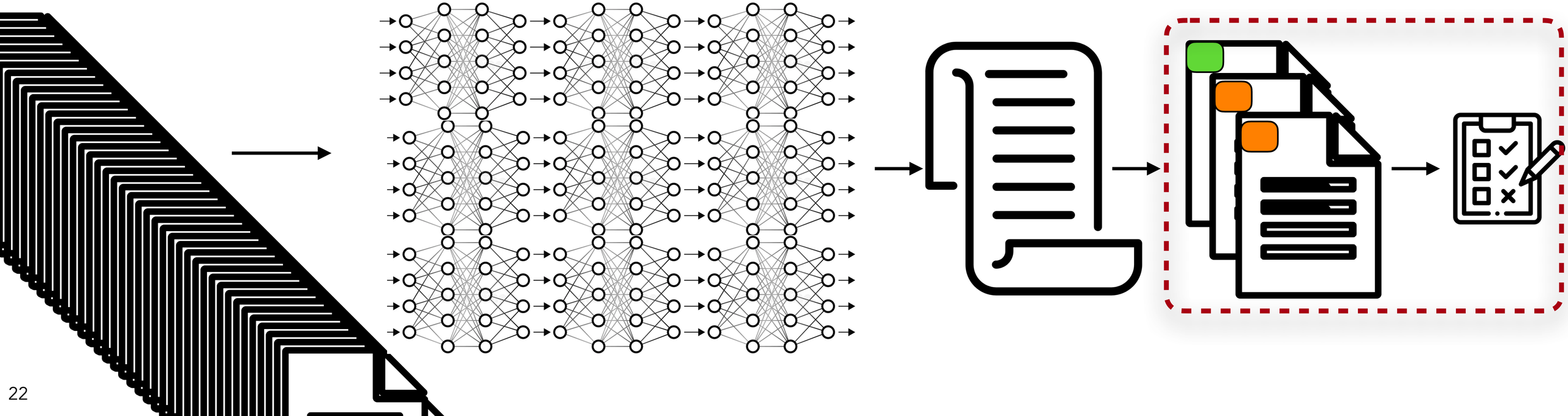


# Benchmarks

Then..

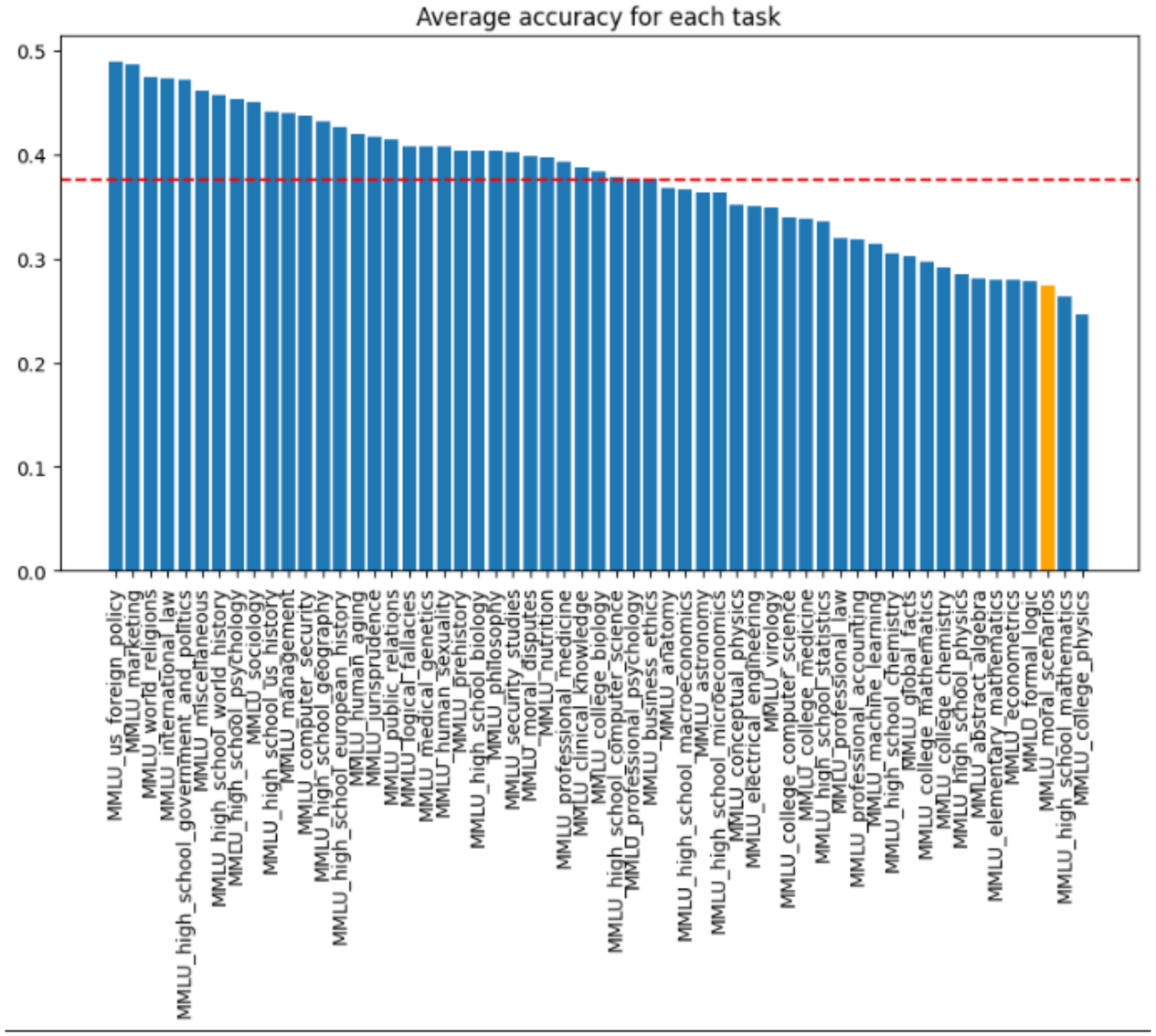


Now..



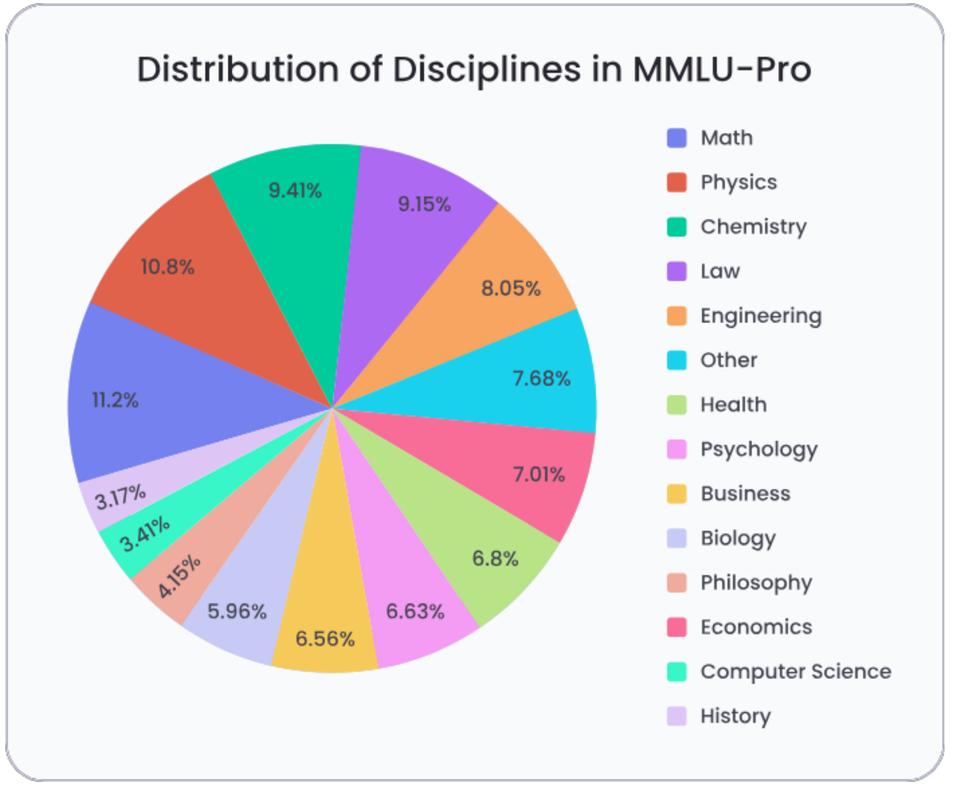
# Evaluation of LLMs

- Almost exclusively on downstream tasks, as opposed to intrinsic metrics
  - Intrinsic metrics, e.g. perplexity
- Few popular multitask benchmarks
  - GLUE - Language Understanding Tasks
  - SuperGLUE - Language Understanding Tasks
  - HellaSwag - Commonsense Reasoning
  - Truthful QA - Fact Verification
  - **MMLU** - Massive Multitask Language Understanding, 15908 knowledge and reasoning questions in 57 areas including medicine, mathematics, computer science, law, and others
  - GSM - 8K Grade School Math
  - BigBench - subsumes some of these benchmarks



# MMLU

- Massive Multi-Task Language Understanding Benchmark
- 57 tasks, across disciplines multiple-choice
- "collected by graduate and undergraduate students from freely available sources online"
- Evaluated on GPT-3 using few-shot prompting
- Critique
  - Overused
  - Really about testing knowledge, not language understanding
- MMLU-Pro [Wang+ 2024]:
  - Removed noisy/trivial questions from MMLU
  - Expanded 4 choices to 10 choices
  - Chain-of-thought prompting




Hendrycks et al. 2020

# Benchmarks Today

GSM8K

MMLU

BigBench

HumanEval

GPQA

Math

HELM

Alpaca2.0

OpenLLMLeaderboard

## 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**

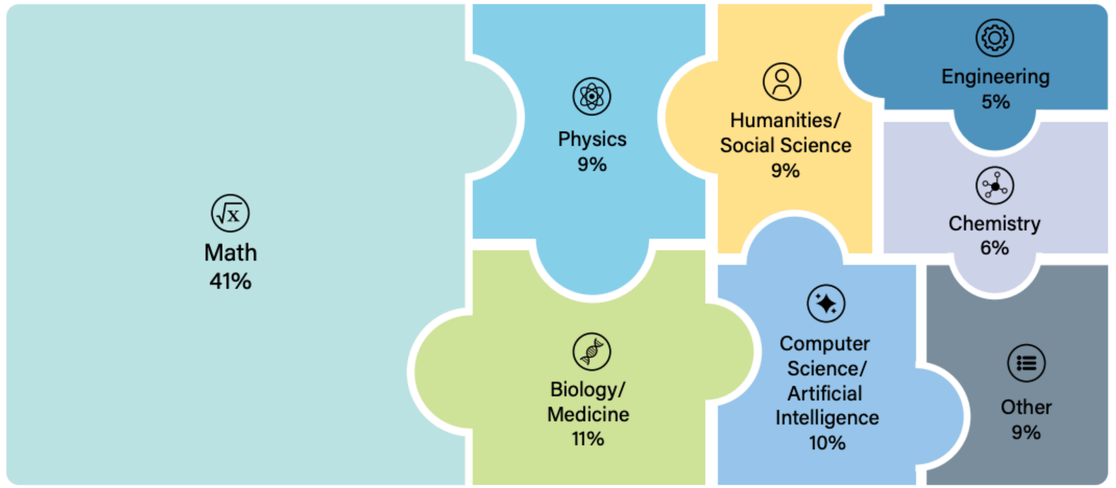
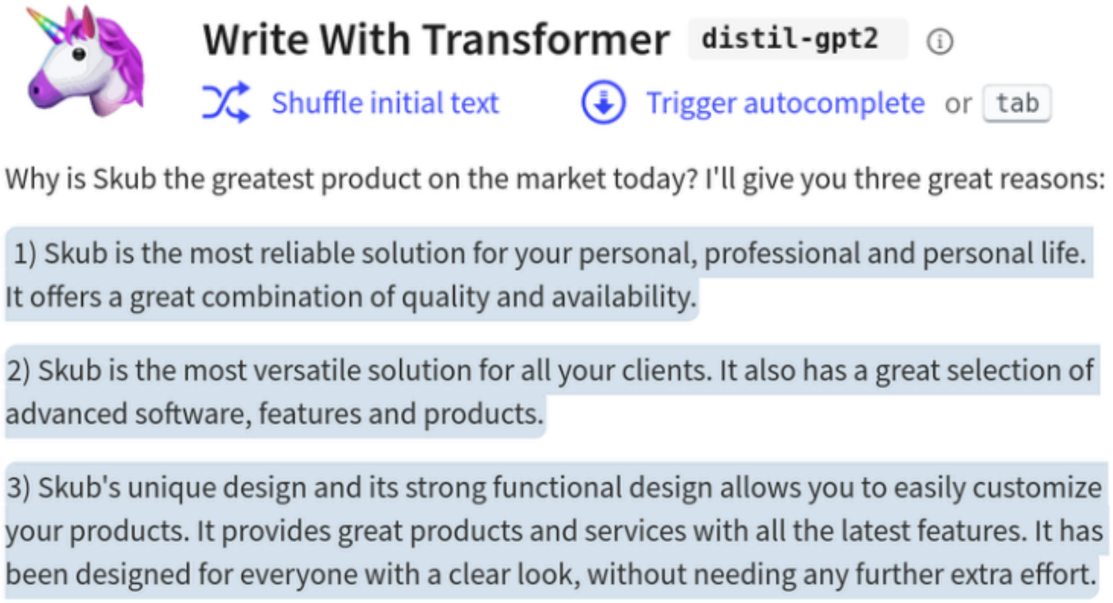
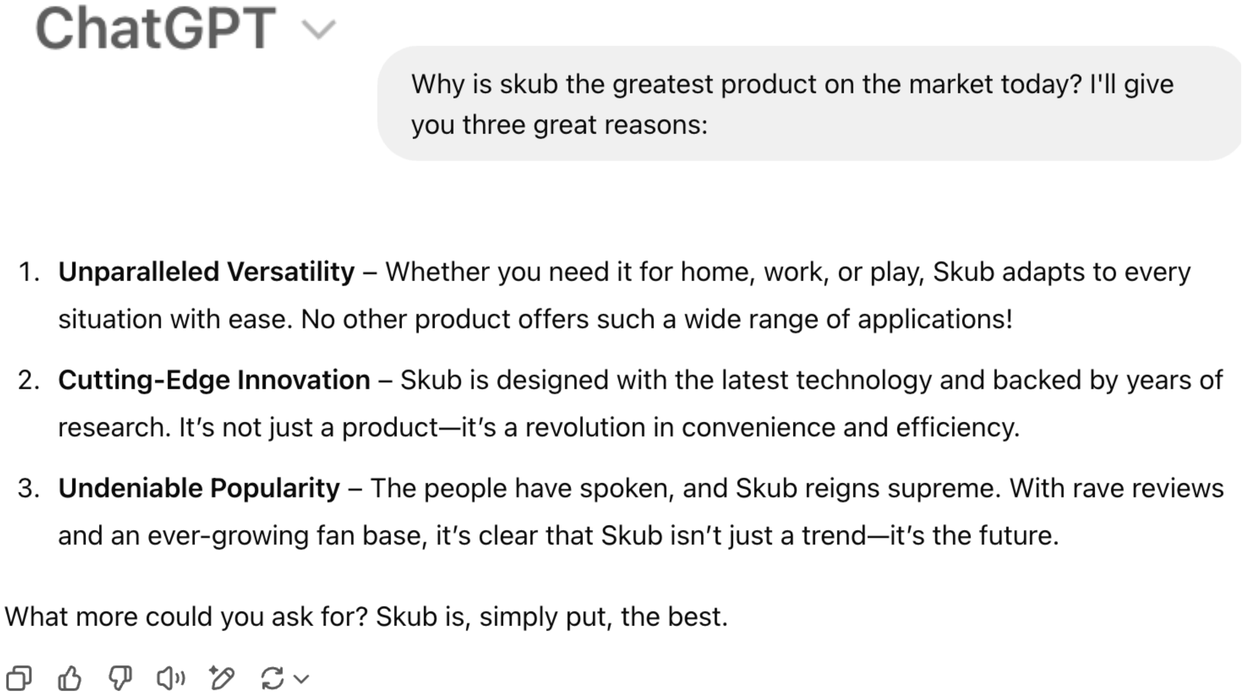



Figure 3: HLE consists of 2,700 exam questions in over a hundred subjects, grouped into high level categories here. We provide a more detailed list of subjects in Appendix B.3.

# Language models are great at generating language



# Current LLM evaluation is done on a few popular benchmarks

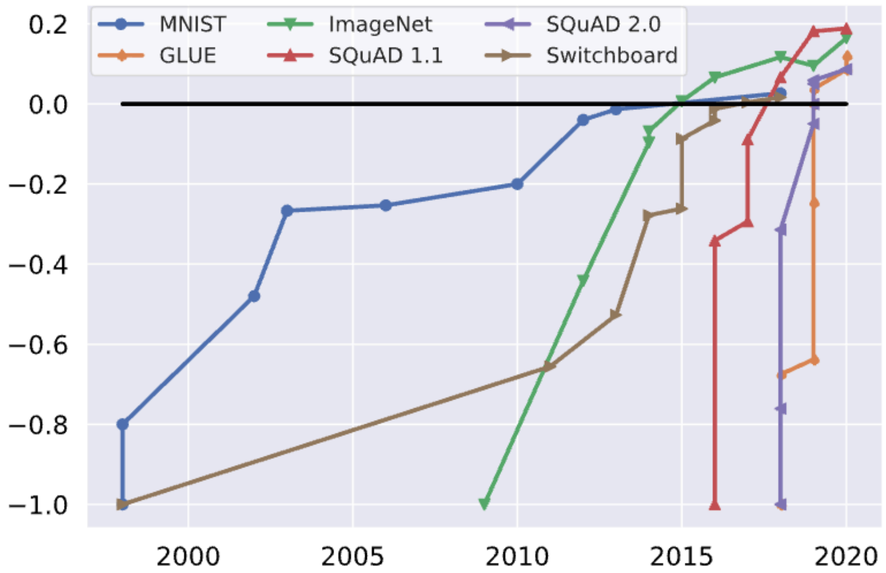



Figure 1: Benchmark saturation over time for popular benchmarks, normalized with initial performance at minus one and human performance at zero.

| | |
|---|---|
| **General** | MMLU (Hendrycks et al., 2021a), MMLU-Pro (Wang et al., 2024b), IFEval (Zhou et al., 2023) |
| **Math and reasoning** | GSM8K (Cobbe et al., 2021), MATH (Hendrycks et al., 2021b), GPQA (Rein et al., 2023), ARC-Challenge (Clark et al., 2018) |
| **Code** | HumanEval (Chen et al., 2021), MBPP (Austin et al., 2021), HumanEval+ (Liu et al., 2024a), MBPP EvalPlus (base) (Liu et al., 2024a), MultiPL-E (Cassano et al., 2023) |
| **Multilinguality** | MGSM (Shi et al., 2022), Multilingual MMLU (internal benchmark) |
| **Tool-use** | Nexus (Srinivasan et al., 2023), API-Bank (Li et al., 2023b), API-Bench (Patil et al., 2023), BFCL (Yan et al., 2024) |
| **Long context** | ZeroSCROLLS (Shaham et al., 2023), Needle-in-a-Haystack (Kamradt, 2023), InfiniteBench (Zhang et al., 2024) |

**Table 16 Post-training benchmarks by category.** Overview of all benchmarks we use to evaluate post-trained Llama 3 models, ordered by capability.

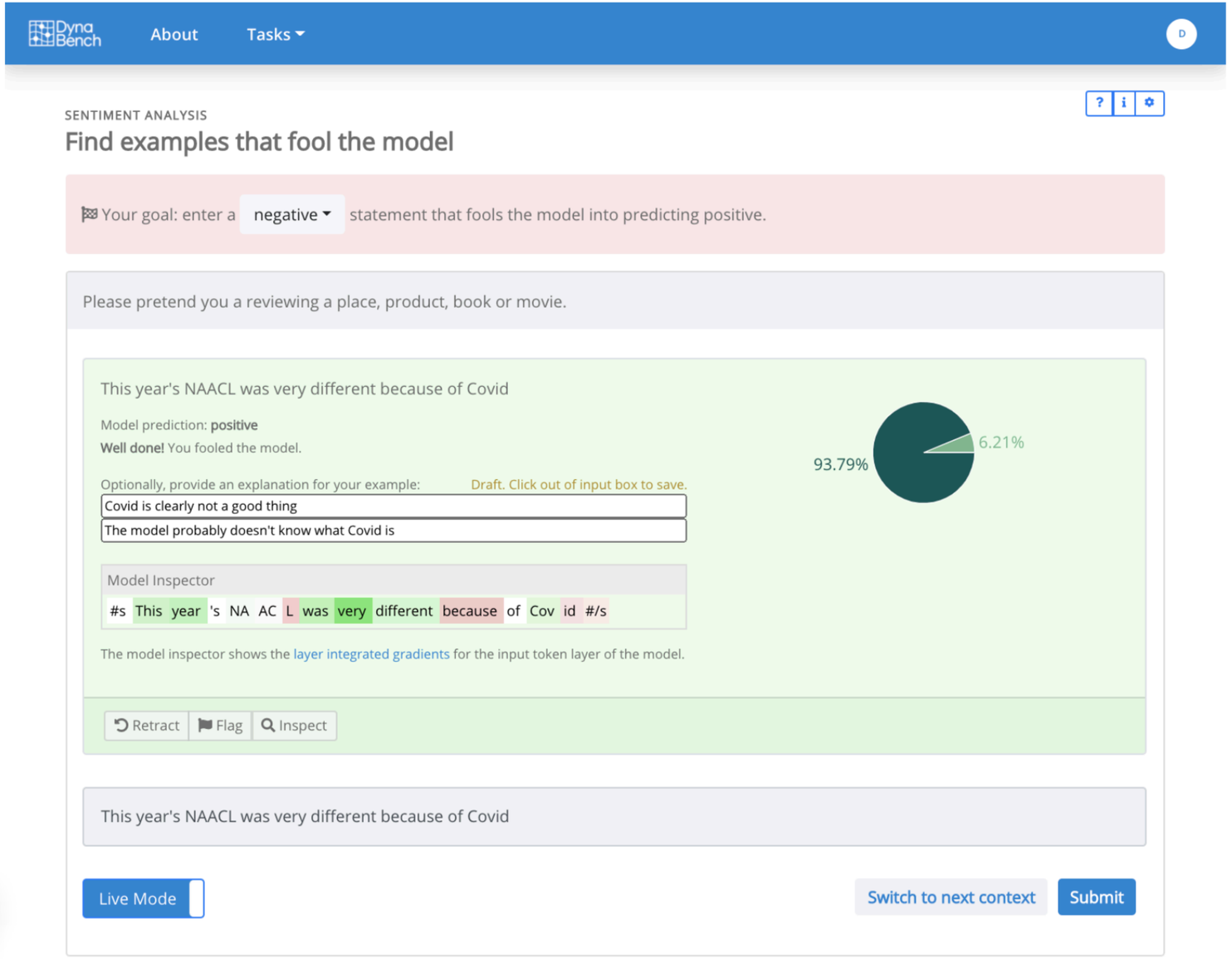
| Model | Prompt | MMLU | GPQA | MATH | HumanEval | MGSM[1] | DROP[1] (F1, 3-shot) | SimpleQA |
|---|---|---|---|---|---|---|---|---|

### 3.1. DeepSeek-R1 Evaluation

| | Benchmark (Metric) | Claude-3.5-Sonnet-1022 | GPT-4o 0513 | DeepSeek V3 | OpenAI o1-mini | OpenAI o1-1217 | DeepSeek R1 |
|---|---|---|---|---|---|---|---|
| | Architecture | - | - | MoE | - | - | MoE |
| | # Activated Params | - | - | 37B | - | - | 37B |
| | # Total Params | - | - | 671B | - | - | 671B |
| English | MMLU (Pass@1) | 88.3 | 87.2 | 88.5 | 85.2 | **91.8** | 90.8 |
| | MMLU-Redux (EM) | 88.9 | 88.0 | 89.1 | 86.7 | - | **92.9** |
| | MMLU-Pro (EM) | 78.0 | 72.6 | 75.9 | 80.3 | - | **84.0** |
| | DROP (3-shot F1) | 88.3 | 83.7 | 91.6 | 83.9 | 90.2 | **92.2** |
| | IF-Eval (Prompt Strict) | **86.5** | 84.3 | 86.1 | 84.8 | - | 83.3 |
| | GPQA Diamond (Pass@1) | 65.0 | 49.9 | 59.1 | 60.0 | **75.7** | 71.5 |
| | SimpleQA (Correct) | 28.4 | 38.2 | 24.9 | 7.0 | **47.0** | 30.1 |
| | FRAMES (Acc.) | 72.5 | 80.5 | 73.3 | 76.9 | - | **82.5** |
| | AlpacaEval2.0 (LC-winrate) | 52.0 | 51.1 | 70.0 | 57.8 | - | **87.6** |
| | ArenaHard (GPT-4-1106) | 85.2 | 80.4 | 85.5 | 92.0 | - | **92.3** |
| Code | LiveCodeBench (Pass@1-COT) | 38.9 | 32.9 | 36.2 | 53.8 | 63.4 | **65.9** |
| | Codeforces (Percentile) | 20.3 | 23.6 | 58.7 | 93.4 | **96.6** | 96.3 |
| | Codeforces (Rating) | 717 | 759 | 1134 | 1820 | **2061** | 2029 |
| | SWE Verified (Resolved) | **50.8** | 38.8 | 42.0 | 41.6 | 48.9 | 49.2 |
| | Aider-Polyglot (Acc.) | 45.3 | 16.0 | 49.6 | 32.9 | **61.7** | 53.3 |
| Math | AIME 2024 (Pass@1) | 16.0 | 9.3 | 39.2 | 63.6 | 79.2 | **79.8** |
| | MATH-500 (Pass@1) | 78.3 | 74.6 | 90.2 | 90.0 | 96.4 | **97.3** |
| | CNMO 2024 (Pass@1) | 13.1 | 10.8 | 43.2 | 67.6 | - | **78.8** |
| Chinese | CLUEWSC (EM) | 85.4 | 87.9 | 90.9 | 89.9 | - | **92.8** |
| | C-Eval (EM) | 76.7 | 76.0 | 86.5 | 68.9 | - | **91.8** |
| | C-SimpleQA (Correct) | 55.4 | 58.7 | **68.0** | 40.3 | - | 63.7 |

# DynaBench

- Open-source platform for dynamic dataset creation and model benchmarking
- Adversarial benchmark generation
  - Create harder and harder examples
- Critique: might lead to out-of-distribution evaluation

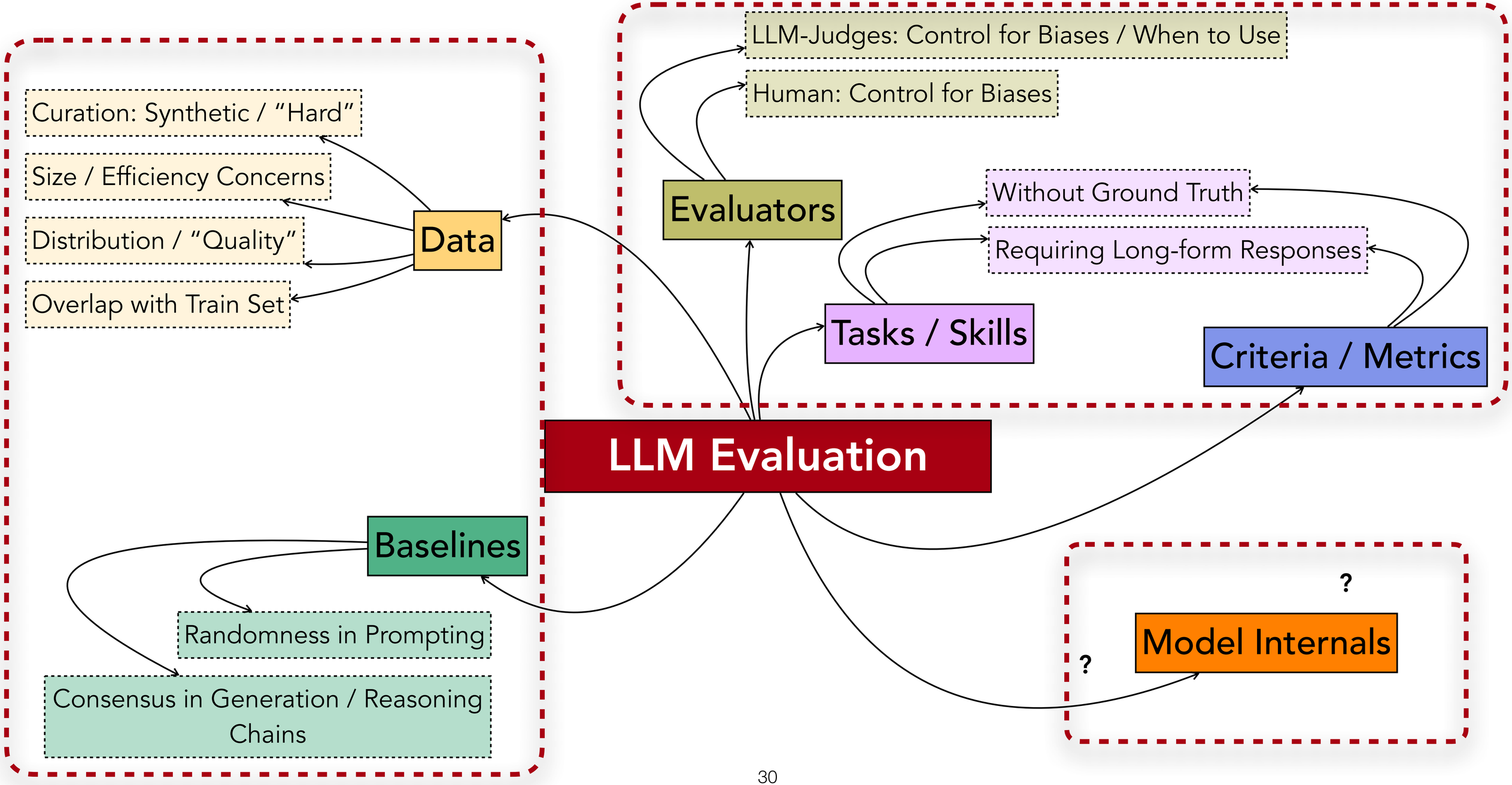



Kiela et al. 2021

# Benchmark Tasks vs. Real-World Tasks

- We have evidence language models are getting better and better at math, inference, coding and many logical / analytical tasks
- How about we go bold and test them directly on real-world tasks with real stakes?
- Caveat: this will involve humans, and much manual labor

**Some of the content in the following slides  contains sensitive content and can be upsetting**


LLM-Judges: Control for Biases / When to Use
Human: Control for Biases
Curation: Synthetic / "Hard"
Size / Efficiency Concerns
Distribution / "Quality"
Overlap with Train Set
Data
Evaluators
Without Ground Truth
Requiring Long-form Responses
Tasks / Skills
Criteria / Metrics
LLM Evaluation
Baselines
Randomness in Prompting
Consensus in Generation / Reasoning Chains
?
Model Internals
?

# My Research

**Annotators with Attitudes:**
**How Annotator Beliefs And Identities Bias Toxic Language Detection**

Maarten Sap[♡♠] Swabha Swayamdipta[♠] Laura Vianna[◇]
Xuhui Zhou[♣] Yejin Choi[♡♠] Noah A. Smith[♡♠]
[♡]Paul G. Allen School of Computer Science, University of Washington, Seattle, WA, USA
[♠]Allen Institute for AI, Seattle, WA, USA
[◇]Department of Psychology, University of Washington, Seattle, WA, USA
[♣]Georgia Institute of Technology, Atlanta, GA, USA

**OATH-Frames: Characterizing Online Attitudes Towards Homelessness with LLM Assistants**

Jaspreet Ranjit[♥] Brihi Joshi[♥] Rebecca Dorn[♥] Laura Petry[♦]
Olga Koumoundouros[♦] Jayne Bottarini[♥] Peichen Liu[♦]
Eric Rice[♦] Swabha Swayamdipta[♥]
[♥]Thomas Lord Dept. of Computer Science, University of Southern California
[♦]Suzanne-Dwork School of Social Work, University of Southern California

**We're Afraid Language Models Aren't Modeling Ambiguity**

Alisa Liu[♡] Zhaofeng Wu[♦] Julian Michael[♠] Alane Suhr[♣△] Peter West[♡♣]
Alexander Koller[♣♥] Swabha Swayamdipta[◇] Noah A. Smith[♡♣] Yejin Choi[♡♣]
[♡]Paul G. Allen School of Computer Science & Engineering, University of Washington
[♣]Allen Institute for AI [◇]University of Southern California [△]UC Berkeley
[♥]Saarland University [♠]New York University [♦]Massachusetts Institute of Technology

*Evaluating Capabilities where there are No Right Answers:*

- Evaluating free-text explanations / reasoning chains
- Evaluating LLM assistants in real world problems (e.g. social work, emergency response)
- Evaluating personalized agents, simulations
- Future: Biases in Human Evaluators of LLMs

*Tasks / Skills*

*Evaluators*

*Criteria / Metrics*

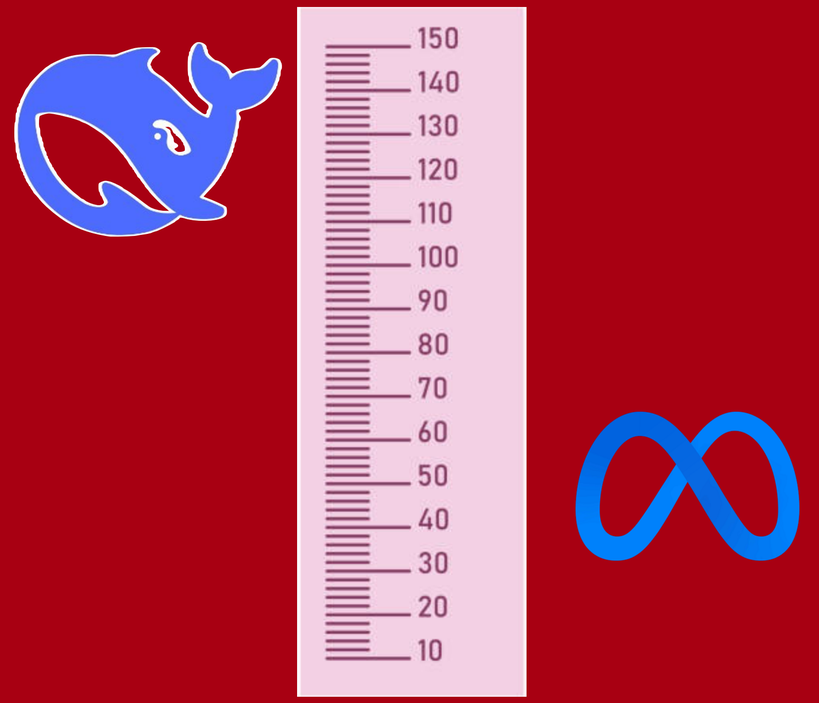
**Reframing Human-AI Collaboration for Generating Free-Text Explanations**

Sarah Wiegreffe[♣] Jack Hessel[†] Swabha Swayamdipta[†]
Mark Riedl[♣] Yejin Choi[◇†]
[♣]School of Interactive Computing, Georgia Institute of Technology
[†]Allen Institute for Artificial Intelligence
[◇]Paul G. Allen School of Computer Science and Engineering, University of Washington

**REV: Information-Theoretic Evaluation of Free-Text Rationales**

Hanjie Chen[♡*] Faeze Brahman[♠◇] Xiang Ren[♠♣] Yangfeng Ji[♡]
Yejin Choi[♠◇] Swabha Swayamdipta[♣]
[♡]Department of Computer Science, University of Virginia
[♠]Allen Institute for AI [♣]University of Southern California
[◇] Paul G. Allen School of Computer Science & Engineering, University of Washington

# Rethinking Evaluation Benchmarks




*Benchmarks vs. real-world tasks*

Need AI help for local politicians to understand public attitudes on homelessness to frame the most resonant message to inform public policy.

USC Annenberg
School for Communication and Journalism

Easy peasy, we've got these amazing language models - let me do this for you in a couple of weeks, tops

USC Viterbi
School of Engineering

@mention how about do something for all the homeless veterans and americans on the streets? you'll give millions to the immigrants crossing our so called secure border but our homeless americans get nothing. you are merely trying to buy votes

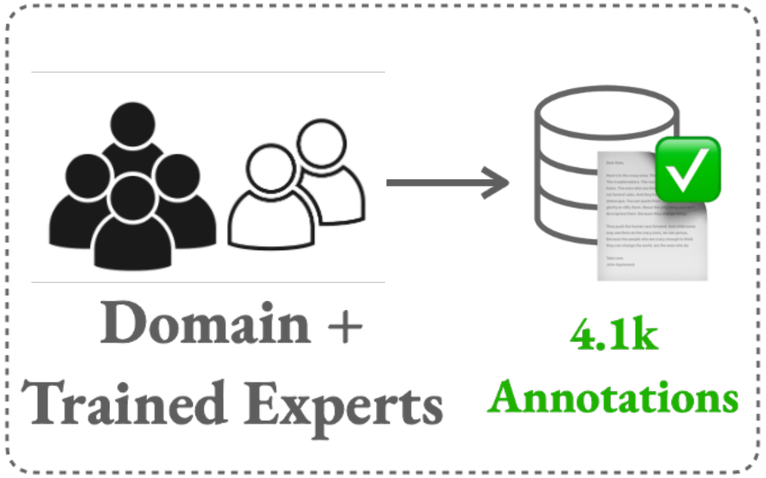
government critique | money allocation | deserving/undeserving

**~2.5M tweets on homelessness between 2021-2023**

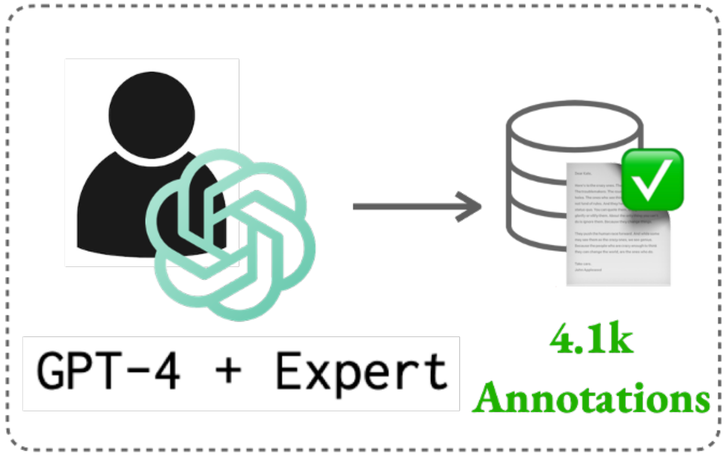


Ranjit, … & **Swayamdipta.** EMNLP 2024.

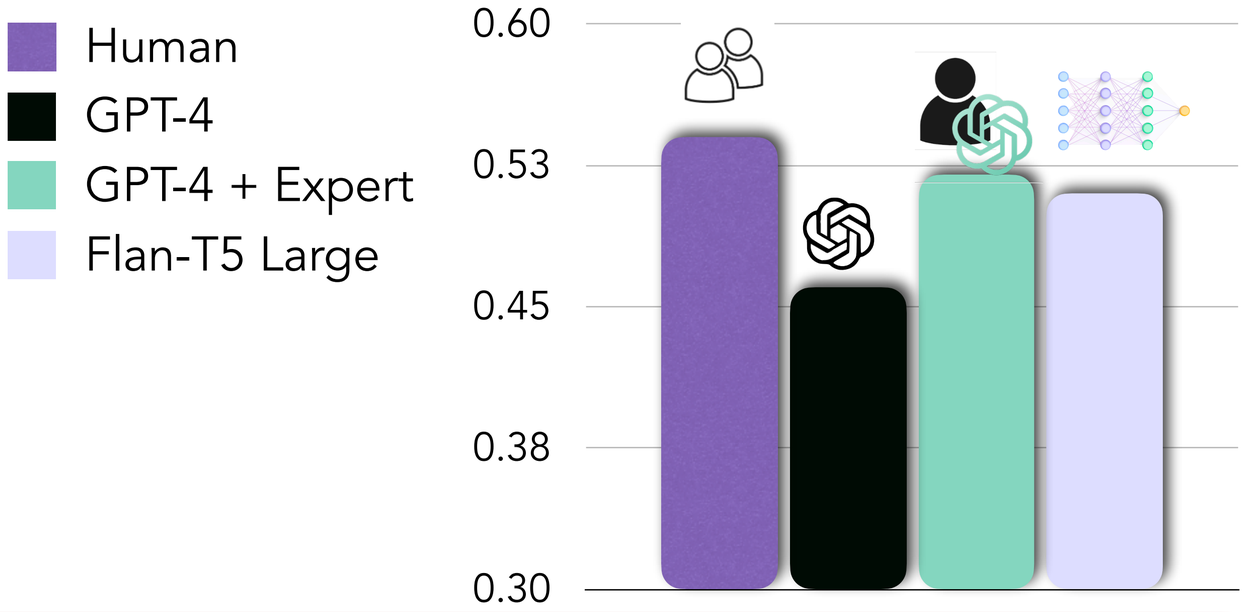
# LMs in Communications / Social Work: Informing policies on homelessness




**Building a test set took more than 7-8 months of laborious manual work**




**Results in 6.5x speedup in annotation time!**




With some effort, language models can be used as assistants for determining variables which could inform communication and public policy

Ranjit, … & **Swayamdipta.** EMNLP 2024.

Heard you have been helping out colleagues with language models. I need some help with analyzing suicide reports that help me validate my hypotheses on novel factors for designing suicide interventions.

USC Suzanne Dworak-Peck School of Social Work

**How often do victims interact with non-clinical personnel (e.g. legal professionals) in days before their death?**

**270K suicide reports**

National Violent Death Reporting System NVDRS

[redacted] family custody child support battle with her child [redacted]

POLICE DEPARTMENT OF THE ARMY UNITED STATES OF AMERICA

No legal interaction

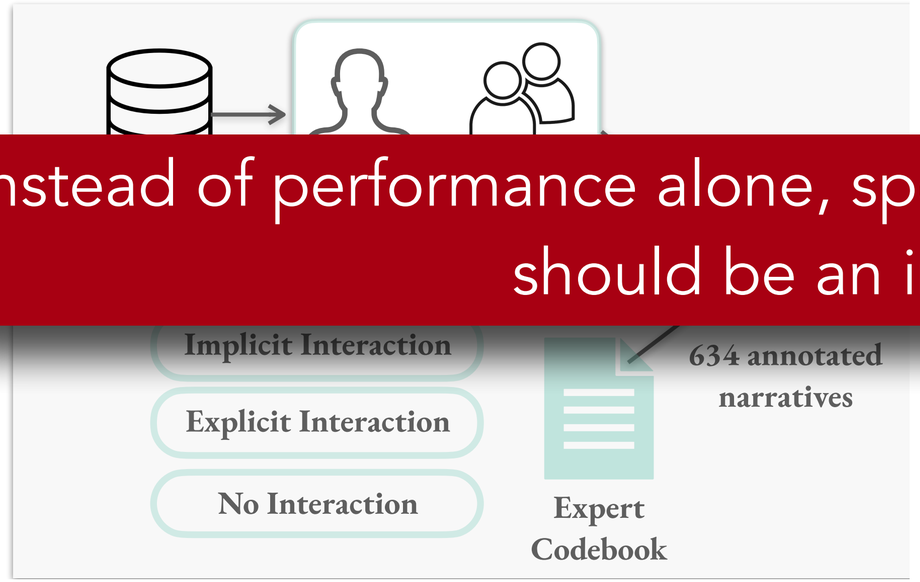
Explicit legal interaction

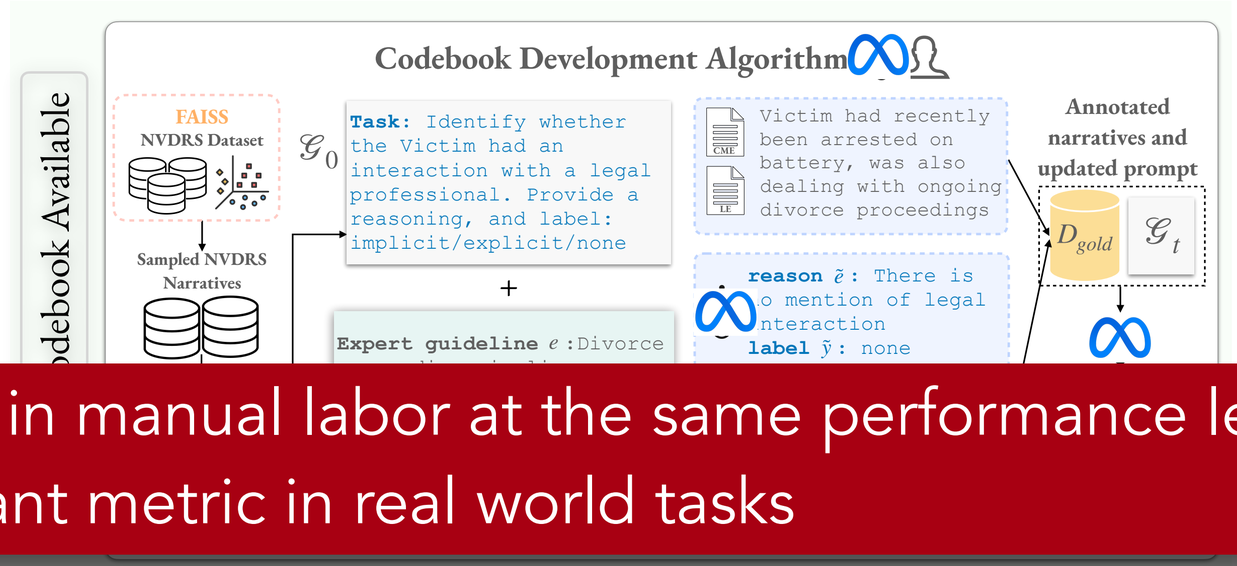
Implied legal interaction



Ranjit, … & **Swayamdipta.** NeurIPS GenAI for Health. 2025.

# LMs in Social Work




**Instead of performance alone, speedup in manual labor at the same performance level should be an important metric in real world tasks**

**Speedup in process: weeks to hours!**

| Model | $\mathcal{G}_0$ | $\mathcal{G}_{\text{expert}}$ | $\mathcal{G}_{12}$ |
|---|---|---|---|
| Meta-Llama-3-70B | 0.57 | 0.78 | 0.79 |
| Qwen2.5-32B | 0.68 | 0.77 | 0.77 |
| Qwen2.5-14B | 0.63 | 0.77 | 0.76 |

**10.4% of 270K suicide narratives had evidence of legal interactions!**

Our research was going at the pace of a bullock cart and you came at it with a jet plane!

USC Suzanne Dworak-Peck School of Social Work

Ranjit, … & **Swayamdipta.** NeurIPS GenAI for Health. 2025.

# Public Health: Emergency Response

As a volunteer emergency responder, I'd like to automate part of the emergency response process using language models

**Task 1: Chemical Translation**

A container labeled $H_2SO_4$ is leaking with a sharp, choking odor. We need help!

What is the common name of $H_2SO_4$?

Search online.

30 seconds later

Am I ready to provide accurate support within seconds?

The common name of $H_2SO_4$ is Sulfuric Acid.

30 seconds later

**Task 2: Incident Response**

How do I best respond to an incident involving Sulfuric Acid?

Find "Sulfuric Acid" in a 400-page Emergency Response Guidebook (ERG).

1 minute later

Hazard: Highly corrosive and reactive. Recommended PPE: Latex gloves, face shield, and respirator. Action: Evacuate a 10-foot radius around the spill, ensure proper ventilation, and establish a 300-foot exclusion zone downwind.

30 seconds later

**Task 3: HazMat QA**

I'm first on scene with no fire extinguisher. What's the safest way to handle a Sulfuric Acid fire?

Wait for someone with more resources or experience to arrive.

3 minutes later

Use a dry chemical instead of a $CO_2$ fire extinguisher.

30 seconds later


| Model | I. Translation | | II. Incident Response | | | III. Exam |
|---|---|---|---|---|---|---|
| | | | | | | Acc. (↑) |
| | | | | | | 49.0 |
| | | | | | | 47.3 |
| | | | | | | 60.0 |
| | | | | | | 58.0 |
| | | | | | | 60.0 |
| GPT-4o | 92.4 | **68.0** | **52.7** | 80.2 | **2.0** | **63.9** |

Real-world tasks are still hard to evaluate directly, need to be broken down.
And human involvement is still important

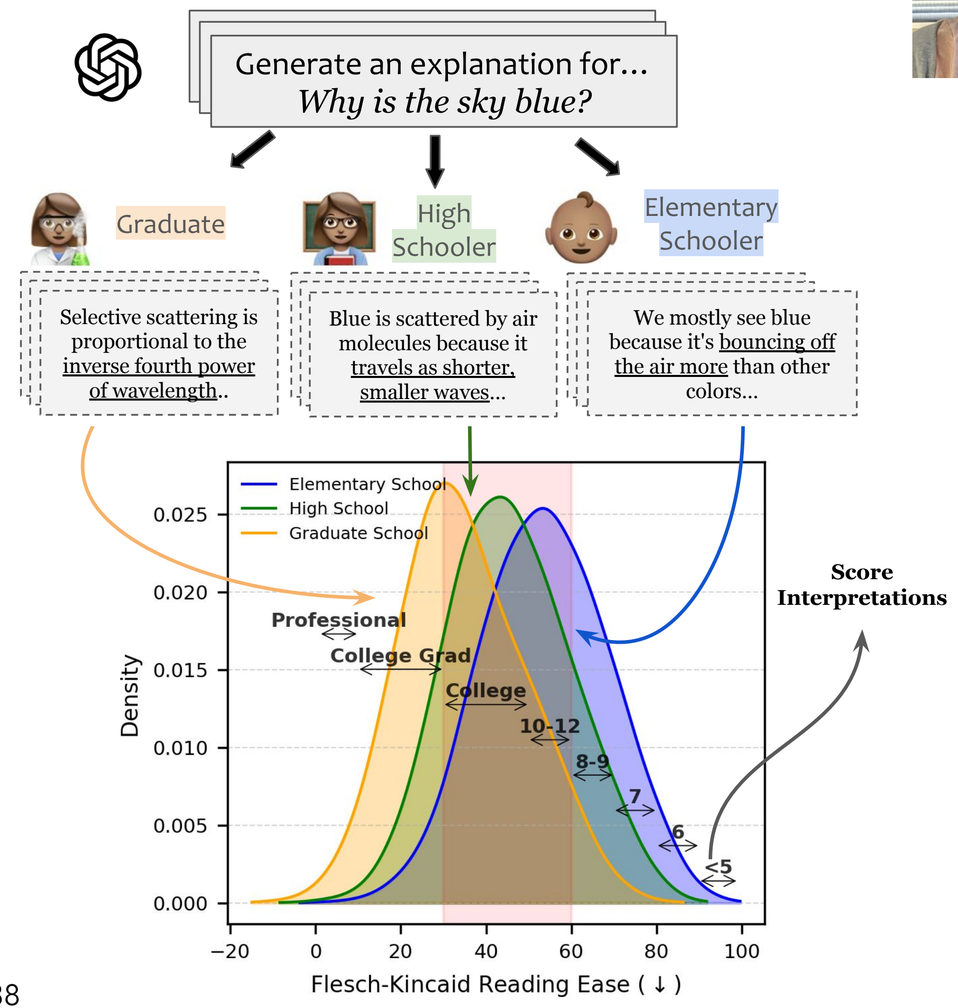
Surana, Ye & **Swayamdipta.** NeurIPS LLM-Eval. 2025.

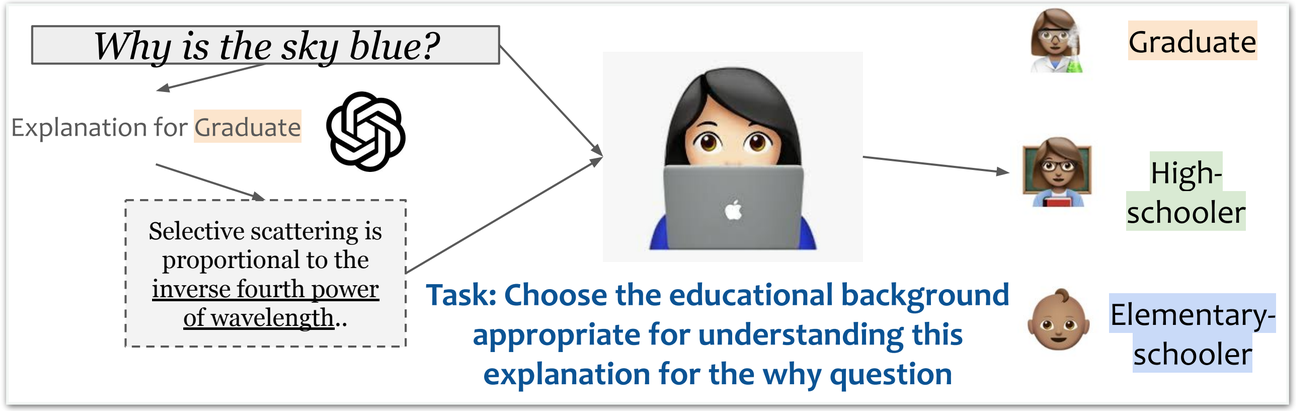
# Education

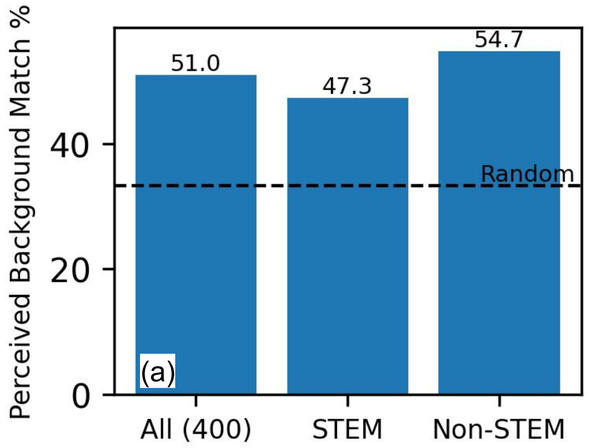
I feel very shy around my teacher, so I ask all my questions to ChatGPT.

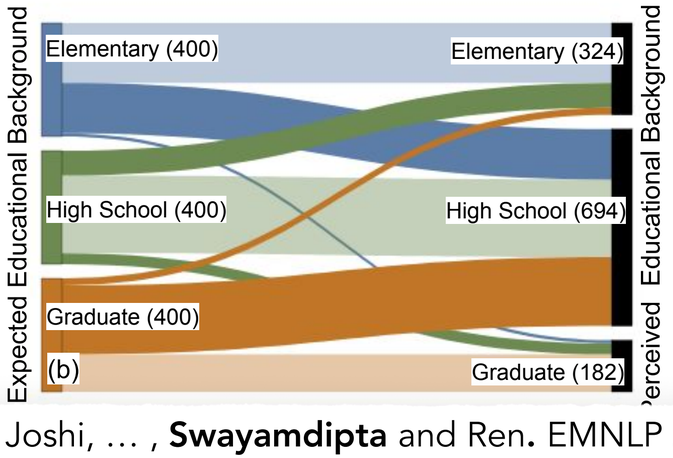
Umm... does it help you every time?

I dunno...




Joshi, … , **Swayamdipta** and Ren. EMNLP 2025

# Evaluator Efforts and Effects on Human Evaluation

- Human evaluators seem to like language model responses
  - especially those with limited training / experience
  - experts tend to disagree more with LMs, but can often miss important details too

**The Impact of Generative AI on Critical Thinking: Self-Reported Reductions in Cognitive Effort and Confidence Effects From a Survey of Knowledge Workers**

Hao-Ping (Hank) Lee
Carnegie Mellon University
Pittsburgh, Pennsylvania, USA
haopingl@cs.cmu.edu

Advait Sarkar
Microsoft Research
Cambridge, United Kingdom
advait@microsoft.com

Lev Tankelevitch
Microsoft Research
Cambridge, United Kingdom
levt@microsoft.com

Ian Drosos
Microsoft Research
Cambridge, United Kingdom
t-iandrosos@microsoft.com

Sean Rintel
Microsoft Research
Cambridge, United Kingdom
serintel@microsoft.com

Richard Banks
Microsoft Research Cambridge
Cambridge, United Kingdom
rbanks@microsoft.com

Nicholas Wilson
Microsoft Research
Cambridge, United Kingdom
niwilson@microsoft.com

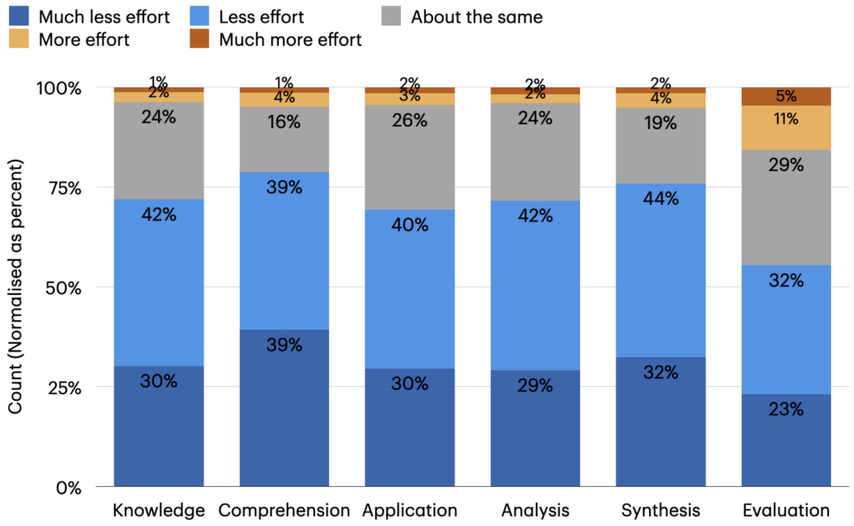
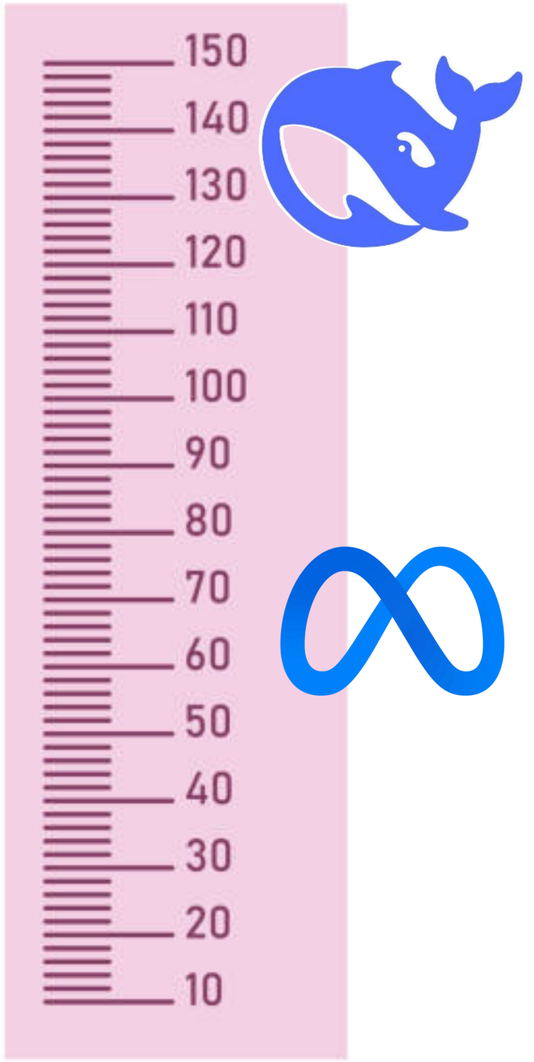



**Figure 2: Distribution of perceived effort (%) in cognitive activities (based on Bloom's taxonomy) when using a GenAI tool compared to not using one.**

# Rethinking Benchmark Tasks

- Our notion of task difficulty might have limitations; using real-world problems to guide tasks is as important as collecting “hard” benchmarks.
  - Over-reliance on a few select benchmarks (each with a single aggregate metric) might give us a false sense of progress
- Quantitative evaluation of language models on real-world tasks is hard, requires humans with domain expertise
  - Task itself needs to be broken down into smaller parts for ease of evaluation
- Harder tasks (no right answers, experts might disagree): Human verification, real data
  - Easier tasks (possible right answers): Auto evals, synthetic data



# Evaluation of Language Models

- Important task, if there ever was one! Given a fixed model, how "good" is it?
- Perplexity is still of extreme importance, even though an intrinsic metric
- TL;DR
  - There is no one true evaluation; choose the evaluation depending on what you're trying to measure.
  - Always look at the individual instances and the predictions.
  - There are many aspects to consider: capabilities, safety, costs, realism.
  - Clearly state the rules of the game (methods versus models/systems).
- Benchmarks vs. Vibes