

Duke CS590-06 Building Intelligent Agents with Frontier Models

Lecture 4: Benchmarking and Evaluation

Shuyan Zhou

9/4/2025

Duke CS590-06

Part of this lecture drew inspiration from and adapted material from
Stanford's CS224n (Spring 2025) Lecture 11

Why evaluate? to compare, measure, understand, and drive progress

- **Compare models** for a specific task.
- **Measure progress** in the field over time.
- Understand a model's **behavior and limitations**.
- **Drive improvements** via incentives (benchmarks).

Objective metrics drive better model selection

Scenario: You need to translate user reviews from Spanish to English.

You have two models: **Model A** and **Model B**.

How do you decide? 🤔

Method: Evaluate both on a test set using a quality metric like BLEU score.

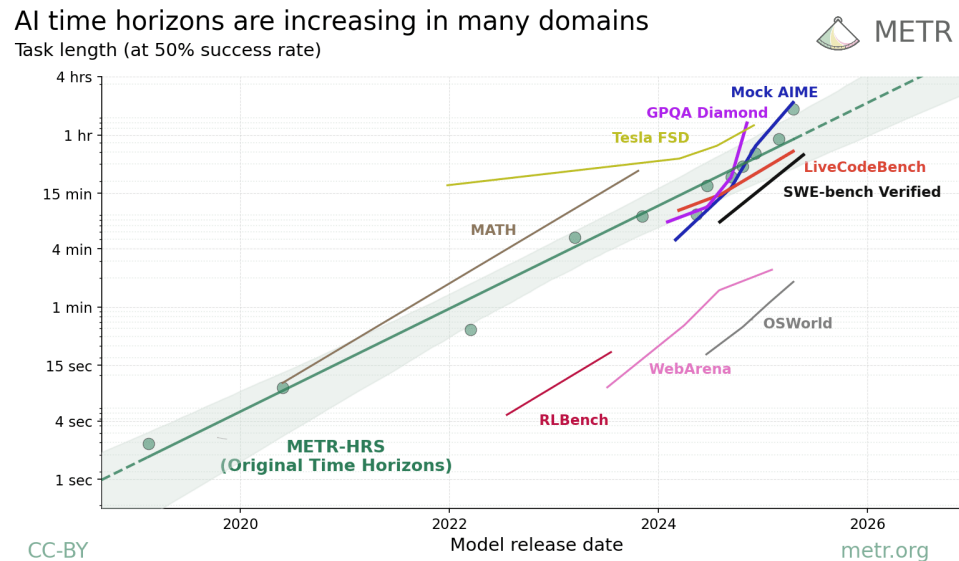
Model A: BLEU score = 35

Model B: BLEU score = 30

Decision: Deploy Model A. ✅

Benchmarks reveal exponential progress in AI

Benchmarks allow us to map AI performance to human labor costs and track progress over time.



How Does Time Horizon Vary Across Domains?

Evaluation reveals weaknesses and guides research

- Evaluating models on **low-resource languages** often reveals their limitations.
- Models excelling at English often perform poorly on languages like Swahili or Nepali due to a lack of training data.
- This poor benchmark performance highlights critical problems and **drives research** into new areas.

Public benchmarks create incentives and accelerate innovation

Public leaderboards create a clear target for the research community.

- The ImageNet Challenge spurred a revolution in computer vision (e.g., AlexNet in computer vision)
- Ambitious benchmarks like ARC-AGI or the "Human's Last Exam" push for fundamentally new approaches.

A well-defined goal focuses community effort and accelerates innovation.

Effective evaluation requires a clear task and metric

All evaluation boils down to two components:

1. **Task Definition:** What is the *exact* task? (Inputs & Outputs)
2. **Metrics:** How do you measure success? (The scoring rule)

NLP evaluation is shifting from close-ended to open-ended tasks

Historically: **Close-ended tasks**

Now: **Open-ended tasks**

These two types of tasks require different evaluation design and methodologies.

Close-ended tasks

Characteristics:

- A limited, predictable set of possible answers.
- Often, only a few correct answers exist.
- Evaluation can be easily and reliably automated.

Classification: a common close-ended task

- **Sentiment Classification:** Is this movie review positive or negative?
- **Natural Language Inference (NLI):** Does a hypothesis contradict, entail, or is it neutral to the premise?

"I love this movie.
I've seen it many times
and it's still awesome."



"This movie is bad.
I don't like it at all.
It's terrible."



Premise	Label	Hypothesis
A man inspects the uniform of a figure in some East Asian country.	contradiction	The man is sleeping.
An older and younger man smiling.	neutral	Two men are smiling and laughing at the cats playing on the floor.
A soccer game with multiple males playing.	entailment	Some men are playing a sport.

Extractive QA: evaluation by finding the exact text span

In SQuAD (Rajpurkar et al, 2016), the model answers by extracting a text span from a passage.

Passage: "Gemini is a family of multimodal models developed by Google..."

Question: "Who developed Gemini?"

Correct Answer: "Google"

Metrics: **Exact Match (EM)** and **F1-Score** (word overlap).

Multi-task benchmarks (SuperGLUE) push for generalization

Measures general capabilities across a diverse set of close-ended tasks.

BoolQ	Passage: <i>Barq's – Barq's is an American soft drink. Its brand of root beer is notable for having caffeine. Barq's, created by Edward Barq and bottled since the turn of the 20th century, is owned by the Barq family but bottled by the Coca-Cola Company. It was known as Barq's Famous Olde Tyme Root Beer until 2012.</i> Question: <i>is barq's root beer a pepsi product</i> Answer: No
CB	Text: <i>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?</i> Hypothesis: <i>they are setting a trend</i> Entailment: Unknown
COPA	Premise: <i>My body cast a shadow over the grass.</i> Question: <i>What's the CAUSE for this?</i> Alternative 1: <i>The sun was rising.</i> Alternative 2: <i>The grass was cut.</i> Correct Alternative: 1

The final score is an **aggregated metric** (average score) across all tasks.

MMLU raises the bar: testing for expert-level knowledge

The Massive Multitask Language Understanding (MMLU) benchmark is a much harder challenge.

- 57 tasks in a multiple-choice format.
- Subjects like US history, law, computer science, and physics.
- Requires significant world knowledge, not just pattern matching.

Microeconomics	One of the reasons that the government discourages and regulates monopolies is that	
	(A) producer surplus is lost and consumer surplus is gained.	✗
	(B) monopoly prices ensure productive efficiency but cost society allocative efficiency.	✗
	(C) monopoly firms do not engage in significant research and development.	✗
	(D) consumer surplus is lost with higher prices and lower levels of output.	✓

Figure 3: Examples from the Microeconomics task.

Conceptual Physics	When you drop a ball from rest it accelerates downward at 9.8 m/s^2 . If you instead throw it downward assuming no air resistance its acceleration immediately after leaving your hand is	
	(A) 9.8 m/s^2	✓
	(B) more than 9.8 m/s^2	✗
	(C) less than 9.8 m/s^2	✗
	(D) Cannot say unless the speed of throw is given.	✗
College Mathematics	In the complex z -plane, the set of points satisfying the equation $z^2 = z ^2$ is a	
	(A) pair of points	✗
	(B) circle	✗
	(C) half-line	✗
	(D) line	✓

Figure 4: Examples from the Conceptual Physics and College Mathematics STEM tasks.

Challenges: choosing the right metrics

Scenario: A fraud detector, where only 0.1% of transactions are fraudulent.

- **Model A (Useless):** Always predicts "not fraud".
 - **Accuracy: 99.9%, Recall: 0%**
- **Model B (Useful):** Catches 80% of fraud, some false positives.
 - **Accuracy: 99.8%, Recall: 80%**

Model A has higher accuracy but is worthless. We need better metrics like Precision, Recall, and F1-Score.

Challenges: Models can cheat by exploiting spurious correlations

Models might learn superficial patterns (shortcuts) instead of true understanding.

Premise	A woman selling bamboo sticks talking to two men on a loading dock.
Entailment	There are at least three people on a loading dock.
Neutral	A woman is selling bamboo sticks to help provide for her family .
Contradiction	A woman is not taking money for any of her sticks.

Model achieves non-trivial performance with hypothesis only (no premise presented!) (Gururangan et al 2017)

Practice: Test against simple baselines (e.g., a model that only sees the hypothesis).

Open-ended tasks

Characteristics:

- A vast, essentially infinite number of possible correct answers.
- Multi-facet evaluation: some correct answers are better than others.
- Evaluation is much harder.

Modern NLP is defined by open-ended generation

- **Summarization:** Condense a long article into a few sentences.
- **Machine Translation:** Translate a sentence from one language to another.
- **Code Generation:** Write a function based on a natural language description.
- **Dialogue:** Have a coherent, engaging conversation.

Evaluating open-ended tasks: three main approaches

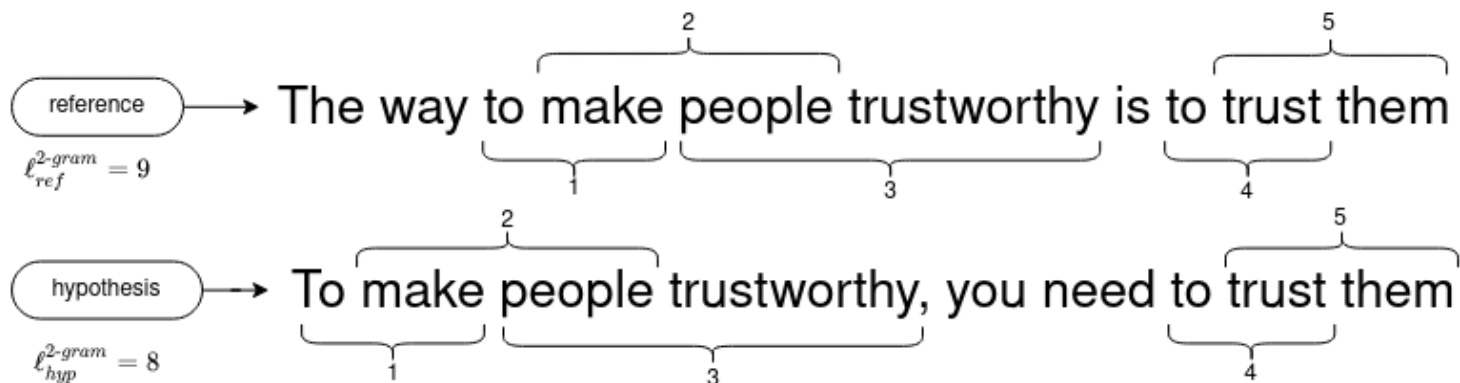
1. **Reference-Based:** Compare model output to a "golden" reference answer.
2. **Reference-Free:**
 1. Judge the output on its own intrinsic qualities.
 2. Comparing model outputs
 3. Outcome-Based: Check if the output achieves a real-world goal.

Reference-based metrics: simple n-gram overlap (BLEU, ROUGE)

(Prompt, Reference answer, generation)

Measure **word or phrase** overlap between the generated text and a reference text.

- BLEU (Papineni et al, 2002): For translation
- ROUGE (Lin et al, 2004): for summarization



N-gram metrics fail on semantics and paraphrasing

Metrics like BLEU and ROUGE are easily fooled because they don't understand semantic

Reference: "Global leaders convened to discuss climate change, focusing on carbon emission reduction strategies."

Generated A (Good): "World heads of state met to talk about global warming and how to lower carbon output."

(Low word overlap → low ROUGE score)

Generated B (Bad): "Global leaders discuss carbon emission reduction."

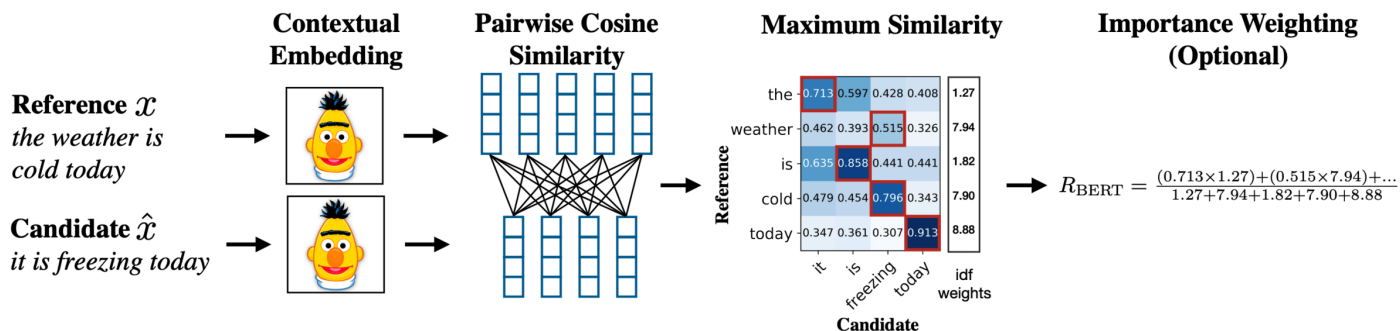
(High word overlap → high ROUGE score)

Model-based metrics capture semantic similarity

Idea: Use a model to compute **semantic similarity** between the generated and reference texts.

BERTScore (Zhang et al 2017):

1. Embed tokens from both texts using a pretrained model like BERT.
2. Compute cosine similarity between token embeddings.
3. Aggregate the similarity scores.

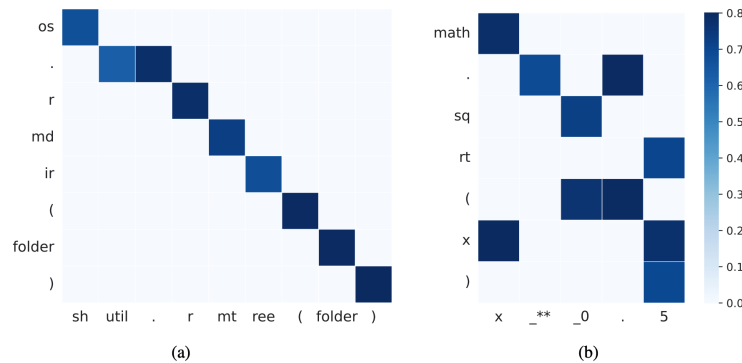
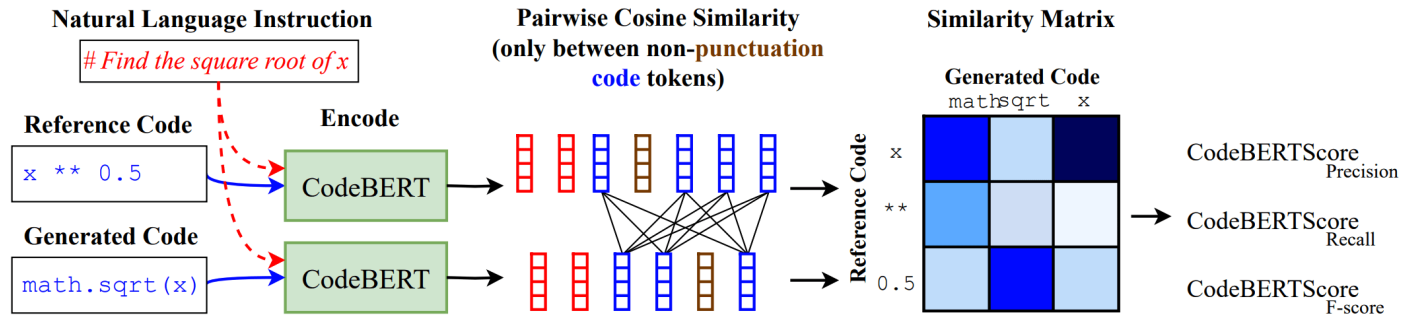


Model-based metrics has higher correlation with human judgements

Metric	en↔cs (5/5)	en↔de (16/16)	en↔et (14/14)	en↔fi (9/12)	en↔ru (8/9)	en↔tr (5/8)	en↔zh (14/14)
BLEU	.970/. 995	.971/. 981	.986/.975	.973/. 962	.979/. 983	.657 /.826	.978/.947
ITER	.975/.915	.990/. 984	.975/. 981	.996/.973	.937/.975	.861 /.865	.980/ –
RUSE	.981/ –	.997/ –	.990/ –	.991/ –	.988/ –	.853/ –	.981/ –
YiSi-1	.950/. 987	.992/. 985	.979/. 979	.973/.940	.991/.992	.958/.976	.951/. 963
P_{BERT}	.980/. 994	.998/.988	.990/.981	.995/.957	.982/. 990	.791/.935	.981/.954
R_{BERT}	.998/.997	.997/. 990	.986/. 980	.997/.980	.995/.989	.054/.879	.990/.976
F_{BERT}	.990/.997	.999/.989	.990/. 982	.998/.972	.990/.990	.499/.908	.988/.967
F_{BERT} (idf)	.985/. 995	.999/.990	.992/.981	.992/. 972	.991/.991	.826/.941	.989/.973

(Zhang et al 2017)

Model-based metrics beyond natural language



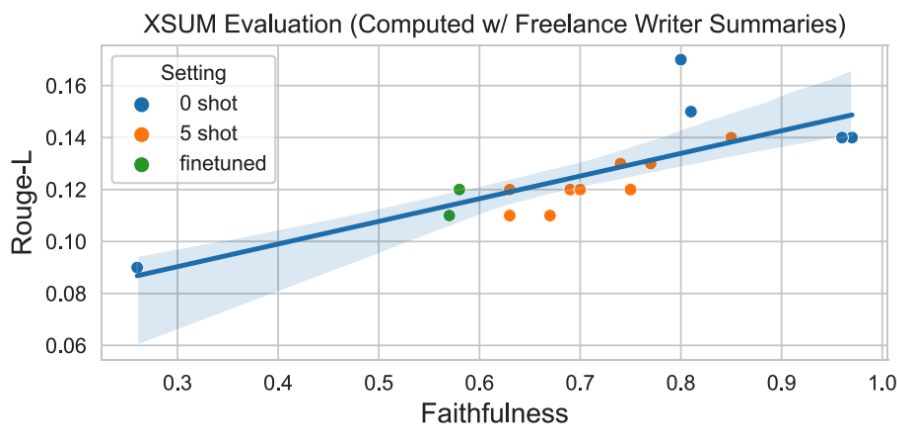
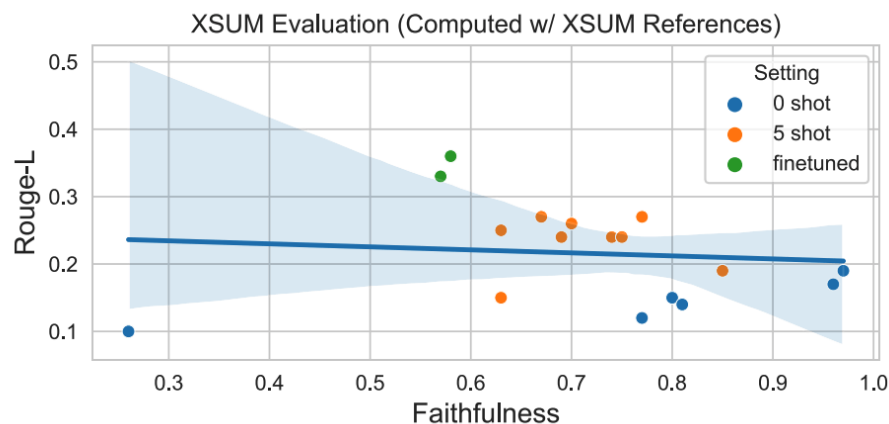
CodeBertScore (Zhou et al, 2022)

Semantic metrics can miss critical nuances

- Even powerful metrics can fail to capture small but critical semantic differences.
- e.g., In code generation if $x \geq 0$ vs if $x > 0$
 - A metric like CodeBERTScore might see these as 99% similar, but one is correct and the other may miss edge conditions.

Reference-based metrics penalize novel, correct answers

A fundamental flaw: they assume the provided reference is the *only* way to be correct.



(Zhang et al 2024)

- Human ratings don't correlated with the reference in the original dataset
- Correlate with the references provided by experts

Moving beyond references: judging output directly

- **Problem:** Relying on limited "golden" references is too limiting.
- **Solution:** Evaluate the generated output on its own merits.
- **Two main paths:**
 - Rate its intrinsic qualities (fluency, helpfulness).
 - Check its extrinsic effect (does it complete a task?).

The gold standard: human evaluation

Ask human annotators to rate model outputs on several axes:

- Overall Quality
- Fluency & Coherence
- Relevance & Helpfulness
- Factual Correctness

All automatic metrics are ultimately trying to be a cheap proxy for human judgment.

Human evaluation can be flawed

While it is the ground truth, human evaluation is:

- **Slow:** Takes days or weeks to collect.
- **Expensive:** You have to pay annotators. 💰
- **Hard to Standardize:** Instructions, interfaces, and annotator pools vary.

This has led to a "reproducibility crisis" where it's hard to compare human evaluation results across papers.

Reproducibility crisis in human evaluations

Non-Repeatable Experiments and Non-Reproducible Results: The Reproducibility Crisis in Human Evaluation in NLP

Anya Belz^{a,b}

Craig Thomson^b

Ehud Reiter^b

Simon Mille^a

^aADAPT, Dublin City University
Dublin, Ireland

^bUniversity of Aberdeen
Aberdeen, UK

{anya.belz,simon.mille}@adaptcentre.ie

{c.thomson,e.reiter}@abdn.ac.uk

- Are the experiments repeatable?
- Are the results reproducible?

Only 5% experiments are repeatable, 20% when authors helped

- Practical barrier
- Lack of information

Reference-free automation: using models to score generations

Can we use a model to approximate human judgment *without a reference*?

BARTScore:

- **Key Idea:** A good generation is one that a powerful model thinks is *likely* given the source input.

$$\text{BARTSCORE} = \sum_{t=1}^m \omega_t \log p(\mathbf{y}_t | \mathbf{y}_{<t}, \mathbf{x}, \theta)$$

(Yuan et al 2021)

BartScore correlates well with human judgement


	Rank19	Q-CNN	Q-XSUM
	Acc.	Pearson	
ROUGE-1	0.568	0.338	-0.008
ROUGE-2	0.630	0.459	0.097
ROUGE-L	0.587	0.357	0.024
BERTScore	0.713	0.576	0.024
MoverScore	0.713	0.414	0.054
PRISM	0.780	0.479	0.025
FactCC [30]	0.700	–	–
QAGS [67]	0.721	0.545	0.175
Human [14]	0.839	–	–
BARTSCORE	0.684	0.661†	0.009
+ CNN	0.836†	0.735†	0.184†
+ CNN + Para	0.788	0.680†	0.074
+ CNN + Prompt	0.796	0.719†	0.094

(Yuan et al 2021)

Finetuning on related corpus helps the shape the probability distributions

Put models in wild comparisons

- Top models have very close scores on standard benchmarks, making them hard to distinguish.
- These benchmarks often don't capture the full range of real-world use cases.



Center for Research on Foundation Models HELM Capabilities Leaderboard

Leaderboard: MMLU-Pro

MMLU-Pro

Model	COT correct
Claude 4 Opus (20250514, extended thinking)	0.875
Gemini 2.5 Pro (03-25 preview)	0.863
GPT-5 (2025-08-07)	0.863
Claude 4 Opus (20250514)	0.859

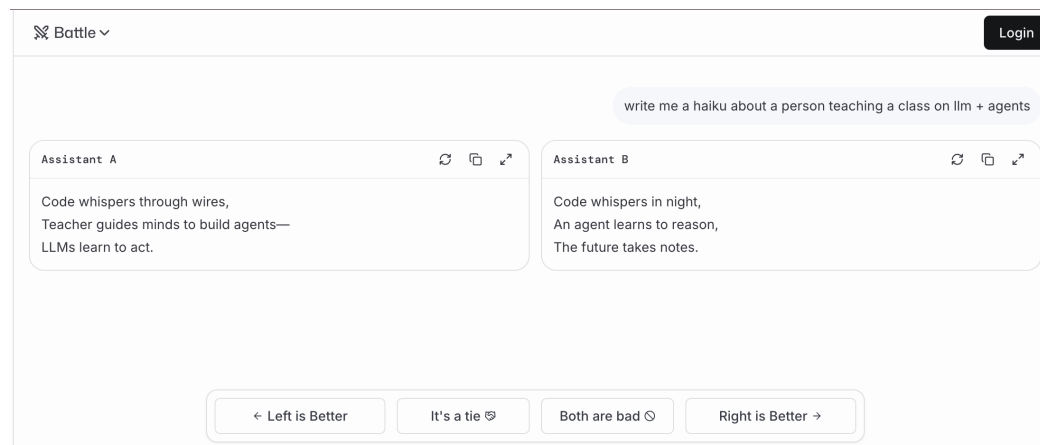
- A better question: not "Is this good?" but **"Is A better than B?"**

Crowdsourcing preference: the chatbot arena

Platforms like the **Chatbot Arena** use side-by-side comparison:

1. A user enters a prompt.
2. Two anonymous models generate responses.
3. The user votes for which response is better (A, B, Tie, Bad).

These pairwise comparisons are used to calculate an **Elo rating** for each model, creating a dynamic, public leaderboard.



The screenshot displays the Chatbot Arena interface. At the top, there is a "Battle" dropdown menu and a "Login" button. The main area shows a prompt: "write me a haiku about a person teaching a class on llm + agents". Below the prompt, two assistants are compared side-by-side. Assistant A's response is: "Code whispers through wires, Teacher guides minds to build agents— LLMs learn to act." Assistant B's response is: "Code whispers in night, An agent learns to reason, The future takes notes." At the bottom, there are four buttons for voting: "Left is Better", "It's a tie", "Both are bad", and "Right is Better".

The "arena" paradigm extends beyond chatbots

Side-by-side, preference-based evaluation is now being used for more specialized tasks.

- **WebDev Arena:** Judge which model writes better HTML/CSS code.

The screenshot shows a web application titled "Podcast Transcript Generator". It has a form for creating a transcript. The form includes fields for "Episode title", "Host name" (with "Alex Ruiz" entered), and "Guest names" (with "Jamie Lee, Morgan Patel" entered). There are also checkboxes for "Future of Audio", "Storytelling in podcasts", "Monetization strategies", and "Audience engagement". A "Duration (minutes)" field is set to "45", and a "Tone" dropdown is set to "Informative". A "Generate transcript" button is at the bottom right. The interface is clean and modern, with a light blue and white color scheme.

VISION				text-to-image			
Rank (UB)	Model	Score	Votes	Rank (UB)	Model	Score	Votes
1	gemi-2.5-pro	1248	22,173	1	gemi-2.5-flash-image-previ...	1147	220,674
1	chatgpt-4o-latest-20250326	1235	12,519	2	imagen-4.0-ultra-generate-pr...	1135	193,895
2	gpt-5-chat	1224	10,019	3	gpt-image-1	1129	128,710
2	gpt-4.5-preview-2025-02-27	1220	2,946	4	imagen-4.0-generate-preview...	1119	196,696
2	o3-2025-04-16	1219	18,834	5	qwen-image-prompt-extend	1082	123,596
3	gemi-2.5-flash	1208	15,401	5	seedream-3	1077	159,028
3	gpt-4.1-2025-04-14	1206	14,404	6	flux-1-kontext-max	1075	78,017
3	gpt-5-high	1205	12,032	8	imagen-3.0-generate-002	1062	256,225
3	claude-opus-4-20250514-think...	1200	1,417	9	flux-1-kontext-pro	1056	165,377
3	claude-sonnet-4-20250514-thi...	1197	1,302	10	qwen-image	1051	79,245
View all				View all			

This allows for nuanced evaluation in complex domains.

Recap on different evaluation conditions

- Reference-based: (prompt, reference, generation)
- Reference-free:
 - Absolute rating (prompt, generation)
 - Comparison
(prompt, Model A generation, Model B generation)

Use powerful language models to judge!

Absolute rating

[System]

Please act as an impartial judge and evaluate the quality of the response provided by an AI assistant to the user question displayed below. Your evaluation should consider factors such as the helpfulness, relevance, accuracy, depth, creativity, and level of detail of the response. Begin your evaluation by providing a short explanation. Be as objective as possible. After providing your explanation, please rate the response on a scale of 1 to 10 by strictly following this format: "[[rating]]", for example: "Rating: [[5]]".

[Question]

{question}

[The Start of Assistant's Answer]

{answer}

[The End of Assistant's Answer]

Figure 6: The default prompt for single answer grading.

(Zheng et al, 2023)

May not tell the subtle differences between answers

Comparison

[System]

Please act as an impartial judge and evaluate the quality of the responses provided by two AI assistants to the user question displayed below. You should choose the assistant that follows the user's instructions and answers the user's question better. Your evaluation should consider factors such as the helpfulness, relevance, accuracy, depth, creativity, and level of detail of their responses. Begin your evaluation by comparing the two responses and provide a short explanation. Avoid any position biases and ensure that the order in which the responses were presented does not influence your decision. Do not allow the length of the responses to influence your evaluation. Do not favor certain names of the assistants. Be as objective as possible. After providing your explanation, output your final verdict by strictly following this format: "[[A]]" if assistant A is better, "[[B]]" if assistant B is better, and "[[C]]" for a tie.

[User Question]

{question}

[The Start of Assistant A's Answer]

{answer_a}

[The End of Assistant A's Answer]

[The Start of Assistant B's Answer]

{answer_b}

[The End of Assistant B's Answer]

(Zheng et al, 2023)

Scalability: $O(N^2)$ comparisons with N models

LLM-as-a-judge has reasonable agreement with human

Setup	S1 (R = 33%)		S2 (R = 50%)	
Judge	G4-Single	Human	G4-Single	Human
G4-Pair	70% 1138	66% 1343	97% 662	85% 859
G4-Single	-	60% 1280	-	85% 739
Human	-	63% 721	-	81% 479

Pair-wise comparison can distinguish nuance difference (tie generation) more accurately

(Zheng et al, 2023)

LLM judges have biases too

- **Positional Bias:** Tends to prefer the first response it sees.
- **Length Bias:** Tends to prefer longer, more verbose responses.
- **Self-Preference Bias:** Tends to prefer outputs from its own model family.

Judge	Prompt	Consistency	Biased toward first
Claude-v1	default	23.8%	75.0%
	rename	56.2%	11.2%
GPT-3.5	default	46.2%	50.0%
	rename	51.2%	38.8%
GPT-4	default	65.0%	30.0%
	rename	66.2%	28.7%

switch answer position

Table 3: Failure rate under “repetitive list” attack for different LLM judges on 23 answers.

Judge	Claude-v1	GPT-3.5	GPT-4
Failure rate	91.3%	91.3%	8.7%

Repeat the same content

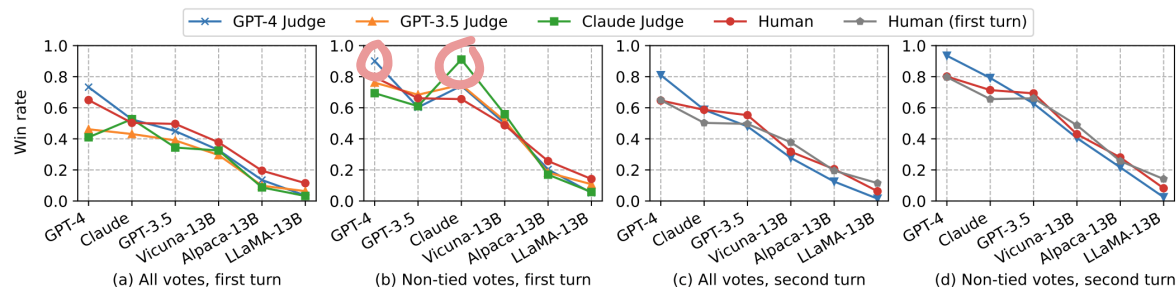


Figure 3: Average win rate of six models under different judges on MT-bench.

(Zheng et al, 2023)

LLM judges have biases too

- **Positional Bias:** Tends to prefer the first response it sees.
 - Conservative rating (evaluate twice with swapped position)
 - Randomized position
 - Few-shot demo
- **Length Bias:** Tends to prefer longer, more verbose responses.
 - Length normalization (Dubios et al 2024)
- **Self-Preference Bias:** Tends to prefer outputs from its own model family.
 - Multi-agent debating
 - Finetuning etc

Models can be optimized to hack human preferences

Humans aren't perfect judges either. We are biased by:

- Confident, assertive, or formal language.
- Longer responses that seem more comprehensive.
- Nicely formatted outputs (e.g., with markdown).

Models can be fine-tuned to be overly sycophantic or verbose to please human raters, sometimes at the cost of factuality.

A more robust paradigm: outcome-based evaluation

- Instead of judging *how* an answer is written, we check if it *works*.
- This moves evaluation from subjective quality to objective, verifiable results.
- Did the model's output achieve the desired goal?

The evolution of code benchmarks: from snippets to full repositories

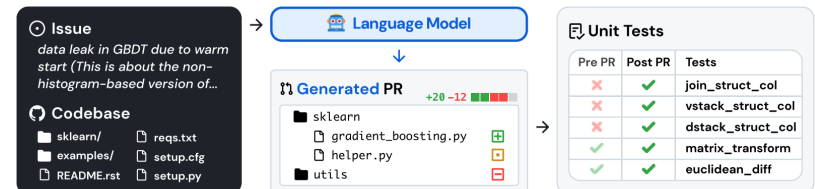
- **CoNaLa** (Yin et al 2018): One liner, n-gram matching eval
- **HumanEval** (Chen et al 2021): Simple programming tasks
- **Codeforce** (Li et al 2022): Advanced algorithmic reasoning.
- **SWE-bench** (Jimenez et al 2023): Solving actual GitHub issues in large codebases.

I₅: Converting integer to string in Python?

URL: <https://stackoverflow.com/questions/961632/>

Top Predictions:

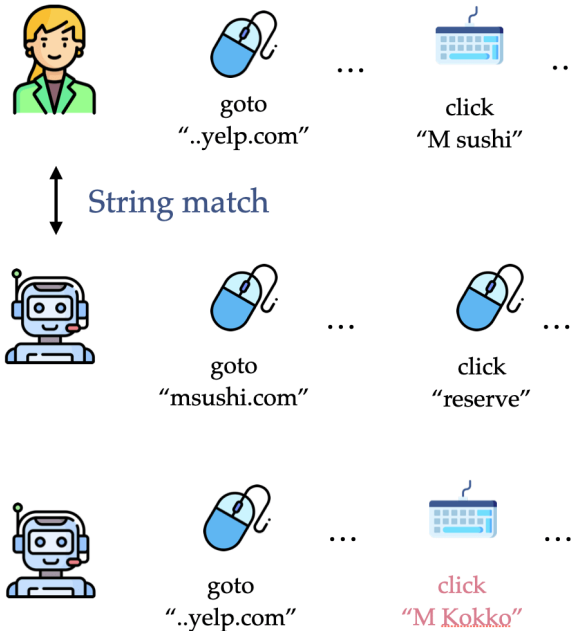
S₁ `int('10')` ✗
S₂ `str(10); int('10')` ✗
S₃ `a.__str__()` ✓



- From n-gram matching to execution-based evaluation
- Task complexity has skyrocketed.

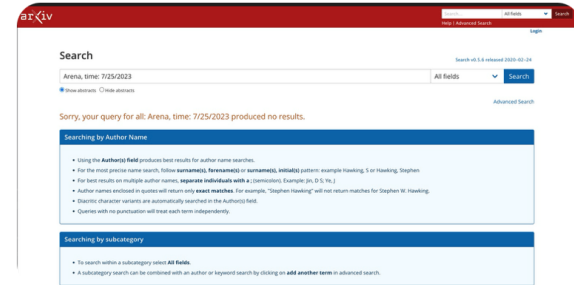
Exercise: web navigation evaluation

- **Task:** Book me a table at M sushi this Friday



- Falsely penalize alternative solutions
- High similarity != task success

LLM-as-a-judge for web navigation



Help me to judge if this trajectory is correct or not given the intent. You will answer with yes or no

Intent: Find a paper whose title includes "Arena" and was submitted to arxiv on 7/25/2023

Trajectory:

Action 1: goto [arxiv.com]

Action 2: search [Arena, time: 7/25/2023]

Action 3: I have review the page and there is no results. I think there is no paper whose title includes "Arena" and was submitted to arxiv on 7/25/2023. The answer is "There is no paper that satisfied your query"

Screenshot of last step:



Yes, the trajectory is correct given the intent.

The judge also only has partial information of the environment
Easy to fool

Outcome-based evaluation

Function	ID	Intent	Eval Implementation
$r_{\text{info}}(a^*, \hat{a})$	1	Tell me the name of the customer who has the most cancellations in the history	<code>exact_match(\hat{a}, "Samantha Jones")</code>
	2	Find the customer name and email with phone number 8015551212	<code>must_include(\hat{a}, "Sean Miller")</code> <code>must_include(\hat{a}, "sean@gmail.com")</code>
	3	Compare walking and driving time from AMC Waterfront to Randyland	<code>fuzzy_match(\hat{a}, "Walking: 2h58min")</code> <code>fuzzy_match(\hat{a}, "Driving: 21min")</code>
$r_{\text{prog}}(s)$	4	Checkout merge requests assigned to me	<code>url = locate_last_url(s)</code> <code>exact_match(url, "gitlab.com/merge_requests?assignee_username=byteblaze")</code>
	5	Post to ask "whether I need a car in NYC"	<code>url = locate_latest_post_url(s)</code> <code>body = locate_latest_post_body(s)</code> <code>must_include(url, "/f/nyc")</code> <code>must_include(body, "whether I need a car in NYC")</code>

WebArena (Zhou et al 2023)

Evaluation harnesses standardize and simplify benchmarking

Running all these different benchmarks can be complex.

Frameworks exist to provide a single entry point to run a model against multiple benchmarks:

- HELM (Holistic Evaluation of Language Models)
- EleutherAI Language Model Evaluation Harness
- Agentic tasks: Terminal Bench

But harnesses aren't perfect: prompting details matter

- Different harnesses may use slightly different prompt templates for the same task.
- Modern LLMs are highly sensitive to small changes in prompting (e.g., few-shot examples, formatting).
- The same model can get different scores on the same benchmark depending on which harness is used.

Practice: Always report the evaluation framework, prompting and other settings used for reproducibility.

The LLM evaluation lifecycle

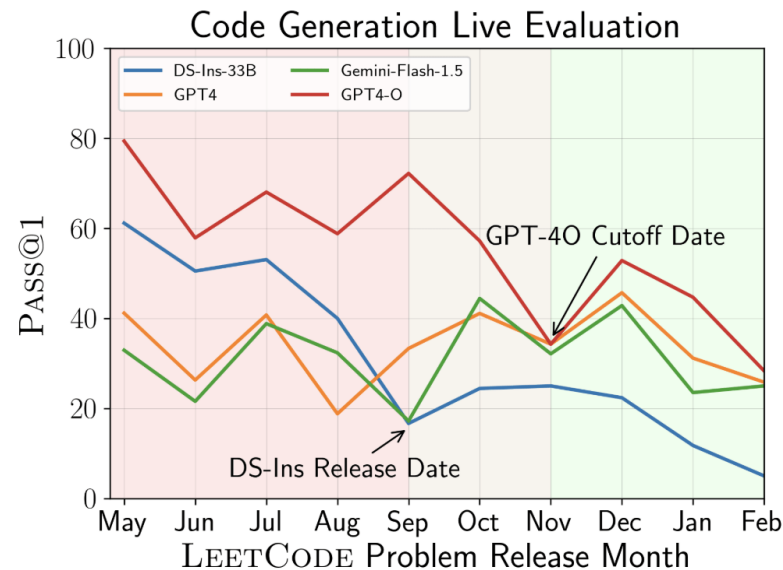
- **Pre-training:** Internal metrics like perplexity (PPL) measure next-word prediction ability.
- **Post-training / Fine-tuning:**
 - **Close-ended:** Accuracy, F1 on benchmarks like MMLU.
 - **Open-ended**
 - **Reference-based**
 - **Reference-free**
 - **Outcome-based**

A pervasive threat: test set contamination

- Contamination (or overfitting) is hard to avoid
 - Many datasets are stored as plain text on the internet
- Hard to detect (e.g., trained on paraphrased test set)
- **Solution:** Dynamic benchmarks with new, unseen data.

Fighting contamination with dynamic benchmarks

- LiveCodeBench (Jain et al 2024):
 - Continuously sources new problems from platforms like LeetCode



Performance drop for question set released after model's release date

Fighting contamination with dynamic benchmark

- **FreshQA** (Vu et al 2023):
 - Questions with changing answers over time

Type	Question	Answer (as of this writing)
never-changing	Has Virginia Woolf's novel about the Ramsay family entered the public domain in the United States?	Yes , Virginia Woolf's 1927 novel <i>To the Lighthouse</i> entered the public domain in 2023.
never-changing	What breed of dog was Queen Elizabeth II of England famous for keeping?	Pembroke Welsh Corgi dogs.
slow-changing	How many vehicle models does Tesla offer?	Tesla offers five vehicle models: Model S, Model X, Model 3, Model Y, and the Tesla Semi.
slow-changing	Which team holds the record for largest deficit overcome to win an NFL game?	The record for the largest NFL comeback is held by the Minnesota Vikings .
fast-changing	Which game won the Spiel des Jahres award most recently?	Dorfromantik won the 2023 Spiel des Jahres.
fast-changing	What is Brad Pitt's most recent movie as an actor	Brad Pitt recently starred in Babylon , directed by Damien Chazelle.
false-premise	What was the text of Donald Trump's first tweet in 2022, made after his unbanning from Twitter by Elon Musk?	He did not tweet in 2022.
false-premise	In which round did Novak Djokovic lose at the 2022 Australian Open?	He was not allowed to play at the tournament due to his vaccination status.

FreshQA

[FreshQA August 27, 2025](#)

Next update: September 3, 2025

We update our dataset weekly or upon request. If you find any updates or misclassifications in our FreshQA questions or answers that we may have overlooked, please notify us by commenting on the dataset spreadsheet above or sending an email to freshilms@google.com.

Older versions:

[FreshQA August 18, 2025](#)

[FreshQA August 11, 2025](#)

[FreshQA July 28, 2025](#)

[FreshQA July 22, 2025](#)

[FreshQA July 14, 2025](#)

[FreshQA July 7, 2025](#)

The cultural & linguistic bias of benchmarks

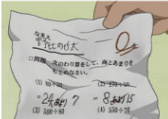
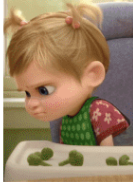

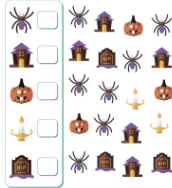


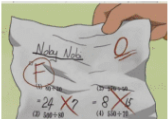


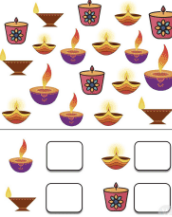


- Most popular benchmarks are English-centric.
- They often reflect Western cultural norms, values, and knowledge.

This creates a cycle where models are optimized for a narrow slice of the world's population, potentially harming equity.

Category	Task	Number of			Metric
		Datasets	Language	System Outputs	
Text Classification	Text Classification	127	12	199	Accuracy
Sequence Labeling	Named Entity Recognition	78	60	450	F1
	Word Segmentation	1	1	-	F1
	Chunking	2	1	-	F1
Cloze	Generative	10	1	-	CorrectCount
	Multiple Choice	21	2	-	Accuracy
Text Pair Classification	Text Pair Classification	57	30	96	Accuracy
Span Text Classification	Span Text Classification	4	1	-	Accuracy
Text Editing	Grammatical Error Correction	10	1	-	SeqCorrectCount
Question Answering	Extractive	80	18	185	F1
	Multiple Choice	72	2	-	Acc.
	Open Domain	4	2	-	ExactMatch
Conditional Generation	Machine Translation	242	60	170	Bleu
	Summarization	251	55	-	Bleu
	Code Generation	4	5	-	Bleu
KG Prediction	KG Prediction	3	1	28	Hits
Language Modeling	Language Modeling	-	-	-	Perplexity

GlobalBench (Song et al, 2023)

Cross-culture image translation

Audiovisual Media		Education		Advertisements	
Doraemon	Pixar <i>left: Inside Out; right: Zootopia</i>	Addition	Counting	Ferrero Rocher	Coca-Cola
 Japan	 US	 US	 US	 China	 Multiple (India, Pak..)
 US	 Japan	 India	 India	 Multiple (US, UK ..)	 Egypt

(Khanuja et al, 2024)

Evaluation is more than just performance

Performance (e.g., accuracy, Elo score) is just one axis. Other critical dimensions for real-world deployment include:

- **Efficiency:** Latency, computational cost, memory usage.
- **Fairness & Bias:** Does it perform equally well for all user groups?
- **Robustness:** How does it handle adversarial or out-of-distribution inputs?

A slightly worse model that is 10x faster and cheaper can be the better choice sometimes

Evaluating superhuman performance: scalable oversight

- **Problem:** How do we evaluate a model's reasoning on a task too complex for a human to verify? (e.g., finding a subtle security flaw in code).
- **One possible way:**
 1. Use a powerful model to *assist* a human judge.
 2. The human's job isn't to solve the problem, but to find flaws in the model's proposed solution.
 3. Break down complex problems into smaller, verifiable steps.

Key takeaways on evaluation

- Shift from simple, close-ended metrics to complex, open-ended comparisons.
- Human/LLM-based preference are gaining tractions, but is susceptible to biases and "hacking".
- Outcome-based evaluation is the most objective and robust method when applicable at the current stage.
- Beware of pitfalls: data contamination, cultural bias, and focusing only on performance.

Course logistics

Paper presentations

- Each presenter will present independently for one paper group
- Please discuss with your classmate who present at the same day to decide who will present which paper group.
- Prepare discussion questions.

Example projects

We will be discussing potential ideas for final projects in the coming weeks.

Google cloud credit

CS cluster tutorial