

# Building LLM Reasoners

## Lecture 11: LLM Evaluation

Greg Durrett



Administrative details and recap

What does accuracy mean?

Statistical Significance

Repeated Testing

Human (and LLM) Evaluation

Principles of Benchmarking

Benchmark Contamination



Administrative details and recap

What does accuracy mean?

Statistical Significance

Repeated Testing

Human (and LLM) Evaluation

Principles of Benchmarking

Benchmark Contamination

# Administrative details and recap

# Administrivia

---

- ▶ Compute issues for final project: email me
- ▶ Final project check-ins due next Wednesday April 22
- ▶ Final project presentations in two weeks
- ▶ Final projects due May 8
- ▶ A3, quiz 3 grading underway

# Today: Evaluation

---

- ▶ During your projects, you will think about doing your own freeform evaluation of how well your system works
- ▶ Today: we will discuss principles of evaluation: understanding metrics, significance testing, how to think about benchmarks, and more

Administrative details and recap

What does accuracy mean?

Statistical Significance

Repeated Testing

Human (and LLM) Evaluation

Principles of Benchmarking

Benchmark Contamination

# What does accuracy mean?

# Interpreting Accuracy

---

Suppose I told you we had a hard benchmark dataset and a system gets 25% on it. What do you want to know?

## Baselines

- ▶ Random
- ▶ Majority class
- ▶ Simpler method
- ▶ Prompted LLM

# Baselines

---

Suppose it's a math dataset. Assume a grader like that in Assignment 3.

Random baseline (pick an integer from 0 to 10) gets 1% accuracy

Majority baseline (answering 0) is 2%

System from prior work: 25%

Our system: 27%

Is our system better? What do you need to know?

# Baselines

---

Suppose the prior work model gets 25%, our system gets 27%, and our system really is better (dataset is very large).

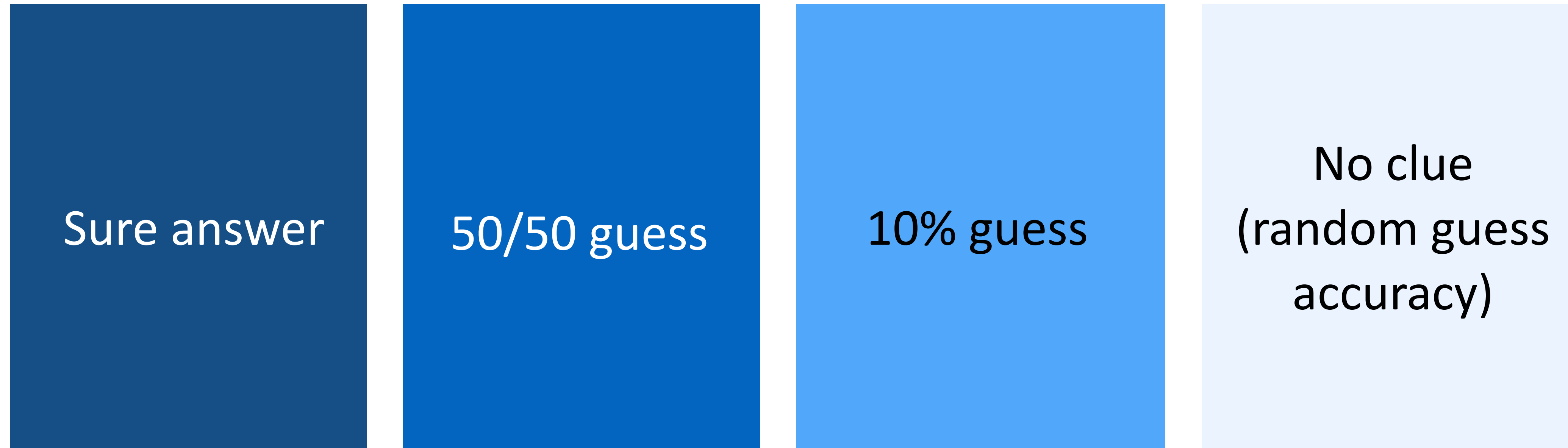
Can we tell \*why\* by looking at the examples our system gets right and the other system gets wrong?

What if:

- ▶ Our systems gets the same 25% correct as the base model, +2% more?
- ▶ 10% are correct from both but the other (15% / 17%) are only gotten right by that system?

# Baselines

---



- ▶ There's some progression of example difficulty across most datasets
- ▶ \*True\* random might be low...but maybe there are a few plausible answers for a lot of the questions

# Analyzing Errors

---

Looking at differences of example sets is likely to lead to a lot of noise

What you can do:

- ▶ Find meaningful subsets of the data where model performance differs. E.g., “on number theory examples, the baseline gets 20% and the better system gets 50%”. (But how many number theory examples are there?)
- ▶ Character the behavior of the model independent of its answer (e.g., how it reasons in a chain-of-thought)

Administrative details and recap

What does accuracy mean?

Statistical Significance

Repeated Testing

Human (and LLM) Evaluation

Principles of Benchmarking

Benchmark Contamination

# Statistical Significance

# Statistical Significance

---

Now suppose I get 25% accuracy and my baseline gets 23% accuracy.

- ▶ What if I have 10000 examples?
- ▶ What if I have 100 examples?
- ▶ Why is this a problem? We did better right?

# Distribution of Test Statistics

---

Datapoints  $D = (x, y) \sim P(x)P(y|x)$ ,  $x$ 's are datapoints and  $y$ 's are labels

Suppose our test set is 100 examples

**Question:** If a model gets 60% on this dataset, how reliable is that value?

$\text{is\_correct}(y, y') = 1$  if  $y=y'$ , else 0

60% =  $\text{mean}([\text{is\_correct}(y_1, \text{model}(x_1)), \text{is\_correct}(y_2, \text{model}(x_2)), \dots])$  on dataset D

Call this statistic  $\mu = \text{acc}(D)$  (for shorthand)

**What is the distribution of  $\mu$ ?**

# Bootstrap

---

Bootstrap: estimate the sample mean by **resampling the underlying data with replacement**

Original sample  $D = (x_i, y_i)$  for  $i=1$  to 100

For each of 10,000 samples (some large number):

Resample  $D' = \{(x_i', y_i')\}$  from  $D = \{(x_i, y_i)\}$

Compute  $\mu' = \text{mean}(\text{acc}(D'))$

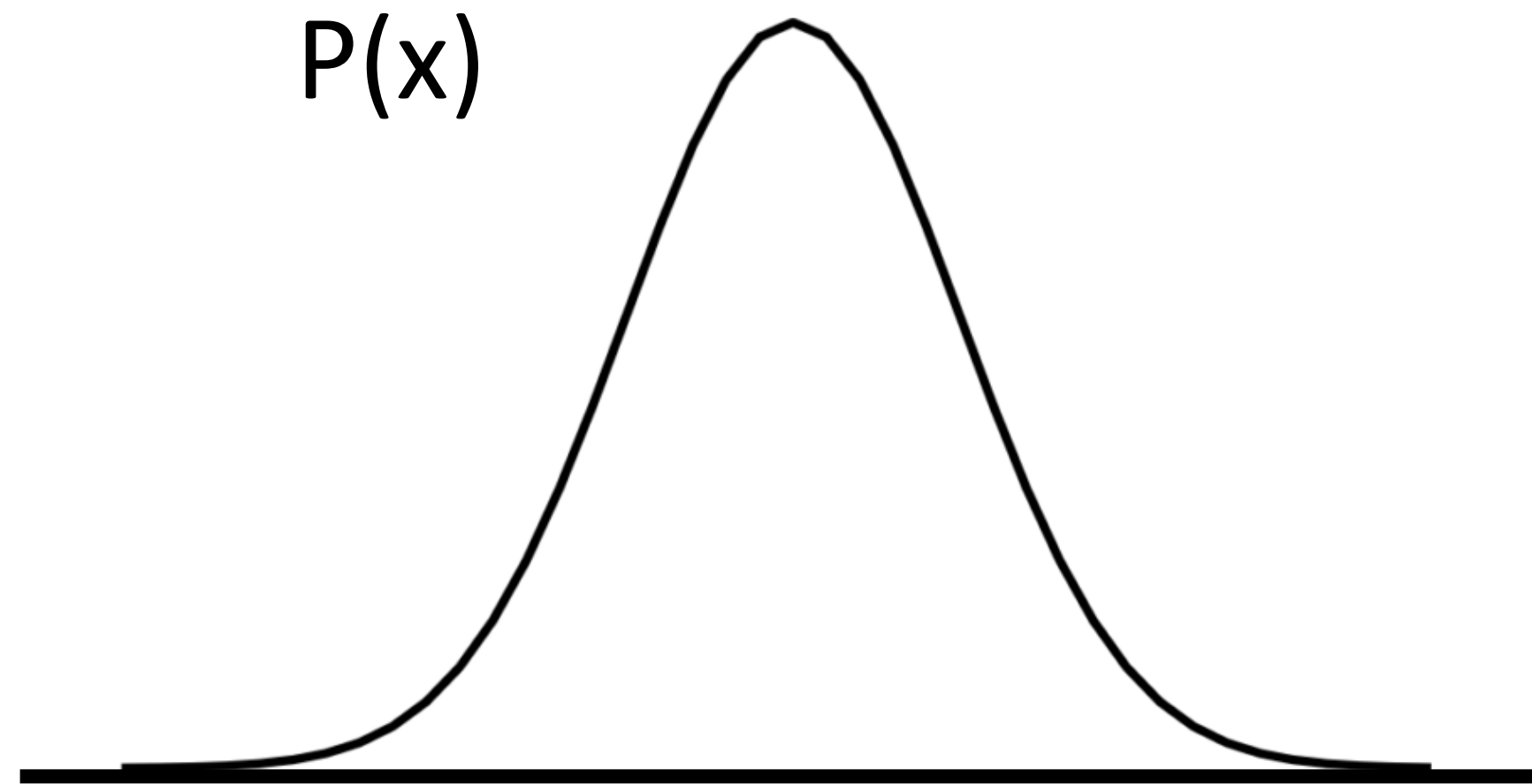
Store  $\mu'$

Compute a 95% confidence interval over the stored  $\mu'$ s

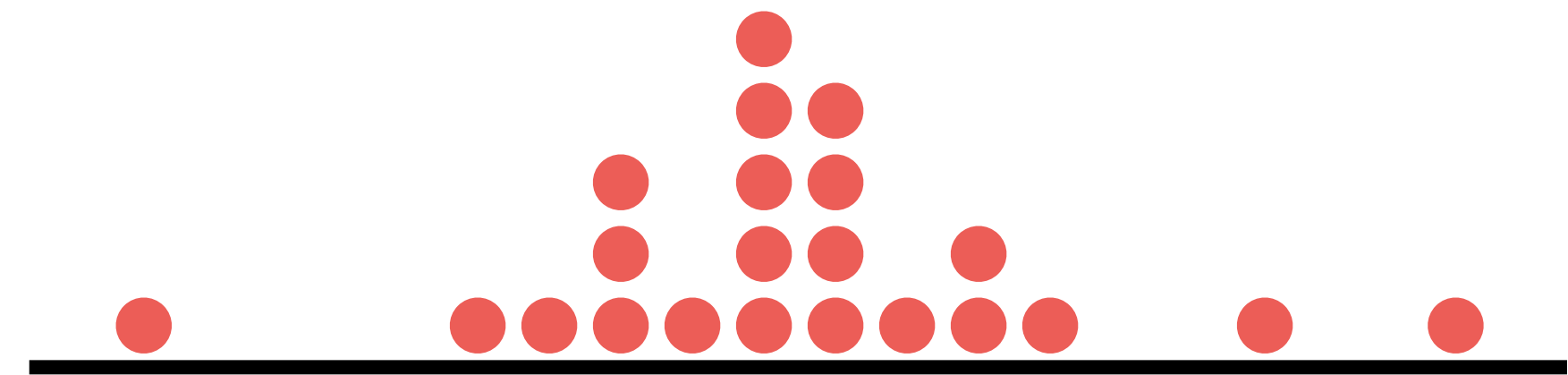
# Bootstrap: Intuition

---

What we really want to do: draw many different samples  $\mu' = \text{mean}(D')$  by drawing a new  $D' \sim P(x)P(y|x)$  from the underlying distribution. But we don't have real access to that, so instead we resample.



Empirical distribution D



We substitute resamples from  $D$  in place of resamples from  $P(x)$ . This is called a plug-in estimator: we're plugging in the empirical distribution  $D$  instead of the real distribution  $P(x)$  (and  $D$  is sampled from  $P(x)$ , so asymptotically this is fine)

# Bootstrap Confidence Interval: Closed Form

---

Our 100 data points look like [0, 0, 1, 1, 1, 0, 1, 1, 0, 0 ...] with 60% 1s

What's the distribution of  $\mu'$  if  $\mu'$  is the mean of 100 samples from this dataset?

What's the std dev of  $\mu'$ ?

Suppose our two systems differ by 5%. How big a test set do we need?

# Paired Bootstrap

---

Suppose we have two systems:

System 1: 60% accuracy, 95% CI (50%, 70%)

System 2: 65% accuracy, 95% CI (55%, 75%)

Is System 2 really no better than System 1?

# Paired Bootstrap

---

Null hypothesis: System 2 is not better than System 1

Original means  $\mu_1, \mu_2$ , suppose  $\mu_2 - \mu_1 = \Delta$

Original sample  $D = (x_i, y_i)$  for  $i=1$  to 100

For each of 10,000 samples (some large number):

Resample  $D' = \{(x_i', y_i')\}$  from  $D = \{(x_i, y_i)\}$

Compute  $\mu_1' = \text{mean}(\text{acc}(\text{System 1}(D')))$

Compute  $\mu_2' = \text{mean}(\text{acc}(\text{System 2}(D')))$

Store 1 if  $(\mu_2' - \Delta) - \mu_1' > \Delta$

p-value = the number of 1s / 10000

# Paired Bootstrap: Intuition

---

Performance on 10 examples:

System 1: 0 0 1 0 0 1 1 1 1 1

System 2: 0 0 0 0 1 0 0 1 1 1

Across many resamples, System 2 will usually do better than System 1, but p-value is only around 0.21 (0.05 is normal significance threshold)

P-value would be higher if the 1s were less correlated

P-value would be lower if the 1s were even more correlated, or if the dataset were bigger

# Interpreting p-values

---

- ▶ Bootstrap will give you a p-value based on the test statistic. If we do 10,000 trials of the bootstrap, the p-value will be  $N/10,000$  for some  $N$  (how many samples exhibited the unusual behavior, e.g., a difference greater than delta by chance)
- ▶ What does a  $p = 0.01$  mean?
  - ▶ Probability of obtaining results that are at least as extreme as the result observed, under the null hypothesis
  - ▶ It does NOT mean that the null hypothesis has a 1% chance of being true. (Determining this would require a Bayesian model, prior distribution over models/parameters, etc.)

# Student's t-test

---

- ▶ Test for the paired differences between observations (e.g.,  $\mu_2 - \mu_1$ )
- ▶ Assumes these are approximately normally distributed. This is a **parametric** test (unlike bootstrap which is **nonparametric**)
- ▶ You can do this, if:
  - ▶ You have a large enough test set (if test set is 50 examples, you can use the bootstrap but it's not large enough for the Central Limit Theorem to apply)
  - ▶ Your metric is well-behaved: some metrics like BLEU score are computed over the entire test set and don't assess examples independently like accuracy

# Other Tests

---

Wilcoxon Signed Rank Test: like the bootstrap, it's a *nonparametric* test

- ▶ Operates over the *rank* of differences and magnitudes of those differences are lost
- ▶ Differences in the power of the test based on what underlying distributions are, but overall similar to bootstrap

McNemar's test: looks at a binary matrix

{System 1, System 2} x {Correct count, Incorrect count}

and tests whether the distributions are the same

# Statistical Significance: Principles

---

- ▶ Significance helps establish that your test set is big enough and that your deltas didn't arise by chance.
- ▶ ...but it doesn't establish much else about robustness, including what happens if we:
  - ▶ Change distribution of test data
  - ▶ Change distribution of train data
  - ▶ Retrain our system on the same data
- ▶ You can find that a 1% gain on a single dataset is significant, but that model is worse on all other datasets

Administrative details and recap

What does accuracy mean?

Statistical Significance

Repeated Testing

Human (and LLM) Evaluation

Principles of Benchmarking

Benchmark Contamination

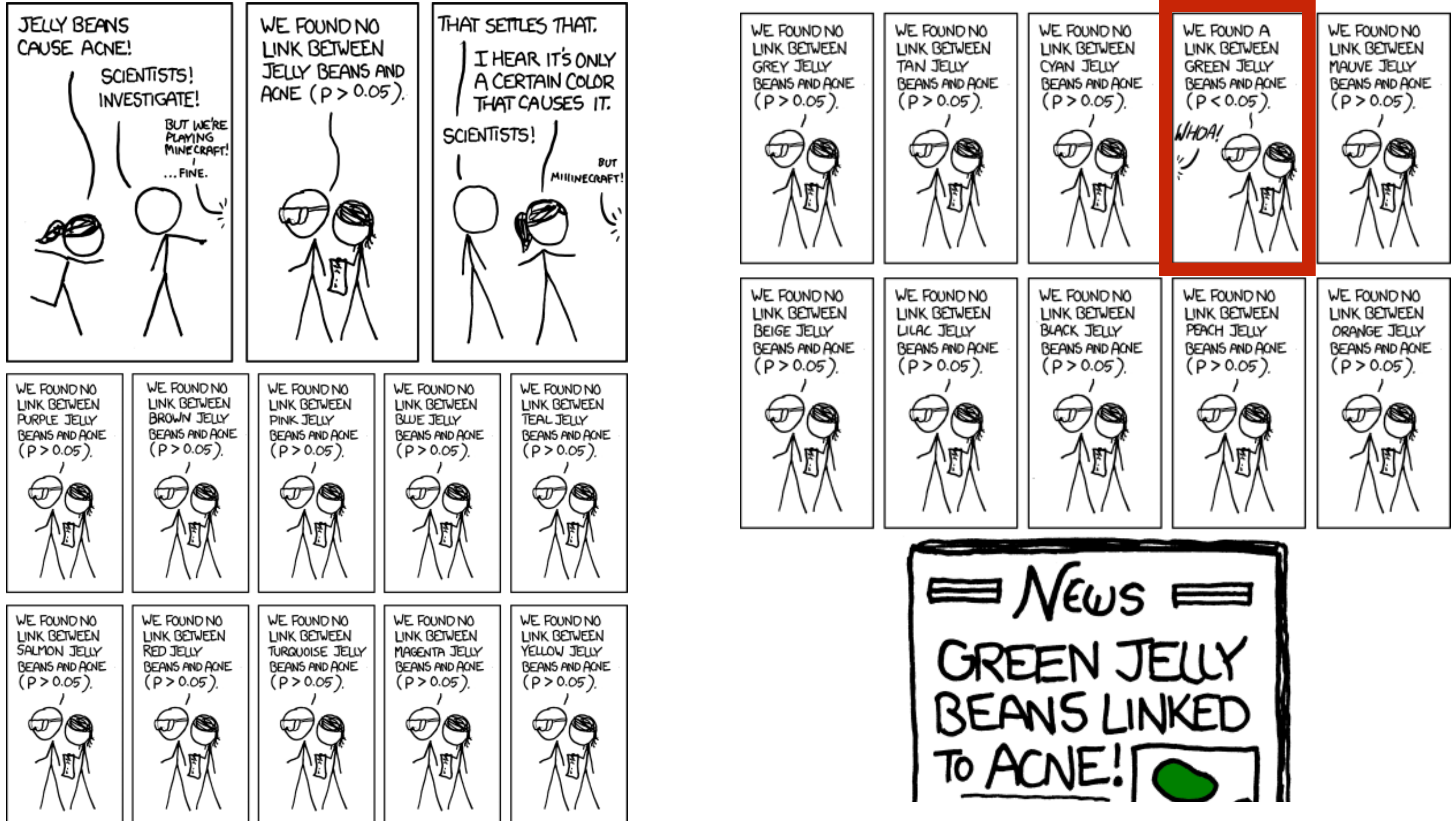
# Repeated Testing

# Multiple Testing

---

- ▶ Classic significance testing (e.g., a psychology experiment):  $p < 0.05$  is the threshold for significance
- ▶ What's wrong with the following: I try 20 hyperparameters, and one of them leads to a significant improvement with  $p > 0.0001$  (easily significant regardless of correction)

# Multiple Testing



# Bonferroni Correction

---

- ▶ Divide threshold by number of tests conducted
  - $p < 0.05$  becomes  $p < 0.0025$  when 20 tests are done
- ▶ Advantages and disadvantages of this?
- ▶ Suppose  $p < 0.0001$  even after this correction. Will this model necessarily be robust?

# File Drawer Effect

---

- ▶ Suppose each of the 20 Jelly Bean studies is done by a different researcher. 19 of those researchers will not submit their studies to be published, but the 20th will (and the 20th didn't need to do any Bonferroni because they only ran one test...)
- ▶ What are the ramifications when we think about this at the scale of an academic field?
- ▶ When doing research: be aware when you're working on a problem that seems unpublished, but which many people have likely thought about...there could be a reason it's not published

Administrative details and recap

What does accuracy mean?

Statistical Significance

Repeated Testing

Human (and LLM) Evaluation

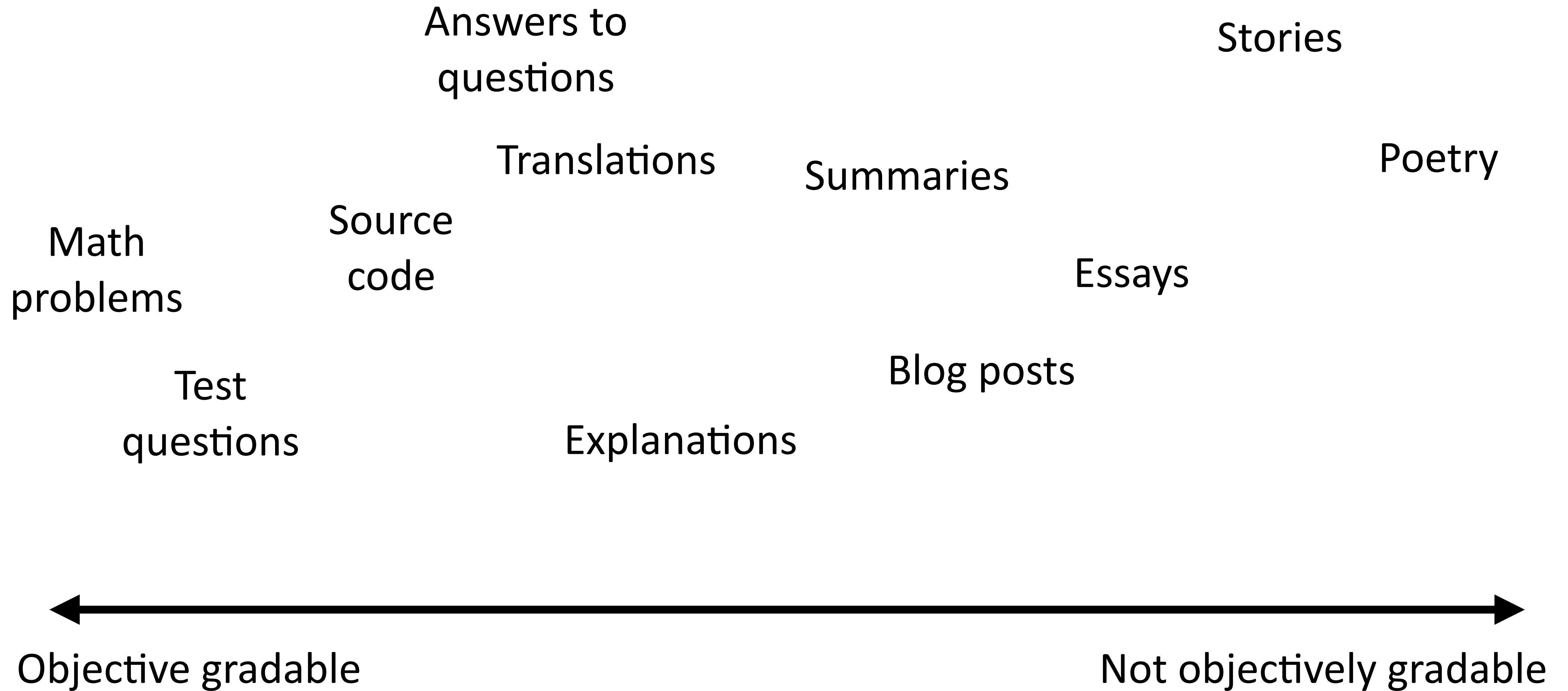
Principles of Benchmarking

Benchmark Contamination

# Human (and LLM) Evaluation

# Evaluation Beyond Accuracy

---



# Evaluation Beyond Accuracy

---

Translations, answers to questions, explanations, summaries

- ▶ We sometimes have a reference (e.g., a reference translation or a ground truth answer)
- ▶ Can look for features of the right answer (e.g., a piece of information in a summary, a keyword in an answer)

In general: some kind of word overlap can work, but is a bit suspect

# Evaluation Beyond Accuracy

---

Blog posts, essays, stories, poetry

- ▶ What features should these be graded on?

# Human (and LLM) Evaluation

---

Suppose we have a set of criteria: (1) grammaticality; (2) coherence; (3) relevance of content; (4) interestingness. How do we get ratings for these for our system outputs?

Prior to 2023: hire humans, give them instructions, have them judge.

Advantage: ultimate ground truth for how linguistic output will be understood

Disadvantage: humans might not do a careful job, could be inconsistent (for good reasons or bad reasons)

Post 2023: LLM-as-a-judge

Advantages: cheap, easy, pretty good

# LLM Evaluation

---

## Large Language Models Are State-of-the-Art Evaluators of Translation Quality

**Tom Kocmi** and **Christian Federmann**

Microsoft, One Microsoft Way, Redmond, WA-98052, USA

{tomkocmi, chrife}@microsoft.com

Score the following translation from {source\_lang} to {target\_lang} **with respect to the human reference** on a continuous scale from 0 to 100, where score of zero means "no meaning preserved" and score of one hundred means "perfect meaning and grammar".

```
{source_lang} source: "{source_seg}"  
{target_lang} human reference: {reference_seg}  
{target_lang} translation: "{target_seg}"  
Score:
```

# LLM Evaluation

**GPT-4**

**Small LMs (BERT) trained  
on labeled translations  
with human evals**

**Zero-shot BERT-based similarity metric**

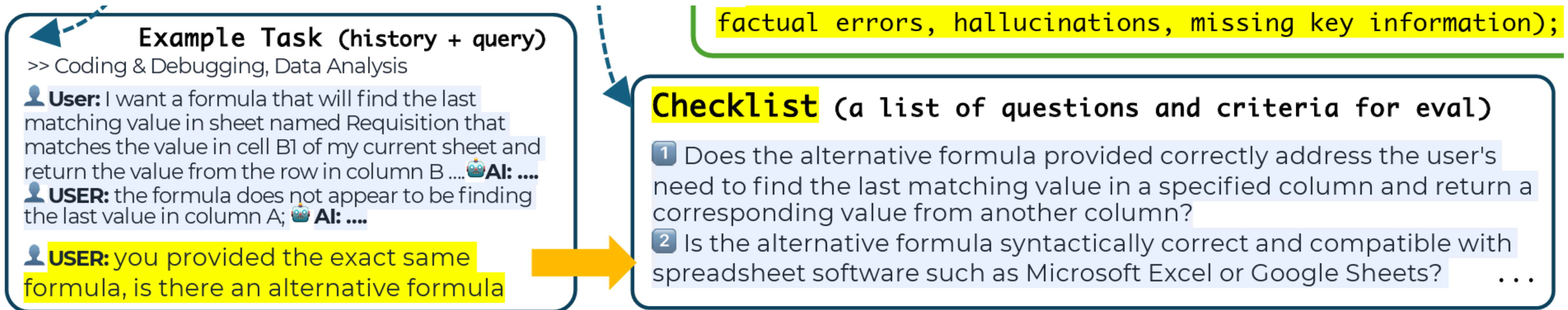
**Lexical matching metrics**

Metric	Accuracy
<b>GEMBA-GPT4-DA</b>	89.8%
<b>GEMBA-GPT4-DA[noref]</b>	87.6%
MetricX XXL	85.0%
BLEURT-20	84.7%
COMET-22	83.9%
COMET-20	83.6%
UniTE	82.8%
MS-COMET-22	82.8%
MATESE	81.0%
YiSi-1	79.2%
COMETKiwi[noref]	78.8%
COMET-QE[noref]	78.1%
BERTScore	77.4%
UniTE-src[noref]	75.9%
MS-COMET-QE-22[noref]	75.5%
MATESE-QE[noref]	74.8%
f200spBLEU	74.1%
chrF	73.4%
BLEU	70.8%

# Sources of Rubrics

For most tasks and criteria, we need **rubrics**: guides for how to grade responses that serve as instructions, for either humans or LLMs!

WildBench:



Criteria can come from: (1) decomposing the given instruction into criteria;  
(2) generation from an LLM

# EVALAGENT: Discovering Implicit Evaluation Criteria from the Web

Manya Wadhwa<sup>♠</sup>, Zayne Sprague<sup>♠</sup>, Chaitanya Malaviya<sup>◇</sup>, Philippe Laban<sup>♡</sup>  
Junyi Jessy Li<sup>♠</sup>, Greg Durrett<sup>♠</sup>

<sup>♠</sup>The University of Texas at Austin <sup>◇</sup>University of Pennsylvania <sup>♡</sup>Microsoft Research  
[manya.wadhwa@utexas.edu](mailto:manya.wadhwa@utexas.edu)





# How do we define evaluation criteria?

MW

*Write a New Yorker style fiction piece given the plot below: An observer becomes entranced by a seemingly ordinary couple on the street, follows them home, and watches them from outside in the rising floodwaters, drawing an eerie connection between the woman and a discard, burned chair they'd noticed earlier.*



*Title: The Couple on Delancey*

*On Delancey Street, just past the Polish bakery and across from the shuttered pawn shop with its cracked neon promise—Cash for Gold—the couple appeared. Not in any overt sense, no lightning bolt of entrance, no cinematic swell. They simply were, standing at the curb, inspecting the schedule on a bus stop as if the city were not pulsing and frayed around them.[...]*

*The observer—whose name we do not learn, not because it is unknown, but because they had long since stopped using it—watched from a distance of five sidewalk squares.*

*But the woman—her presence disturbed the air like heat off asphalt. Not beauty. Not elegance. Just... familiarity, though she was a stranger.*

*[...]*

*Upstairs, the woman raised her hand in acknowledgment. Not a wave. Not a beckon. Something else. A gesture like closing a door.*

*Then she turned from the window.*

*And the rain continued.*

- ▶ Too wordy
- ▶ Doesn't follow the "show don't tell" of writing
- ▶ Doesn't have plot twists

**(a) How do we automate?**  
**(b) Find reliable criteria?**



# How do we define evaluation criteria?

MW

*Write a New Yorker style fiction piece given the plot below: An observer becomes entranced by a seemingly ordinary couple on the street, follows them home, and watches them from outside in the rising floodwaters, drawing an eerie connection between the woman and a discard, burned chair they'd noticed earlier.*

Criteria derived from the instruction:

*The response should...*

- 1. Be a New Yorker style fiction piece*
- 2. Have details about the plot and characters*

**Evaluate the validity of the response.**

**Quality criteria are ambiguous**

Prompting an LLM for criteria:

*The response should...*

- 1. Have a clear depiction of the observer following the couple home*
- 2. Eerie connection drawn between the woman and a discarded chair*
- 3. Have a coherent structure and flow*
- 4. Have characters and theme developed in a way that aligns with literary fiction standards*



# What makes a good criterion?

*The response should have a coherent structure and flow*

*The response should be a New Yorker style fiction*

*The response should have physical details and environmental interactions that reveal character traits*

Specificity

*Precise dimension of quality*



Non-obviousness

*Implicit/unspoken principles*



Actionability

*Tangible improvements*



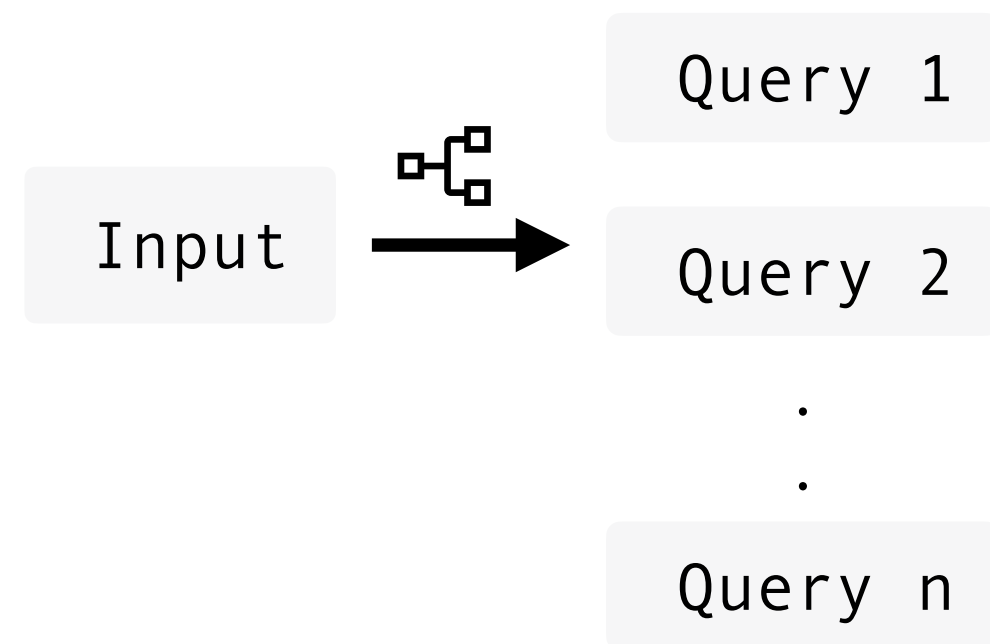
**EvalAgent generates specific, implicit and actionable criteria!**



# EvalAgent

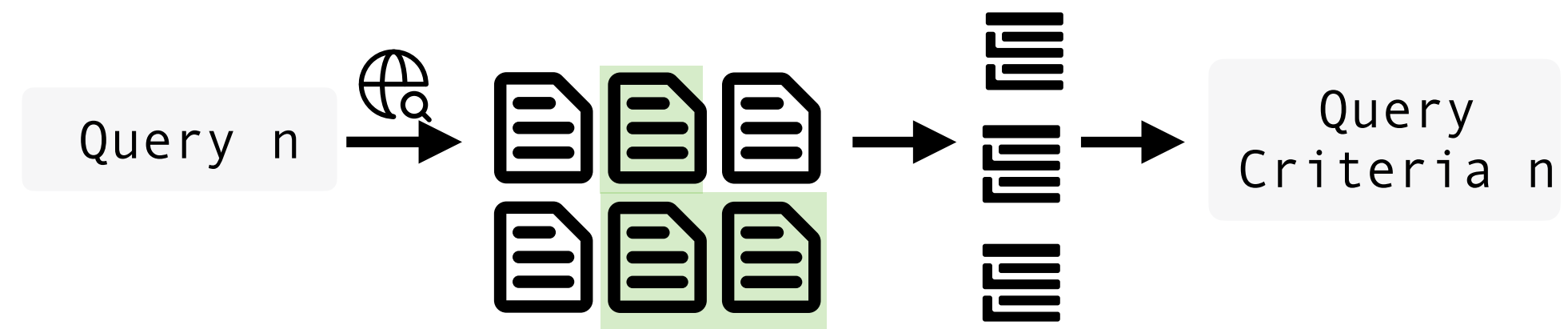
**EvalAgent generates implicit and task-specific criteria based on web documents**

## Step 1: Query Generator



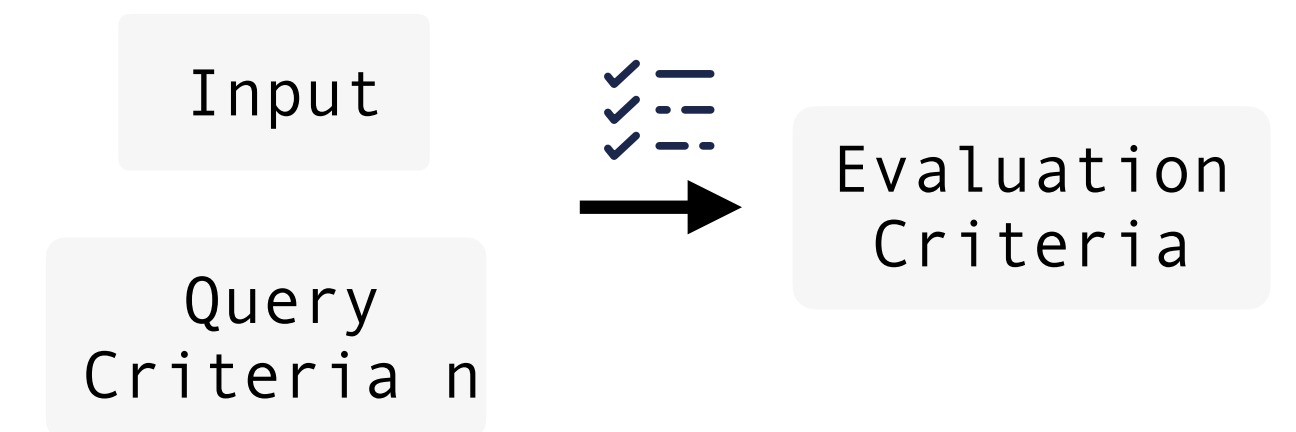
Queries capture *implicit* aspects of the task

## Step 2: Expert Retriever



Retrieve *instructional* documents, filter, summarize each document and aggregate to query specific checklist

## Step 3: Criteria Generator



Aggregates checklists into *task specific* criteria, filter and rank



# EvalAgent: Example

*Write a New Yorker style fiction piece given the plot below:[...]*

MW

## Step 1: Query Generator

Q1) how to write good fiction

Q2) how to build suspense

Q3) how to reveal character through observation

**Queries capture implicit aspects of the task!**



# EvalAgent: Example

Write a New Yorker style fiction piece given the plot below:[...]

MW

**Step 1:** Query Generator

**Step 2:** Expert Retriever

Q1) how to write good fiction

## 2. FICTION WRITING TIP: CREATING CHARACTERS

Think far back to 9th grade English, and you might remember the basic types of story conflicts: man vs. nature, man vs. man, and man vs. self. The conflicts that occur within stories happen to its characters—there can be no story without its people. Sometimes, your story needs to start there: in the middle of a conversation, a disrupted routine, or simply with what makes your characters special.

There are many ways to craft characters with depth and complexity. These include writing backstory, giving characters goals and fatal flaws, and making your characters contend with complicated themes and ideas. [This guide on character development](#) will help you sort out the traits your characters need, and how to interweave those traits into the story.

## 3. FICTION WRITING TIP: GIVE LIFE TO LIVING WORLDS

Whether your story is set on Earth or a land far, far away, your setting lives in the same way your characters do. In the same way that we read to get inside the heads of other people, we

also read to escape to a world outside of our own. Consider starting the story with what makes your world live: a pulsing city, the whispered susurrus of orchards, hills that roil with unsolved mysteries, etc. Tell us where the conflict is happening, and the story will follow.



# EvalAgent: Example

Write a New Yorker style fiction piece given the plot below:[...]

MW

## Step 1: Query Generator

## Step 2: Expert Retriever

### Q1) how to write good fiction

Eight Rules for Writing Fiction (<https://www.newyorker.com/humor/daily-shouts/eight-rules-for-writing-fiction>)

How to start writing fiction (<https://writers.com/how-to-start-writing-fiction>)

Create three-dimensional characters. Say you're writing about a hard-charging banker who's having an extramarital affair. This is a good start, but to avoid turning him into a flat character, you need to tell the reader why he's a hard-charger. What's in his chest? What drives him?

#### 3. FICTION WRITING TIP: GIVE LIFE TO LIVING WORLDS

Whether your story is set on Earth or a land far, far away, your setting lives in the same way your characters do. In the same way that we read to get inside the heads of other people, we also read to escape to a world outside of our own. Consider starting the story with what makes your world live: a pulsing city, the whispered susurrus of orchards, hills that roil with unsolved mysteries, etc. Tell us where the conflict is happening, and the story will follow.

### Q2) how to build suspense

How to write a suspense novel (<https://thewritepractice.com/how-to-write-a-suspense-novel/>)

Using Description To Reveal Character (<https://writers.com/feature/using-description-to-reveal-character>)

#### Use Dialogue to Reveal Character

What I want to talk about for this is the way dialogue reveals character without the overt "telling" that is so often maligned (and often necessary) in fiction.

#### 3. Varied pacing

When we both listen and watch characters infer and react to things, the reader infers as well.

You should have a good grasp on pacing, knowing when to ratchet up the tension and when to give readers a break. It's like putting your readers on a roller coaster — you need to use the brakes at times, and a slow uphill crawl at other times to increase **tension and build** anticipation.

Depending on what the author's motivations, and secrets.

Of course, feel free to include dizzying gut-wrenching twists and moments of glorious downhill glide, so your readers can feel the wind in their faces and their hearts in their throats.

**Retrieve and summarize high quality instructional documents to query specific criteria**



# EvalAgent: Example

Write a New Yorker style fiction piece given the plot below:[...]

MW

**Step 1: Query Generator**

**Step 2: Expert Retriever**

**Step 3: Criteria Generation**

Create three-dimensional characters. Say you're writing about a hard-charging banker who's having an extramarital affair. This is a good start, but to avoid turning

him i  
tell th  
a har  
chest  
drives

3. FICTION WRITING TIP: GIVE LIFE TO LIVING WORLDS

Whether your story  
your characters do.  
also read to escape  
your world live: a p  
mysteries, etc. Tell

What I want to talk about fo  
that is so often maligned (a

When we both listen *and* w  
characters infer and react t  
things), the reader infers as

Depending on what the aut  
motivations, and secrets.

*Use Dialogue to Reveal Character*

3. Varied pacing

You should have a good grasp on pacing, knowing when to ratchet up the tension and when to give readers a break. It's like putting your readers on a roller coaster — you need to use the brakes at times, and a slow uphill crawl at other times to increase **tension and build** anticipation.

Of course, feel free to include dizzying gut-wrenching twists and moments of glorious downhill glide, so your readers can feel the wind in their faces and their hearts in their throats.

- The characters should have goals
- Character actions should have consequences
- Story should maintain a narrative voice and tone complementary to the character
- Story should reveal character observations through actions, dialogues and environmental interactions
- [...]
- [...]

**Aggregate and align criteria with the user prompt to generate evaluation criteria**

Administrative details and recap

What does accuracy mean?

Statistical Significance

Repeated Testing

Human (and LLM) Evaluation

Principles of Benchmarking

Benchmark Contamination

# Principles of Benchmarking

# Desiderata for Benchmarks I

---

- ▶ Dataset must have enough examples (exception: Terminal-Bench 2.0, with <100)
- ▶ High-quality: answers should be right, grading should be correct
- ▶ Simple to understand
- ▶ Easy to run
- ▶ Needs to be a meaningful task
- ▶ (Ideally) doesn't saturate too quickly

<https://www.jasonwei.net/blog/evals>

# Desiderata for Benchmarks II

---

- ▶ Natural
- ▶ Automatically evaluatable (+ single metric)
- ▶ Challenging

# Modern Benchmarking: Muse Spark

	Benchmark	Muse Spark Thinking	Opus 4.6 Max	Gemini 3.1 Pro High	GPT 5.4 Xhigh	Grok 4.2 Reasoning
MULTIMODAL	CharXiv Reasoning Figure Understanding	86.4	65.3 Self-Reported: 61.5	80.2	82.8	60.9
	MMMU Pro Multimodal Understanding	80.4	77.4	83.9	81.2	75.2
	ERQA Embodied Reasoning	64.7	51.6	69.4	65.4	54.1
	SimpleVQA Visual Factuality	71.3	62.2	72.4	61.1	57.4
	ScreenSpot Pro Screenshot Localization - With Python	84.1	83.1	84.4	85.4	—
	ZeroBench Multi-Step Visual Reasoning (pass@5) - With Python	33.0	—	29.0	41.0	—
TEXT/REASONING	Humanity's Last Exam Multidisciplinary Reasoning (No Tools)	42.8	40.0	45.4 Self-Reported: 44.4	43.9 Self-Reported: 39.8	31.6
	Humanity's Last Exam Multidisciplinary Reasoning (With Tools)	50.4	53.1	51.4	52.1	—
	ARC AGI 2 Abstract Reasoning Puzzles (Public)	42.5	63.3	76.5	76.1	53.3
	GPQA Diamond PhD Level Reasoning	89.5	92.7 Self-Reported: 91.3	94.3	92.8	88.5
	LiveCodeBench Pro Competitive Coding	80.0	70.7	82.9 Self-Reported: 78.2	87.5	74.2

# Modern Benchmarking: Muse Spark

HEALTH	HealthBench Hard Open-Ended Health Queries	42.8	14.8	20.6	40.1	20.3
	MedXpertQA (Text) Medical Multiple Choice	52.6	52.1	71.5	59.6	50.2
	MedXpertQA (MM) Medical Multiple Choice	78.4	64.8	81.3	77.1	65.8
AGENTIC	DeepSearchQA Agentic Search	74.8	73.7	69.7	73.6	62.8
	SWE-Bench Verified Agentic Coding	77.4	80.8	80.6	—	76.7*
	SWE-Bench Pro Diverse Agentic Coding	52.4	53.4	54.2	57.7	51.8*
	Terminal-Bench 2.0 Agentic Terminal Coding	59.0	65.4	68.5	75.1	47.1*
	$\tau^2$ -Bench Telecom Agentic Tool Use (Artificial Analysis)	91.5	92.1	95.6	91.5	96.5
	GDPval-AA Elo Office Tasks (Artificial Analysis)	1444	1606	1320	1672	1055

# Exception: ChatbotArena

What's the longest palindrome you can come up with?

Assistant A



Creating a very long, coherent palindrome on the fly is quite a challenge! The best ones are often crafted over time, leveraging clever wordplay and structural symmetry.

One of the longest and most famous English palindromes that actually makes a bit of sense is:

**"Doc, note: I dissent. A fast never prevents a fatness. I diet on cod."**

Let's break it down (ignoring punctuation and spaces):

```
DOCNOTEIDISSENTAFASTNEVERPREVENTSAFATNESSI  
DIETONCOD  
DOCNOTEIDISSENTAFASTNEVERPREVENTSAFATNESSI  
DIETONCOD (read backward)
```

It's **54 letters long** (excluding punctuation and spaces).

While I could try to string together random words, this one demonstrates the



Assistant B



Technically, the longest palindrome I can generate is **infinite**, provided I just repeat the same letter forever (e.g., `aaaaaaaaaaaa...`). However, if you mean a *meaningful* sentence or paragraph, the answer becomes much harder.

There is no single "official" world record for the longest English palindrome because they depend on how strictly you apply rules (case sensitivity, spaces, punctuation). However, here are three categories of palindromes I can provide, ranging from famous classics to a custom-built long version.

## 1. The Famous Classic (Meaningful & Compact)

*"A man, a plan, a canal: Panama"*  
(22 characters)

This is arguably the most famous palindrome in history because it forms a complete, grammatical sentence that tells a mini-story.



2. The "Structural" Long One (Custom-Built)

# Elo

---




- ▶ System of ranking based on wins + losses, used in chess and other competitive matchmaking environments

- ▶ After a “game”:

$$E_A = \frac{1}{1 + 10^{(R_B - R_A)/400}}$$

$$R'_A = R_A + K(S_A - E_A)$$

- ▶ ChatbotArena just fits a Bradley-Terry model

Rank	Model	Elo Rating	Description
1	 vicuna-13b	1169	a chat assistant fine-tuned from LLaMA on user-shared conversations by LMSYS
2	 koala-13b	1082	a dialogue model for academic research by BAIR
3	 oasst-pythia-12b	1065	an Open Assistant for everyone by LAION
4	alpaca-13b	1008	a model fine-tuned from LLaMA on instruction-following demonstrations by Stanford
5	chatglm-6b	985	an open bilingual dialogue language model by Tsinghua University
6	fastchat-t5-3b	951	a chat assistant fine-tuned from FLAN-T5 by LMSYS
7	dolly-v2-12b	944	an instruction-tuned open large language model by Databricks
8	llama-13b	932	open and efficient foundation language models by Meta
9	stablelm-tuned-alpha-7b	858	Stability AI language models

Administrative details and recap

What does accuracy mean?

Statistical Significance

Repeated Testing

Human (and LLM) Evaluation

Principles of Benchmarking

Benchmark Contamination

# Benchmark Contamination

# Kinds of Contamination

---

How could a model be contaminated for a test dataset?

- ▶ Pre-training data contains the examples
- ▶ Post-training data contains examples with labels
- ▶ RL dataset contains the data points

# Detecting Contamination

## Pre-training Data

The music was composed by Hitoshi Sakimoto, who had also worked on the previous Valkyria Chronicles games...

...

Does a frog jump out of boiling water?

Is it possible to create mass from energy?

Is there a movie with  $\theta$  on rotten tomatoes?

Is the jaguar S type rear wheel drive?

...

Highway89 was created out of a highway rerouting in the late 1930s. Originally, it formed the routing...

Test Set  
Contamination

## Contamination Test

### Canonical Order

Does a frog jump out of boiling water?



Is it possible to create mass from energy? ✓



Is there a movie with  $\theta$  on rotten tomatoes? ✓



Is the jaguar S type rear wheel drive? ✓



high model log-probability



low model log-probability

### Shuffled Order

Does a frog jump out of boiling water?



Is it possible to create mass from energy? ✓



Is the jaguar S type rear wheel drive? ✗



Is there a movie with  $\theta$  on rotten tomatoes? ✗

Differences in log-probability between orderings reveal contamination.

- ▶ Likelihood of examples in dataset order should be substantially higher than examples in a random order!

---

**Algorithm 1** Sharded Rank Comparison Test

---

**Require:** Test set examples  $x_1, \dots, x_n$

**Require:** Target model  $\theta$

**Require:** Number of shards  $r$

**Require:** Number of permutations per shard  $m$

1: Partition the examples into shards  $S_1, S_2, \dots, S_r$ , where each shard has at least  $\lfloor n/r \rfloor$  examples, and one extra example is added to the first  $n \bmod r$  shards.

2: **for** each shard  $S_i$  **do**

3:     Compute the log-likelihood of the canonical order:

$$l_{\text{canonical}}^{(i)} := \log p_{\theta}(\text{seq}(x_1^{(i)}, x_2^{(i)}, \dots, x_k^{(i)}))$$

4:     Estimate  $l_{\text{shuffled}}^{(i)} := \text{Mean}_{\pi}[\log p_{\theta}(\text{seq}(x_{\pi(1)}^{(i)}, \dots, x_{\pi(k)}^{(i)}))]$  by computing the sample average over  $m$  random permutations  $\pi$ .

5:     Compute  $s_i = l_{\text{canonical}}^{(i)} - l_{\text{shuffled}}^{(i)}$

6: **end for**

7: Define  $s = \frac{1}{r} \sum_{i=1}^r s_i$  the sample average over the shards.

8: Run a one-sided t-test for  $E[s_i] > 0$ , returning the associated p-value of the test as  $p$ .

Table 1: We report the results of training a 1.4B language model from scratch on Wikitext with intentional contamination. For each injected dataset, we report the number of examples used (size), how often the test set was injected into the pre-training data (dup count), and the p-value from the permutation test and sharded likelihood comparison test. The bolded p-values are below 0.05 and demonstrate in the case of higher duplication counts, such as datasets appearing 10 or more times, we obtain vanishingly small p-values on our test. Finally, rows marked 1e-38 were returned as numerically zero due to the precision of our floating point computation.

Name	Size	Dup Count	Permutation p	Sharded p
BoolQ	1000	1	0.099	0.156
HellaSwag	1000	1	0.485	0.478
OpenbookQA	500	1	0.544	0.462
MNLI	1000	10	<b>0.009</b>	<b>1.96e-11</b>
TruthfulQA	1000	10	<b>0.009</b>	<b>3.43e-13</b>
Natural Questions	1000	10	<b>0.009</b>	<b>1e-38</b>
PIQA	1000	50	<b>0.009</b>	<b>1e-38</b>
MMLU Pro. Psychology	611	50	<b>0.009</b>	<b>1e-38</b>
MMLU Pro. Law	1533	50	<b>0.009</b>	<b>1e-38</b>
MMLU H.S. Psychology	544	100	<b>0.009</b>	<b>1e-38</b>

# Kinds of Contamination

---

- ▶ What kind of contamination does this not detect?

Administrative details and recap

What does accuracy mean?

Statistical Significance

Repeated Testing

Human (and LLM) Evaluation

Principles of Benchmarking

Benchmark Contamination