Evaluations of LLMs

(Optional)

AI with Deep Learning

EE4016

Prof. Lai-Man Po

Department of Electrical Engineering

City University of Hong Kong

https://medium.com/@lmpo/advanced-methods-for-assessing-large-language-models-a60f640e1240

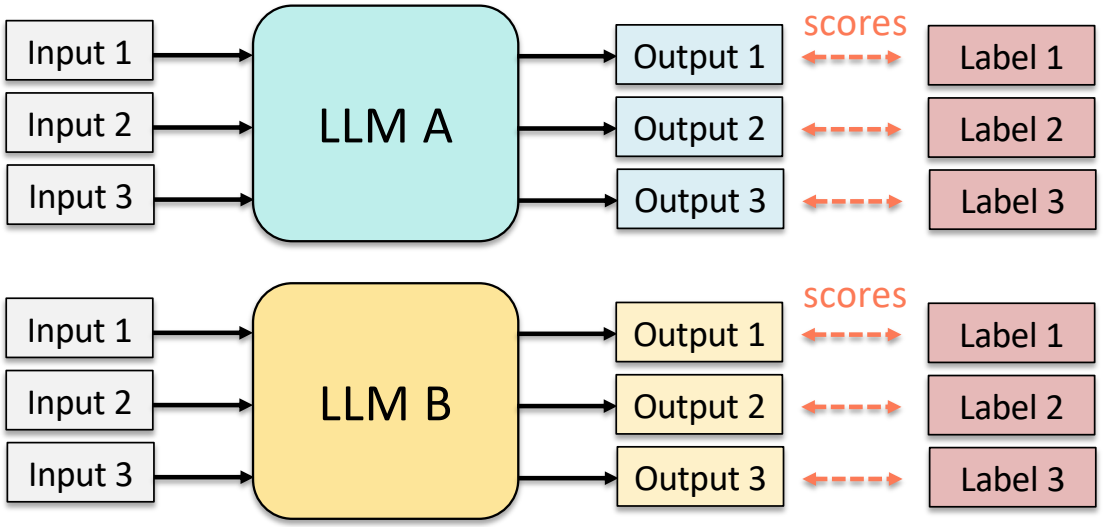
Meta Llama 3 Pre-trained model performance

	Meta Llama 3 8B	Mistral 7B		Gemma 7B	
		Published	Measured	Published	Measured
MMLU 5-shot	**66.6**	62.5	63.9	64.3	64.4
AGIEval English 3-5-shot	**45.9**	--	44.0	41.7	44.9
BIG-Bench Hard 3-shot, CoT	**61.1**	--	56.0	55.1	59.0
ARC-Challenge 25-shot	78.6	78.1	78.7	53.2 0-shot	**79.1**
DROP 3-shot, F1	**58.4**	--	54.4	--	56.3

	Meta Llama 3 70B	Gemini Pro 1.0	Mixtral 8x22B
		Published	Measured
MMLU 5-shot	**79.5**	71.8	77.7
AGIEval English 3-5-shot	**63.0**	--	61.2
BIG-Bench Hard 3-shot, CoT	**81.3**	75.0	79.2
ARC-Challenge 25-shot	**93.0**	--	90.7
DROP 3-shot, F1	**79.7**	74.1 variable-shot	77.6

What is a Metric?

- **Given "supervised data" how do we evaluate?**
  1. Run the model on the inputs to get predictions (outputs)
  2. Define a **metric (or score)** that estimates how well the model predictions reflect the **"gold" outputs** (labels).
  3. Compute the metric
- **How to compute a score?**
  - Human Evaluation
  - Automatic Evaluation



Evaluating LLMs: The Challenges

- **Traditional evaluation metrics**, such as accuracy, recall, precision, and F1 score, are **insufficient** for assessing the complexities of LLMs.
- **The Limitations of Traditional Metrics**
  - LLMs require a deep understanding of context, nuances, and subtleties in language, making traditional metrics inadequate for evaluation.
- **The Complexity of LLM Evaluation**
  - The evaluation of LLMs is further complicated by three key factors:
    1. **Contextual Understanding**: LLMs need to comprehend the context in which language is used, rendering traditional metrics insufficient.
    2. **Open-ended Generation**: LLMs can generate diverse responses, making it difficult to define a single "correct" answer.
    3. **Multifaceted Evaluation**: LLMs require assessment of various aspects, including fluency, coherence, and factual accuracy, which traditional metrics cannot capture.

Traditional NLP Evaluation Metrics

- **The Need for Tailored Evaluation Methods**
  - LLMs are applied in various real-world applications, such as text classification, sentiment analysis, and dialogue systems, each requiring tailored evaluation methods to accurately assess their performance.
- **Traditional NLP Metrics**
  - **Perplexity, BLUE, ROUGE**
  - **GLUE, SuperGLUE**

Task	Metric	Automatic Scoring Function
Classification	Accuracy	Exact Match: Did the model predict the same output as the gold output?
Question Answering	F1 Score	How many words are in common between the prediction and gold output?
Translation	ROUGE/BLEU	How many words/phrases are in common between the prediction and gold output?
Program Synthesis	Accuracy	Does the predicted code produce the same result as the output when run?
…	…	…

Language Model Evaluation: Perplexity

- When evaluating Language Models in isolation, one universal metric can provide valuable insights into their quality. This measure is called **Perplexity**:

$$\mathrm{PPL}(w_1, w_2, \ldots, w_N) = \exp\left(-\frac{1}{N}\sum_{i=1}^{N-1} \log_2 P(w_{i+1}|w_1, w_2, \ldots, w_i)\right)$$

  - w_i represents the i-th word.
  - $P(w_{i+1}|w_1, w_2, \ldots, w_i)$ is the probability of the *i+1*-th word given its preceding words,
  - N is the total number of words in the sequence,
- The intuition behind Perplexity, which is rooted in **entropy**, lies in measuring a model's uncertainty in predicting a sequence.
  - **A lower Perplexity score** indicates that the model is **less uncertain**, and therefore, more accurate in its predictions, making it a **better predictor** of the sample.

BLEU (2002)

- **BLEU** (**B**i-**L**ingual **E**valuation **U**nderstudy) is a metric used to evaluate the quality of machine-generated translations. It measures how closely a candidate translation matches reference translations by analyzing the precision of n-grams and applying a brevity penalty to discourage short translations.

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

  - *BP* is the brevity penalty,
  - N is the maximum n-gram order (typically 4),
  - w_n is the weight assigned to the n-gram precision,
  - p_n is the precision of n-grams of order n*n*.

Components of BLEU Score

1. **N-Gram Precision:** This measures how many n-grams in the candidate translation match those in the reference translations. For example, if the candidate translation contains several unigrams, bigrams, and trigrams that overlap with the references, these are counted to compute precision.
2. **Brevity Penalty (BP):** This penalizes translations that are shorter than their reference counterparts. It is calculated as

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

   - c is the length of the candidate translation,
   - r is the length of the closest reference translation.

BLEU Score Interpretation

- BLEU scores range from 0 to 1, with higher scores indicating better quality translations. A score of 0 means no overlap with reference translations, while a score of 1 indicates perfect overlap. However, achieving a perfect score is rare even for human translators.
- For example,

```
>>> sentence_1 = "I am so cold. It was freezing outside today. I could barely spend more than ten minutes outside of my car."
>>> sentence_2 = "I am freezing. It was so cold today outside my car. I could barely spend more than ten minutes outside."
>>> sentence_3 = "I do declare I am frigid. Just due to the low temperature, I decided to remain inside my vehicle."
```

```
>>> sentence_bleu([sentence_1], sentence_2)
0.8890800579779802
>>> sentence_bleu([sentence_1], sentence_3)
0.2381895638933463
```

- Text similarity with **BLEU drops** drastically by using **different words with similar meanings**

BLEU Score Limitations

- Despite its popularity, BLEU has several limitations:
  - **Synonymy and Paraphrasing:** BLEU does not account for synonyms or paraphrases, which can lead to lower scores for semantically equivalent translations.
  - **Correlation with Human Judgment:** While BLEU generally correlates well with human evaluations, it may not always reflect human preferences accurately, especially in cases requiring nuanced understanding or creativity.

ROUGE (2004)

- **ROUGE** (**R**ecall-**O**riented Understudy for **G**isting **E**valuation) is a set of metrics for evaluating text generation tasks like summarization and machine translation.
- Unlike BLEU, which focuses on precision, ROUGE emphasizes recall, making it useful for tasks where capturing all relevant information is critical.
- **Key variants of ROUGE include:**
  - **ROUGE-N:** Measures n-gram overlap between generated and reference texts.
  - **ROUGE-L:** Evaluates the longest common subsequence (LCS) between texts, capturing sentence-level structure similarity.
  - **ROUGE-S:** Assesses skip-bigram overlap, allowing for gaps between words and capturing semantic similarity beyond exact matches.

ROUGE Sub-Metrics

Metric	Algorithm	Best Use Case
ROUGE-N	Counts the overlap of N-grams between the system and reference summaries to measure precision and recall.	Especially useful for evaluating the precision and recall of specific N-gram matches in the summaries.
ROUGE-L	Finds the Longest Common Subsequence (LCS) between system and reference summaries, focusing on sequence order.	Best applied when assessing the fluency and the correct order of words in the summaries generated.
ROUGE-W	Enhances LCS by accounting for the length of consecutive sequences, giving higher scores to longer matches.	Ideal for scenarios where the significance of longer, contiguous matches outweighs isolated matches.
ROUGE-S	Evaluates skip-bigram co-occurrence statistics, allowing for more flexibility in word order within the summaries.	Effectively used in evaluating the presence of key information or phrases while permitting flexible word order.

The ROUGE scores are calculated using precision, recall, and F1-score:

1. Precision: The ratio of overlapping n-grams in the candidate text to the total number of n-grams in the candidate text.

$$\text{Precision} = \frac{\text{Number of overlapping n grams}}{\text{Total number of n grams in candidate}}$$

2. Recall: The ratio of overlapping n-grams in the candidate text to the total number of n-grams in the reference text

$$\text{Recall} = \frac{\text{Number of overlapping n grams}}{\text{Total number of n grams in reference}}$$

3. F1 Score: The harmonic mean of precision and recall, providing a balance between the two metrics.

$$\text{F1} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

Example

```
import evaluate

rouge = evaluate.load('rouge')

predictions = ["I really really loved reading reading the Hunger Games",
"Police killed the gunman"]
references = ["I really loved reading the Hunger Games",
"the gunman killed the police"]
results = rouge.compute(predictions=predictions, references=references)

# {'rouge1': 0.8819444444444445,
# 'rouge2': 0.7142857142857143,
# 'rougeL': 0.6597222222222223,
# 'rougeLsum': 0.659722222222223}
```

https://medium.com/@milana.shxanukova15/100-nlp-questions-answers-metrics-part-d8622bf73d98

Advantages and Limitations of ROUGE Score

- ROUGE metrics provide a more comprehensive evaluation than BLEU by incorporating recall and considering various forms of text similarity.
- **However, they also have limitations:**
  - **Synonymy and Paraphrasing:** Like BLEU, ROUGE does not account for synonyms or paraphrases, which can affect scoring.
  - **Coherence Assessment:** ROUGE primarily focuses on surface-level matches and may not adequately reflect overall coherence or quality in generated text.

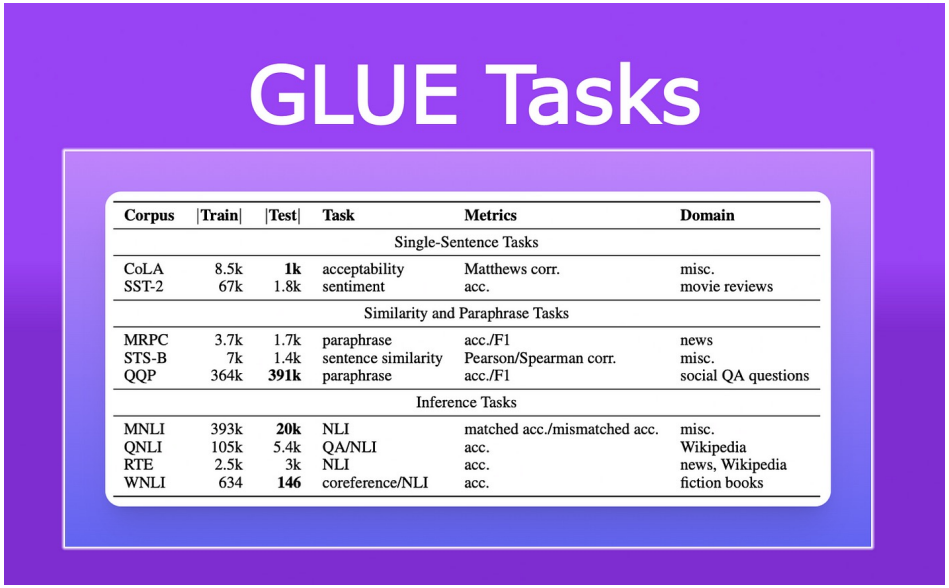
What is the difference between BLEU and ROUGE?

- **BLEU (Bilingual Evaluation Understudy)**
  - A metric commonly used in translation tasks. BLEU measures the similarity between a translated text and a reference translation.
  - It often calculates n-gram values, comparing two texts.
- **ROUGE (Recall-Oriented Understudy for Gisting Evaluation)**
  - A metric for summarization tasks. It also compares two texts (generated summary and source summary), based on the number of overlapped n-grams.
- Both metrics rely on overlapping n-grams.
  - The key difference is that BLEU is **precision orientated** and ROUGE is **recall orientated**, these differences are based on the metrics common usage.

GLUE (2018)

- **GLUE** (**G**eneral **L**anguage **U**nderstanding **E**valuation) is a collection of natural language tasks, such as sentiment analysis and question-answering.
- GLUE was created to encourage the development of models that can **generalize across multiple tasks**, and you can use the benchmark to measure and compare the model performance.

1. Benchmark of 9 sentence-pair language understanding tasks
2. A diverse range of dataset sizes, text genres, and degrees of difficulty



Corpus	|Train|	|Test|	Task	Metrics	Domain
			Single-Sentence Tasks		
CoLA	8.5k	**1k**	acceptability	Matthews corr.	misc.
SST-2	67k	1.8k	sentiment	acc.	movie reviews
			Similarity and Paraphrase Tasks		
MRPC	3.7k	1.7k	paraphrase	acc./F1	news
STS-B	7k	1.4k	sentence similarity	Pearson/Spearman corr.	misc.
QQP	364k	**391k**	paraphrase	acc./F1	social QA questions
			Inference Tasks		
MNLI	393k	**20k**	NLI	matched acc./mismatched acc.	misc.
QNLI	105k	5.4k	QA/NLI	acc.	Wikipedia
RTE	2.5k	3k	NLI	acc.	news, Wikipedia
WNLI	634	**146**	coreference/NLI	acc.	fiction books

SuperGLUE (2019)

- **SuperGLUE** is the successor to GLUE, introduced in 2019.
- To address the limitations of GLUE, **SuperGLUE includes a series of tasks, some of which are new, and some are more challenging versions of tasks found in GLUE.**

Leaderboard Version: **2.0**

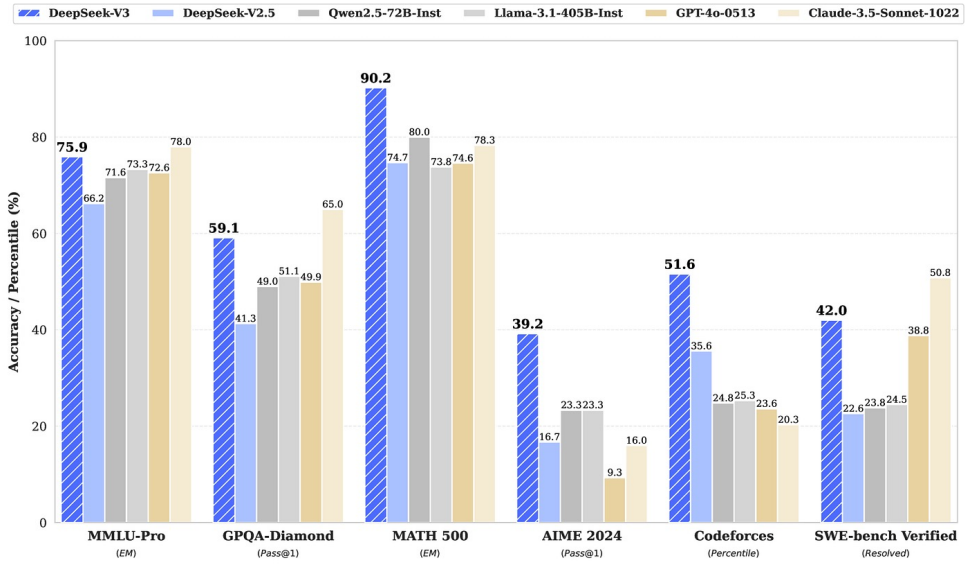
	Rank	Name	Model	Score
	1	JDExplore d-team	Vega v2	91.3
✚	2	Liam Fedus	ST-MoE-32B	91.2
	3	Microsoft Alexander v-team	Turing NLR v5	90.9
	4	ERNIE Team - Baidu	ERNIE 3.0	90.6
	5	Yi Tay	PaLM 540B	90.4
✚	6	Zirui Wang	T5 + UDG, Single Model (Google Brain)	90.4
✚	7	DeBERTa Team - Microsoft	DeBERTa / TuringNLRv4	90.3
	8	SuperGLUE Human Baselines	SuperGLUE Human Baselines	89.8
✚	9	T5 Team - Google	T5	89.3
	10	SPoT Team - Google	Frozen T5 1.1 + SPoT	89.2

Advanced LLM Evaluation: Benchmarks vs. Human Judgment

https://medium.com/@lmpo/llm-evaluation-benchmarks-vs-human-judgment-f1cdd16098c0

LLM Evaluation

- How do we know which AI model is best?
- Benchmarks and leaderboards provide rankings, but evaluating LLMs is more nuanced.
- Evaluation methods fall into two families:
  - **Objective (benchmark-based)**
  - **Subjective (judgment-based)**



Four Core Evaluation Approaches

- LLM evaluation is multifaceted and categorized into two types:
  - **Benchmark-Based Evaluation (objective)**
    1. Multiple-choice accuracy tests
    2. Verifier-based reasoning checks
  - **Judgment-Based Evaluation (subjective)**
    1. Leaderboards and preference ranking
    2. LLM-as-a-Judge systems

Method 1: Evaluating Answer-Choice Accuracy
(Multiple-Choice)

- Measures LLMs by answering predefined **multiple-choice questions (MCQs)**.
- Performance is measured by the **fraction of correct responses**, mirroring traditional exams.
- Examples:
  - **MMLU / MMLU-Pro** — General knowledge across 57 subjects
  - **GPQA / GPQA Diamond** — Expert-level scientific reasoning
  - ARC, HellaSwag — Commonsense and reasoning tasks
- Limitations: Focuses on recall, not reasoning depth.

MMLU (2021)

- **Massive Multitask Language Understanding** (MMLU) is a comprehensive benchmark designed to evaluate the performance of LLMS across a wide array of tasks and domains.
  - **It measures LLM performance on 57 diverse knowledge intensive tasks**

Multiple Choices Questions

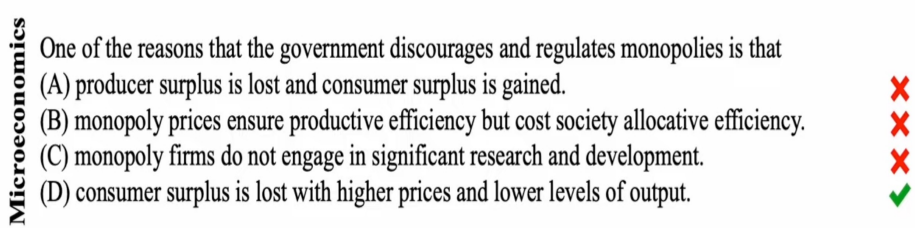


Figure 3: Examples from the Microeconomics task.

Task	Tested Concepts	Supercategory
Abstract Algebra	Groups, rings, fields, vector spaces, ...	STEM
Anatomy	Central nervous system, circulatory system, ...	STEM
Astronomy	Solar system, galaxies, asteroids, ...	STEM
Business Ethics	Corporate responsibility, stakeholders, regulation, ...	Other
Clinical Knowledge	Spot diagnosis, joints, abdominal examination, ...	Other
College Biology	Cellular structure, molecular biology, ecology, ...	STEM
College Chemistry	Analytical, organic, inorganic, physical, ...	STEM
College Computer Science	Algorithms, systems, graphs, recursion, ...	STEM
College Mathematics	Differential equations, real analysis, combinatorics, ...	STEM
College Medicine	Introductory biochemistry, sociology, reasoning, ...	Other
College Physics	Electromagnetism, thermodynamics, special relativity, ...	STEM
Computer Security	Cryptography, malware, side channels, fuzzing, ...	STEM
Conceptual Physics	Newton's laws, rotational motion, gravity, sound, ...	STEM
Econometrics	Volatility, long-run relationships, forecasting, ...	Social Sciences
Electrical Engineering	Circuits, power systems, electrical drives, ...	STEM
Elementary Mathematics	Word problems, multiplication, remainders, rounding, ...	STEM
Formal Logic	Propositions, predicate logic, first-order logic, ...	Humanities
Global Facts	Extreme poverty, literacy rates, life expectancy, ...	Other
High School Biology	Natural selection, heredity, cell cycle, Krebs cycle, ...	STEM
High School Chemistry	Chemical reactions, ions, acids and bases, ...	STEM
High School Computer Science	Arrays, conditionals, iteration, inheritance, ...	STEM
High School European History	Renaissance, reformation, industrialization, ...	Humanities
High School Geography	Population migration, rural land-use, urban processes, ...	Social Sciences
High School Gov't and Politics	Branches of government, civil liberties, political ideologies, ...	Social Sciences

https://arxiv.org/pdf/2009.03300v3.pdf

Example questions from MMLU train datasets

```
{'input': 'Which of the following statements is correct (according to knowledge in 2020)?\n',
 'A': 'Consumers with phenylketonuria must avoid the consumption of the sweetener aspartame',
 'B': 'Consumers with phenylketonuria must avoid the consumption of the sweetener saccharin',
 'C': 'Consumers with phenylketonuria must avoid the consumption of the sweetener sucralose',
 'D': 'Consumers with phenylketonuria must avoid the consumption of the sweetener acesulfame K',
 'target': 'A'}
```

```
{'input': 'Find the characteristic of the ring 2Z.',
 'A': '0',
 'B': '3',
 'C': '12',
 'D': '30',
 'target': 'A'}
```

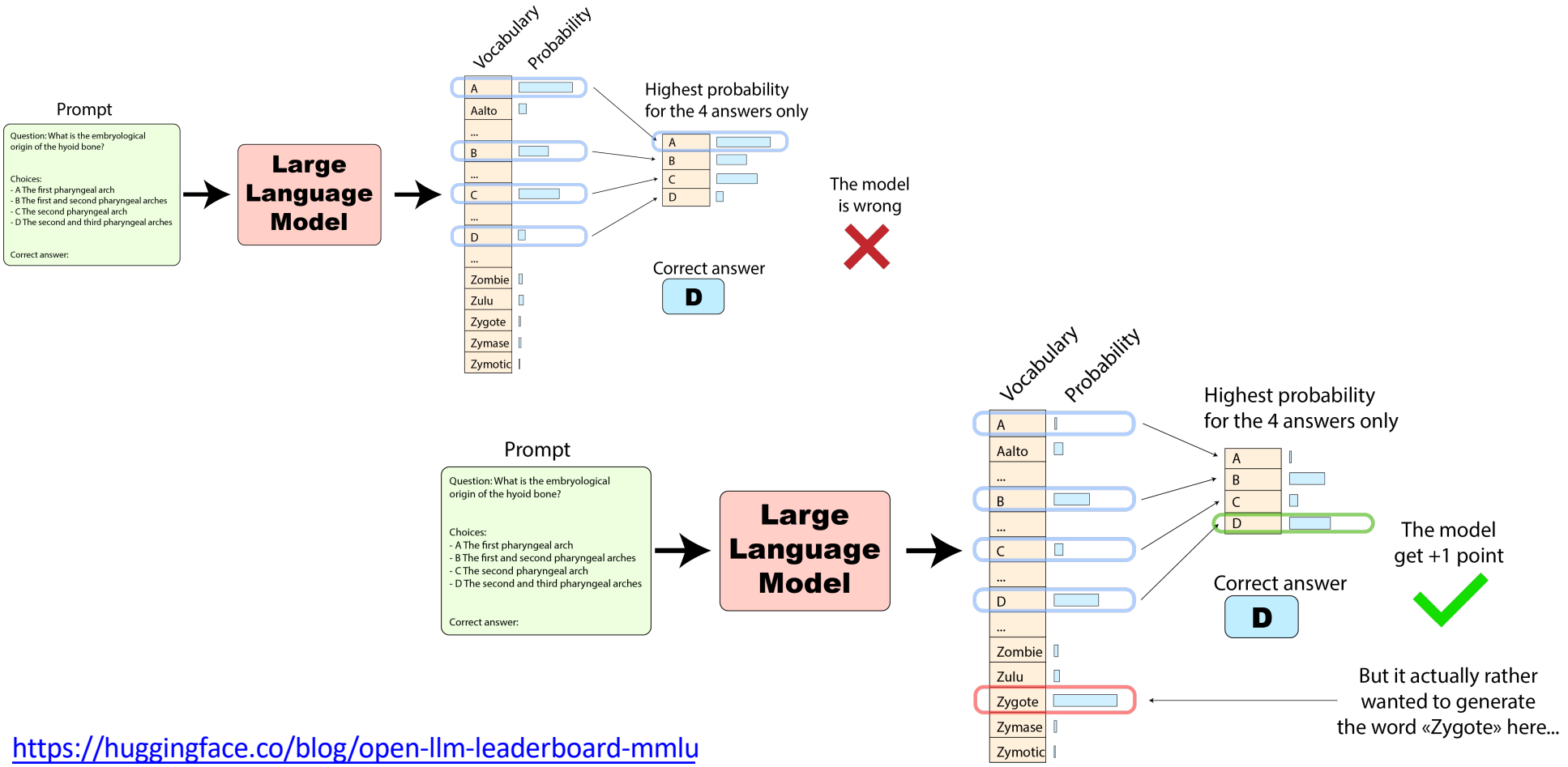
MMUL: Multiple Choice Questions

Find all c in $\mathbb{Z}_3$ such that $\mathbb{Z}_3[x]/(x^2 + c)$ is a field.
(A) 0 **(B) 1** (C) 2 (D) 3

A point pole has a strength of $4\pi \times 10^{-4}$ weber. The force in newtons on a point pole of $4\pi \times 1.5 \times 10^{-4}$ weber placed at a distance of 10 cm from it will be
(A) 15 N. (B) 20 N. (C) 7.5 N. (D) 3.75 N.

From the solubility rules, which of the following is true?
(A) All chlorides, bromides, and iodides are soluble
(B) All sulfates are soluble
(C) All hydroxides are soluble
(D) All ammonium-containing compounds are soluble

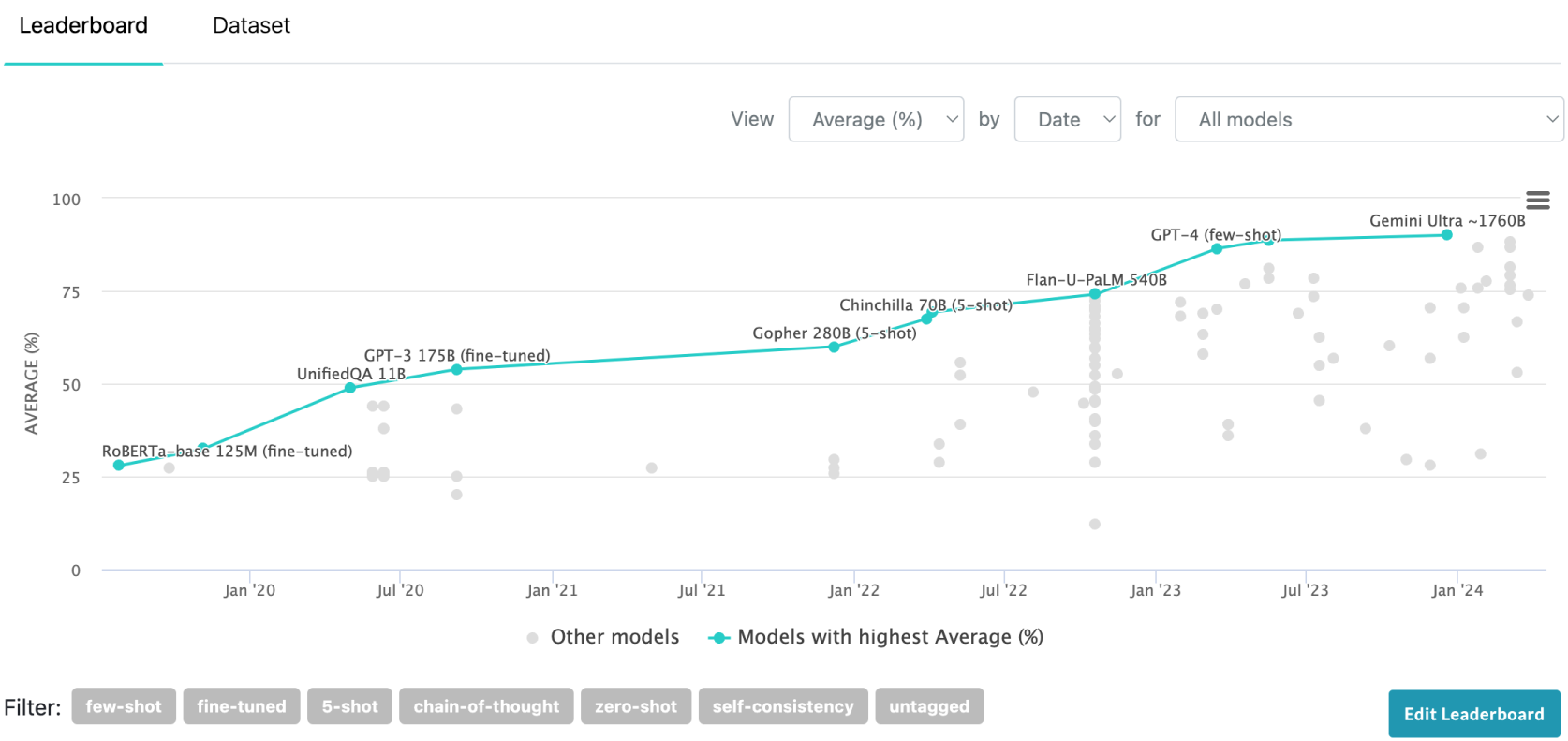
MMLU Implementation



https://huggingface.co/blog/open-llm-leaderboard-mmlu

Multi-task Language Understanding on MMLU

- Leaderboard: https://pub.towardsai.net/llm-benchmarks-in-2024-45226a1fcb54



MMLU using Different Datasets

	MMLU (HELM)	MMLU (Harness)	MMLU (Original)
llama-65b	**0.637**	0.488	**0.636**
tiiuae/falcon-40b	0.571	**0.527**	0.558
llama-30b	0.583	0.457	0.584
EleutherAI/gpt-neox-20b	0.256	0.333	0.262
llama-13b	0.471	0.377	0.47
llama-7b	0.339	0.342	0.351
tiiuae/falcon-7b	0.278	0.35	0.254
togethercomputer/RedPajama-INCITE-7B-Base	0.275	0.34	0.269

AI2 Reasoning challenge (ARC) 2018

- ARC is a set of grade-school science questions with 7,787 **multiple-choice science questions.**
- These questions are designed to be answerable with reasoning and knowledge that a typical 8th grader would be expected to possess.
- Dataset weights 681MB and is divided into 2 sets of questions:
  - ARC-Easy
  - ARC-Challenge
- Shot format

Example questions

```
{'answerKey': 'B',
 'choices': {'label': ['A', 'B', 'C', 'D'],
             'text': ['valleys carved by a moving glacier',
                      'piles of rocks deposited by a melting glacier',
                      'grooves created in a granite surface by a glacier',
                      'bedrock hills roughened by the passing of a glacier']},
 'id': 'Mercury_SC_401653',
 'question': 'Which land form is the result of the constructive force of a '
             'glacier?'}
```

"Think you have Solved Question Answering? Try ARC, the AI2 Reasoning Challenge".

HellaSwag (2019)

- **HellaSwag** is the successor to the previous SWAG dataset, presenting a greater challenge for models to achieve high performance.
  - Harder Endings, Longer contexts, and Low-shot Activities for Situations With Adversarial Generations
- This dataset is designed **to evaluate models' capabilities in commonsense reasoning**, particularly their ability to predict or complete a sentence in a way that makes sense.
- 10-shot format
- Dataset weights 71.5MB.

```
{'activity_label': 'Getting a haircut',
 'ctx': 'He scrubs in the shampoo and then washes it off. He then combs it and '
        'blow dries his hair after styling it with gel. he',
 'ctx_a': 'He scrubs in the shampoo and then washes it off. He then combs it '
          'and blow dries his hair after styling it with gel.',
 'ctx_b': 'he',
 'endings': ['then rinses it off in the sink.',
             'lets his hair down, giving it a final blow dry before putting it '
             'next to his face.',
             'uses an electric clipper to groom the sideburns and the temples.',
             'then sprays some liquid on his hair.'],
 'ind': 31,
 'label': '2',
 'source_id': 'activitynet~v_-JqLjPz-07E',
 'split': 'train',
 'split_type': 'indomain'}
```

Example questions

[HellaSwag: Can a Machine Really Finish Your Sentence?]

Method 2: Using Verifiers to Check Answers

- **Tests open-ended reasoning via automated verification tools.**
- Examples:
  - GSM8K — Math word problems
  - **MATH-500** — Advanced math reasoning
  - **AIME 2024** — Competitive problem solving
  - **Codeforces, SWE-bench**, HumanEval — Programming and code correctness

- **Limitations**: Only applies to verifiable domains and ignores reasoning quality.

Judgment-Based Evaluation

- **Evaluates subjective qualities** like helpfulness, coherence, and style through user or model preferences.
- Key methods:
  - Leaderboards and preference ranking (e.g., Chatbot Arena)
  - LLM-as-a-Judge (GPT-4, Prometheus)

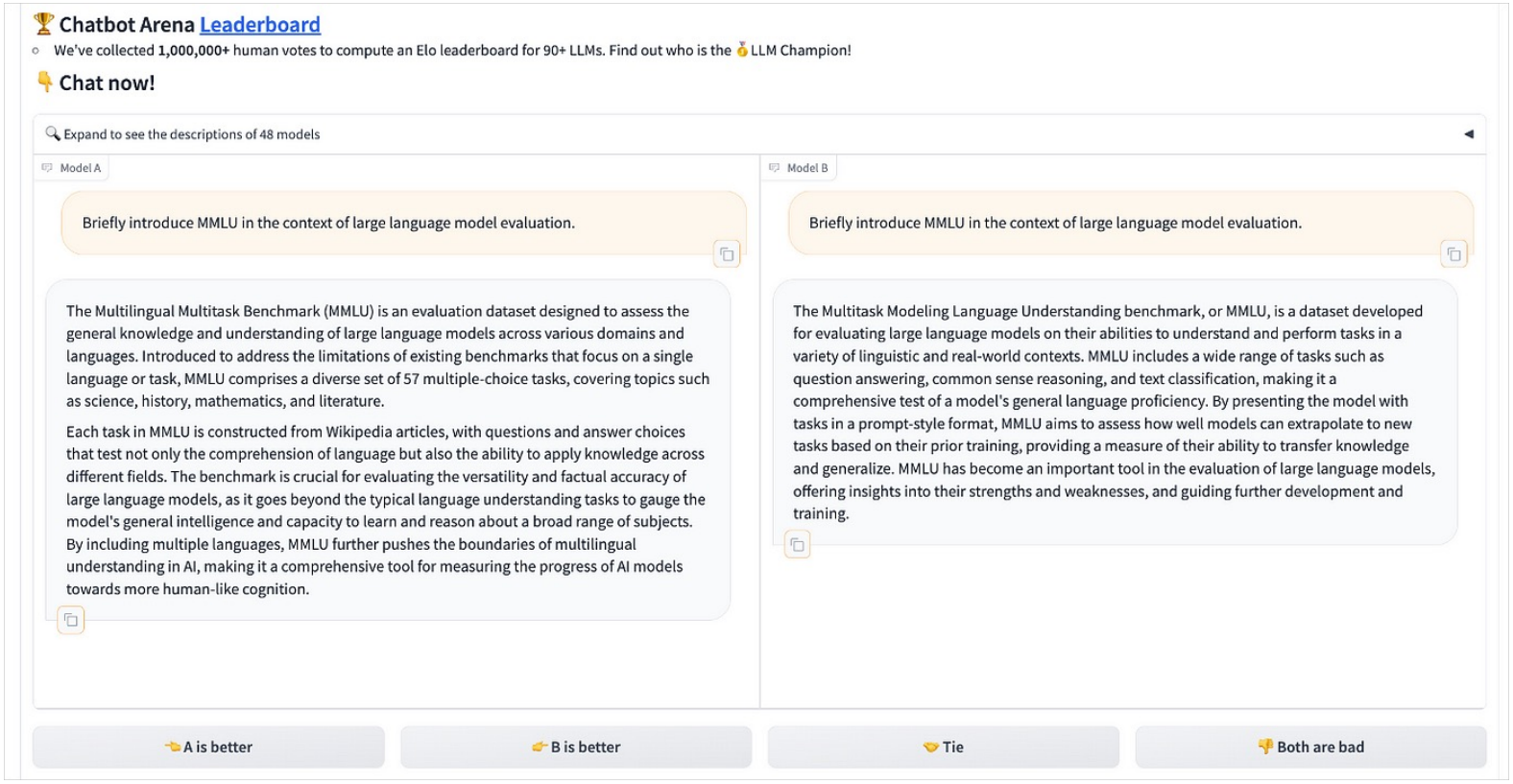
- Focus: 'Is this better?' rather than 'Is this right?'

Method 3: Leaderboards and Preference Ranking

- **Models are compared pairwise and ranked by preference.**
- Systems:
  - **Chatbot Arena / LMSYS** — Human pairwise voting
  - Elo, Bradley–Terry, TrueSkill — Statistical ranking models

- Limitations: Biased by demographics, costly, and subjective.

LMArena



https://lmarena.ai/?mode=side-by-side

LMArena LeaderBoard

Text Arena

View rankings across various LLMs on their versatility, linguistic precision, and cultural context across text.

Last Updated	Total Votes	Total Models
Oct 17, 2025	4,278,480	258

Overall | Search by model name... / | Default

Rank (UB)	Model	Score	95% CI (±)	Votes	Organization	License
1	gemini-2.5-pro	1451	±4	54,087	Google	Proprietary
1	claude-opus-4-1-20250805-thinking-16k	1447	±5	21,306	Anthropic	Proprietary
1	claude-sonnet-4-5-20250929-thinking-32k	1445	±8	6,287	Anthropic	Proprietary
1	gpt-4.5-preview-2025-02-27	1441	±6	14,644	OpenAI	Proprietary
2	chatgpt-4o-latest-20250326	1440	±4	40,013	OpenAI	Proprietary
2	o3-2025-04-16	1440	±4	51,293	OpenAI	Proprietary
2	claude-sonnet-4-5-20250929	1438	±8	6,144	Anthropic	Proprietary
2	gpt-5-high	1437	±5	23,580	OpenAI	Proprietary

https://lmarena.ai/leaderboard/text

LLM-as-a-Judge

- **Automated judgment using advanced LLMs as evaluators.**
- Examples:
  - GPT-4, Prometheus 2, PandaLM, Auto-J frameworks

- **Advantages:** Scalable and fast.
- **Limitations:** Bias from judge model, prompt sensitivity, and reproducibility issues.

Synthesis and Recommendations

- Each evaluation type has trade-offs:
  - Benchmarks — Objective and reproducible, but narrow
  - Verifiers — Good for math/code, limited scope
  - Leaderboards — Capture user perception, subjective
  - LLM-as-a-Judge — Scalable, but dependent on judge reliability

- **Best practice:** Combine multiple evaluation methods for balanced assessment.

Conclusion

A Smarter Way to Think About AI Quality

- A model’s benchmark score isn’t the whole story. True understanding requires knowing how it was evaluated.
- A comprehensive evaluation framework integrates objective benchmarks and human-like judgment to reflect real-world performance and communication quality.